Title

Interpretable and efficient statistical approaches for biomedical data

Author

Li, Xiao

Publication Date

2021

Peer reviewed|Thesis/dissertation

Interpretable and efficient statistical approaches for biomedical data

by

Xiao Li

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Co-chair
Assistant Professor William Fithian, Co-chair
Assistant Professor Gokul Upadhyayula

Spring 2021

Interpretable and efficient statistical approaches for biomedical data

Abstract

Interpretable and efficient statistical approaches for biomedical data

by

Xiao Li

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Co-chair

Assistant Professor William Fithian, Co-chair

Statistics and machine learning have achieved remarkable successes in solving data problems including driving new biomedical discoveries. In particular, prediction and hypothesis testing are two important applications of statistics and machine learning to biomedical data. In this dissertation, we will investigate how appropriate interpretations of prediction algorithms, and scrutiny of efficiency of hypothesis testing techniques can help us extend the capability of statistical and machine learning approaches in biomedical science.

Chapter 1 of this dissertation provides an overview of the topics covered, as well as the background information for the rest of the dissertation. Chapters 2 and 3 introduce the applications of an interpretable machine learning prediction pipeline for two biomedical problems: drug response prediction, and molecular partner prediction in clathrin-mediated endocytosis. In the drug response prediction task, our predictive and stability-driven pipeline achieves the state-of-the-art performance in identifying stable, predictive -omics features for drug response. In the molecular partner prediction task, we developed a interpretable deep learning model that achieves state-of-the-art accuracy in predicting whether a clathrin-coated pit is abortive or valid.

Chapter 4 focuses on the interpretation of a specific algorithm: random forest. Random forest has witnessed numerous applications in biomedical sciences, and its interpretation has become an important topic of research. We derived the first finite sample bound on the bias of Mean Decrease Impurity, one of the most widely used measure of feature importance. To reduce this bias, we proposed a new feature importance measure, called MDI-oob. MDI-oob achieves state-of-the-art performance in feature selection from random forest in biology inspired simulations.

Chapters 5 and 6 aim to provide a more comprehensive understanding of some of the most popular high dimensional tests for biomedical data. Of particular interest is the comparison between special-purpose tests with the Bonferroni correction (or the closely associated max test in global

testing), a simple and transparent test whose Type-I error (false positive) is robust to arbitrary dependence between p-values of univariate null hypotheses. In the context of global testing, we showed that the max test is optimal for detecting sparse signals, provided that the distribution of the signals has Gaussian or heavier tails. We also derived the first general negative results for knockoff methods. We give realistic conditions on the covariance matrix of the design matrix under which the true positive rate of the best achievable knockoff method must converge to zero, even when the true positive rate of Bonferroni correction converges to 1.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

I have had the privilege of working with and learning from many exceptional researchers and professionals during my five years at Berkeley. First and foremost, I would like to thank my two advisors: Bin Yu and Will Fithian. Bin has been a role model for me in many aspects, and I have benefited tremendously from the academic training under her guidance. Her vision for statistics and data science has undoubtedly shaped my understanding of the field and my research taste. Before coming to Berkeley, I did not have much experience working on applied statistics or collaborative projects. Bin provided me with the opportunity and environment to engage in exciting applied projects, and was patient in preparing me, both mentally and technically, for this task. I am deeply indebted to her for the time and effort she spent on me. Working under her in the group is such a fantastic experience. I am constantly in awe of her rigor, creativity, and high standards in driving research projects. During the pandemic, she led the group into an effort to support Response4Life, a non-profit organization working on distributing PPEs to areas that need them the most. Being part of this project, I felt her strong sense of social responsibility, and passion for doing research for the common good. Her mentorship has helped me grow as both a researcher and a person.

I am deeply grateful for Will's guidance and support during my PhD, and have benefited hugely from his high calibre in conducting research. Will guided me into my first research project at Berkeley, and since day one, I have been amazed by his sharpness, instincts, and his deep understanding of statistical inference methods. There were many times when it took me days to understand an argument that Will made during our meetings. I have also learned a lot of skills from him on writing technical papers and giving presentations. Will has been very supportive of my professional development, from helping me to find internship opportunities when I had no work experience, to serving as the chair of my qualifying exam committee, to encouraging me to present at a research conference.

I also owe a lot debt to many other faculty and staff members. I am grateful for Prof. Srigokul Upadhyayula for his advising on the work that forms Chapter 3 of this thesis, and for offering biology insights that help me appreciate the beauty and impact of the problem. I would like to thank Prof. Martin Wainwright for his mentorship and generous support during my research internship at the Voleon Group. Thank you to Prof. Jennifer Listgarten and Sandrine Dudoit for sitting on my qualifying exam committee, and offering valuable feedbacks. Thank you to Dr. Jean-Pierre Kocher for his support on the drug response prediction data. I am also indebted to Prof. Karl Rohe, whose guidance during the final years of my undergraduate study led me to Berkeley. I feel grateful for all the wonderful staff members of the department, especially La Shana Porlaris for always offering solutions and reassurance in time, and Chris Paciorek and Ryan Lovett for their indispensable support on the computing resources.

Next, I would like to thank my brilliant collaborators. Big thanks to Yu Wang, who has been a close collaborator on many projects, and whom I had countless lunches and conversations with in the group office; to Tiffany Tang for working together on the drug response prediction project, and not giving up when the prediction accuracy never seemed to go up; to Chandan Singh for being an incredibly organized and thoughtful colleague on the CME and other projects, and occasional basketball companion; to Karl Kumbier for his support in the early years of my PhD; and to

Xuewei Wang, Xiongtao Ruan, and Sumanta Basu for fruitful discussions on the work presented in Chapters 2-4 of this thesis.

I am also grateful for my other colleagues and friends. The Yu group is my closest academic family and I feel really fortunate to be surrounded with such a group of talented young researchers whom I can always bounce ideas off, and whom I learnt so much technical, communication and presentation skills from. Besides those who collaborated with me directly on the work presented in this thesis, I would like to thank those Yu group members, especially Nick Altieri and Raaz Dwivedi for working closely together on the COVID19 project, and patiently going through many rounds of revisions; Merle Behr for our collaboration on the random forest theory project, and intellectual conversations on many topics along the way; Yuansi Chen, Jamie Murdoch, Simon Walter, Rebecca Barter, Wooseok Ha and Yan Shuo Tan, for insightful discussions. I would like to thank my cohort, especially Eli Ben-Michael, Koulik Khamaru, Bryan Liu, Jake Soloff and Zhiyi You, for struggling with me through the first year sequence, and Joe Borja, Tom Zhiyue Hu, Kenneth Hung, Lihua Lei, Tianyi Lin, Shamindra Shrotriya, Chiao-Yu Yang, Yuting Wei, and Yumeng Zhang for their support and encouragement.

Finally, I would like to express my gratitude to my parents for their unconditional love. You believe in me and support me in the decisions I make. This dissertation would not have been possible without you.

Chapter 1

Overview

Statistics and machine learning have been powerful tools to drive breakthroughs in biomedical sciences. While the scope of their applications to biomedical sciences has been greatly extended in recent years, a significant proportion of these applications can be classified into the following two categories:

1. Prediction of a biologically meaningful target, such as the prediction of 3D protein structure [110] based on sequences of amino acids, and the prediction of individual response to anti-cancer drugs [11] based on the patient's genomic profile

2. Statistical hypothesis testing, often of many hypothesis, whose applications include genome-wide association studies (GWAS), PheWAS [33], burden tests [72], and so on.

Although voluminous literature has been published in each category, we want to emphasize some desiderata in these applications that we believe deserve further attention. First, we want to stress the importance of building interpretation pipelines of predictive models in biomedical applications. Interpretation of a machine learning model is the extraction of relevant knowledge from the model [91]. These knowledge can offer new insights to the domain problem, and provide guidance for follow-up lab experiments. For example, interpreting a drug response prediction model can help identify predictive disease biomarkers, which could potentially be used as therapeutic targets of precision medicine. Appropriate interpretation can thus utilize machine learning models to a greater extent, and build trust with domain experts.

Turning to the hypothesis testing front, we note that although an increasing number of testing approaches for large biomedical data become available, the two types of error (false positive and false negative) of these approaches are usually only investigated under stylized statistical models. In practice, the assumptions underlying these models are usually not, and sometimes could not be checked. As such, it is a crucial task to examine the performance of these approaches when the true model deviates from these idealized models, and identify scenarios where these approaches may fail.

These desiderata serve as an important guideline for the work in this dissertation. Our results will be presented under three thrusts. The first thrust consists of Chapters 2 and 3. In this thrust, we

showcase two examples of interpretable machine learning pipelines for problems arising in drug response prediction and molecular partner prediction, respectively. The data used for drug response prediction is provided by Dr. Jean-Pierre A. Kocher and Xuewei Wang from the Mayo clinic, and the data for molecular partner prediction is provided by Professor Srigokul Upadhyayula and his colleagues at the Advanced Bioimaging Center, UC Berkeley. We interact closely with these domain experts when working on the provided data. The second thrust comprises Chapter 4. In the second thrust, we study the interpretation of Random Forest, a machine learning method that has wide ranging application in biomedical problems such as GWAS [34] and recovering gene regulatory networks [58]. We quantify the "bias" of Mean Decrease Impurity, one of the most popular way to interpret a random forest, in terms of the hyperparameters of the random forest, and suggest a way to reduce this bias. The final thrust consists of Chapters 5 and 6. In this thrust, we investigate the performance of some popular high dimensional testing techniques, and identify realistic scenarios where special-purpose tests such as the high criticism tests and the knockoff filter must have trivial power, even when the power of the Bonferroni correction approaches 1.

The remainder of this introductary chapter provides a more detailed overview on the background of these thrusts and the contributions of this thesis.

1.1 Thrust I: Interpretable prediction pipelines for biomedical data

Interpretable, stability-driven drug response production

Modern cancer -omics and pharmacological data hold great promise in precision cancer medicine for developing individualized patient treatments. However, high heterogeneity and noise in such data pose challenges for predicting the response of cancer cell lines to therapeutic drugs accurately. As a result, arbitrary human judgment calls are rampant throughout a predictive modeling pipeline.

We developed a transparent stability-driven pipeline for drug response interpretable predictions, or staDRIP, which builds upon the PCS framework for veridical data science [135] and mitigates the impact of human judgment calls. Here we use the Predictability, Computability and Stability (PCS) framework for the first time in cancer research to extract proteins and genes that are important for predicting the drug responses and are stable across appropriate data and model perturbations. StaDRIP consists of three steps. In the first step, we fit many models , including those that are state-of-the-art in the literature, to predict drug sensitivity, and filter out models with poor prediction accuracy. Then, for each model with high prediction accuracy, we perturb the data by generating different bootstrap samples, and evaluate the stability of different predictors under this data perturbation. Finally, we extract proteins and genes that are the most stable and important across all models with high prediction accuracy.

We tested our staDRIP pipeline using data from the Cancer Cell Line Encyclopedia (CCLE). Out of the 24 most stable proteins we identified, 18 have been associated with the drug response or identified as a known or possible drug target in previous literature, demonstrating the utility of our PCS-driven pipeline for knowledge discovery in cancer drug response prediction modeling.

Interpretable deep learning for accurate molecular partner prediction in clathrin-mediated endocytosis

Understanding clathrin-mediated endocytosis (CME) is a crucial question in cell biology. One major challenge with analyzing CME is the ability to distinguish whether a clathrin-coated pit is abortive or valid. In this thesis, we address this challenge in two steps. First, we use a secondary marker, auxilin 1, to identify valid events. Second, we build a predictive pipeline using a Long Short Term Memory (LSTM) neural network to predict whether a CME event is abortive or valid. On tracks with lifetime greater than 15 seconds, our LSTM network achieves 84.1% prediction accuracy, which is a 5% improvement upon previous state-of-the-art method. Furthermore, we also proposed an interpretation pipeline that identifies the segments of the tracks that are informative of the network's prediction. We found that gradual accumulation of clathrin towards the peak clathrin signal, and quick fall after the peak both contribute to the prediction of an valid event. Our pipeline potentially obviates the need for the secondary marker (thus simplifying the experimental setup) and/or for manual annotation of events. More generally, the same framework can be used to predict the presence/absence of other molecular partners. All code and models are made openly available on github.[1]

1.2 Thrust II: Understanding the bias of MDI feature importance of random forest

Tree ensembles such as Random Forests (RF) [19] have achieved impressive empirical success across a wide variety of applications. To understand how RF models make predictions, people routinely turn to feature importance measures calculated from tree ensembles. One of the most widely used measures of feature importance is the Mean Decrease Impurity (MDI). MDI computes the total reduction in loss or impurity contributed by all splits for a given feature. This method is computationally efficient and has been widely used in a variety of applications [107, 58]. However, theoretical analysis of MDI, especially finite sample analysis, has remained sparse in the literature [64]. It has long been known that MDI tends to incorrectly assigns high importance to noisy features, leading to systematic bias in feature selection. In Chapter 4, we address the feature selection bias of MDI from both theoretical and methodological perspectives. Based on the original definition of MDI by Breiman et al. [20] for a single tree, we derive a tight non-asymptotic bound on the expected bias of MDI importance of noisy features, showing that the bias of MDI importance is inversely proportional to the minimum leaf node size of random forest. As such, deep trees have higher (expected) feature selection bias than shallow ones. However, it is not clear how to reduce the bias of MDI using its existing analytical expression. We derive a new analytical expression for MDI, and based on this new expression, we are able to propose a new MDI feature importance measure using out-of-bag samples, called MDI-oob. For both the simulated data and a genomic

[1] All code for reproducing, using, and adapting the pipeline here is made available at github.com/Yu-Group/auxilin-prediction, along with data and pre-trained models.

ChIP dataset, MDI-oob achieves state-of-the-art performance in feature selection from Random Forests for both deep and shallow trees.

1.3 Thrust III: Towards a better understanding of popular high-dimensional testing techniques

Optimality of the max test for detecting sparse signals with Gaussian or heavier tail

A fundamental problem in high-dimensional testing is that of *global null testing*: testing whether the null holds simultaneously in all of n hypotheses. The max test, which uses the smallest of the n marginal p-values as its test statistic, enjoys widespread popularity for its simplicity and robustness. However, its theoretical performance relative to other tests, such as the higher criticism and Berk-Jones tests Donoho and Jin [35], has been called into question. In a nutshell, the higher criticism and Berk-Jones tests compare the largest deviation between the empirical distribution of marginal p-values and the theoretical distribution. In the Gaussian sequence version of the global testing problem, Donoho and Jin [35] discovered a so-called "weak, sparse" asymptotic regime in which the higher criticism and Berk-Jones tests achieve a better detection boundary than the max test when all of the nonzero signal strengths are identical.

We study a more general model in which the non-null means are drawn from a generic distribution, and show that the detection boundary for the max test is optimal in the "weak, sparse" regime, provided that the distribution's tail is no lighter than Gaussian. Further, we show theoretically and in simulation that the modified higher criticism of Donoho and Jin [35] can have very low power when the distribution of non-null means has a polynomial tail.

Demystifying fixed-X knockoff

Knockoff filter is a framework [9] originally proposed as a variable selection procedure controlling the FDR in the statistical linear model. A core strength of knockoff methods is their virtually limitless customizability, allowing an analyst to exploit machine learning algorithms and domain knowledge without threatening the method's robust finite-sample false discovery rate control guarantee. While several previous works have investigated regimes where specific implementations of knockoffs are provably powerful, negative results are more difficult to obtain for such a flexible method. In this work we recast the fixed-X knockoff filter for the Gaussian linear model as a conditional post-selection inference method that adds user-generated Gaussian noise to the ordinary least squares estimator $\hat{\beta}$ to obtain a "whitened" estimator $\tilde{\beta}$ with uncorrelated entries, and performs inference using $\text{sgn}(\tilde{\beta}_j)$ as the test statistic for $H_j : \beta_j = 0$. We prove equivalence between our whitening formulation and the more standard formulation based on negative control predictor variables, showing how the fixed-X knockoffs framework can be used for multiple testing on any problem with (asymptotically) multivariate Gaussian parameter estimates. Relying on this perspective, we obtain the first negative results that universally upper-bound the power of all fixed-X

knockoff methods, without regard to choices made by the analyst. Our results show roughly that, if the leading eigenvalues of $\text{Var}(\hat{\beta})$ are large with dense leading eigenvectors, then there is no way to whiten $\hat{\beta}$ without irreparably erasing nearly all of the signal, rendering $\text{sgn}(\tilde{\beta}_j)$ too uninformative for accurate inference. We give reasonable and easy-to-check conditions under which the true positive rate (TPR) for any fixed-X knockoff method must converge to zero even while the TPR of Bonferroni-corrected multiple testing tends to one, and we explore several examples illustrating this phenomenon.

Chapter 2

A stability-driven protocol for drug response interpretable prediction (staDRIP)

2.1 Introduction

A critical goal in precision medicine oncology revolves around predicting a patient's response to therapeutic drugs given the patient's unique molecular profile [104, 67]. Accurate personalized drug response predictions can immediately shed light on therapies that are likely to be ineffective or toxic and aid clinicians in deciding the most promising treatment for their patients [6]. Moreover, interpreting these drug response prediction models can help to improve recommendations of compounds and target genes to prioritize in future preclinical research [26].

While several community-wide, public efforts [11, 32] and many other works have made progress towards improving the predictive accuracy of drug response predictions, identifying the important disease signatures (i.e., proteins, genes, and other biomarkers) that drive the drug response prediction models has received less attention. To date, previous works have typically focused on feature selection within one specific model such as elastic nets [60, 11] and random forest [102]. However, because molecular profiling data is often heterogeneous, noisy, and high-dimensional, these results are highly sensitive to modeling decisions made by humans including the type of model, the amount of training data, and the choice of algorithm.

In this work, we focus on this goal of detecting stable, interpretable, and predictive -omic signatures that drive a cell line's drug response. To overcome the aforementioned challenges, we develop a transparent stability-driven pipeline for drug response interpretable prediction called staDRIP that is rooted in the PCS framework for veridical data science [135]. At its core, the PCS framework builds its foundation on three principles: *predictability* as a reality check, *computability* as an important consideration in algorithmic design and data collection, and *stability* as an overarching principle and minimal requirement for scientific knowledge extraction. These principles were motivated by extensive interdisciplinary research such as Wu et al. [132], which analyzed the gap-gene network of *Drosophila*, and Basu et al. [12], which discovered stable transcription factor interactions in *Drosophila* embryos. Since its conception, the PCS framework has

further demonstrated a strong track record of driving many scientific discoveries including novel gene-gene interactions for the red-hair phenotype [15] and clinically-interpretable subgroups in a randomized drug trial [38].

Here, using integrative -omics and drug response data from the Cancer Cell Line Encyclopedia (CCLE) [11], we employ the PCS framework to develop staDRIP and provide extensive documentation of our modeling choices to arrive at stable biological discoveries of proteins and genes that are predictive of cancer drug responses. Unlike previous works whose results depend heavily on human decisions, staDRIP finds predictive -omic features that are stable across various models and data perturbations, thus mitigating the impact of human judgment calls. We further show that 18 of the top 24 -omic features identified by staDRIP have been previously implicated in the scientific literature, and in doing so, hint at novel candidates for future preclinical research.

2.2 The Cancer Cell Line Encyclopedia (CCLE) dataset

To begin building the personalized drug response models, we leverage data from a panel of 397 human cancer cell lines that have both high-throughput molecular profiling and pharmacological data for 24 anticancer drugs from the Cancer Cell Line Encyclopedia (CCLE) project [11]. Specifically, -omics data from the CCLE was downloaded from DepMap Public 18Q3 [1]. These cell lines encompass 23 different tumor sites and have been profiled for gene expression, microRNA expression, DNA methylation, and protein expression. Note that though the CCLE contains data from 947 cell lines, only 397 of these cell lines had data from all four molecular profiles of interest and pharmacological profiling.

In addition to the molecular profiles, we obtained pharmacological profiling of 24 chemotherapy and target therapy drugs from the CCLE [11]. For each cell line-drug combination, the CCLE incorporated a systematic framework to measure molecular correlates of pharmacological sensitivity in vitro across eight dosages. We refer to Barretina et al. [11] for details on this procedure, but given the fitted dose-response curves of growth inhibition from these experiments, we took the activity area, or AUC, to be the primary response of interest in this work. The AUC is defined as the area between the response curve and 0 (i.e., the no response reference level) and is a well-accepted measure of drug sensitivity [60, 11]. In this case, the AUC is measured on an 8-point scale with 0 corresponding to an inactive compound and 8 corresponding to a compound with 100% inhibition at all 8 dosages.

In Figure 2.1, we provide a graphical summary of the raw molecular and pharmacological profiling data sets.

Data preprocessing

Given the raw data described above, there are a couple areas of initial concern that warrant preprocessing. First, the cancer cell lines encompass 23 different tumor sites, and cell lines from the same tumor site tend to have more similar expression profiles than cell lines from different sites.

[1] https://depmap.org/portal/download/

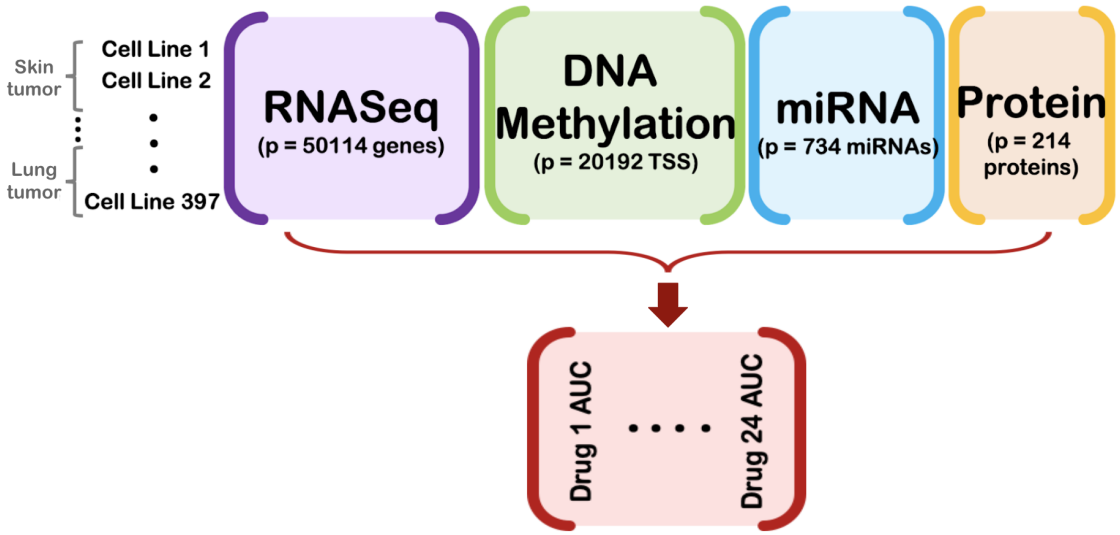


Figure 2.1: A graphical overview of the raw CCLE molecular profiling data sets, which are used to predict the drug responses of 24 therapeutic drugs, as measured via the drug response AUC.

See Table 2.1 for the frequency of cell lines from each tumor site. To illustrate, we observe clusters of cell lines by tumor site when performing both hierarchical clustering and PCA on the RNASeq profile in Figure 2.2. Due to these inherent differences between tumors, we chose to omit the cell lines from tumor sites with < 8 cell lines. This reduces our sample size to 370 cell lines from 16 tumor sites. Here, we chose the threshold 8 to ensure we have at least 2 cell lines from each tumor site in each of the training, validation, and test splits (using a 50-25-25% partitioning scheme).

In addition to reducing the number of samples in our analysis, we reduced the number of features to more manageable sizes before continuing with our analyses. Originally, the molecular profiling data consisted of 734 miRNAs, 50114 genes, 20192 TSS, and 214 proteins. With only 370 cell lines, we aggressively preprocessed the number of genes and TSS by taking the top 10% of genes (or 5000 genes) and top 20% of TSS (or 4000 TSS) with the highest variance. We also transformed the miRNA and RNASeq expression values using the log-transformation log(x+1) in order to mitigate potential problems with highly skewed positive count values.

We recognize however that there were many other reasonable ways to preprocess this data. For instance, we could have taken the top 20% of genes and top 40% of TSS with the highest variance. Another common alternative would have been to filter features using marginal correlations with the response or using a multivariate prediction model (e.g., the Lasso). To assess robustness to these choices, we reran our prediction analysis using these alternative preprocessing procedures and saw that the prediction accuracies are higher using the variance-filtering preprocessing pipelines, as compared to the correlation-filtering and Lasso-filtering pipelines (see PCS documentation). Further, the smaller variance-filtered model gives similar prediction accuracies as the larger variance-filtered model. Thus, for simplicity moving forward, we use and focus primarily on the initially proposed variance-filtering procedure as it is less computationally expensive

Table 2.1: Frequency of cell lines from each tumor site

Tumor Site	**# of Cell Lines**
Lung	72
Haematopoietic and lymphoid tissue	58
Skin	36
Breast	26
Central nervous system	23
Ovary	22
Large intestine	21
Pancreas	20
Endometrium	15
Stomach	14
Oesophagus	13
Liver	12
Urinary tract	12
Autonomic ganglia	9
Soft tissue	9
Bone	8
Kidney	7
Upper aerodigestive tract	6
Thyroid	5
Pleura	4
Prostate	3
Biliary tract	1
Salivary gland	1

than the model with twice as many features and maintains similarly high accuracy.

To summarize, after this preprocessing, we have 370 cell lines with data across the four molecular profiles of interest with 734 miRNAs (log-transformed), 5000 genes (log-transformed), 4000 TSS, and 214 proteins and pharmacological data, measured via the AUC drug response scores, for 24 anticancer compounds. We provide a visual summary of the preprocessed data and plot the overall distribution of features in the four molecular profiles as well as the distribution of the 24 drug responses in Figure 2.3.

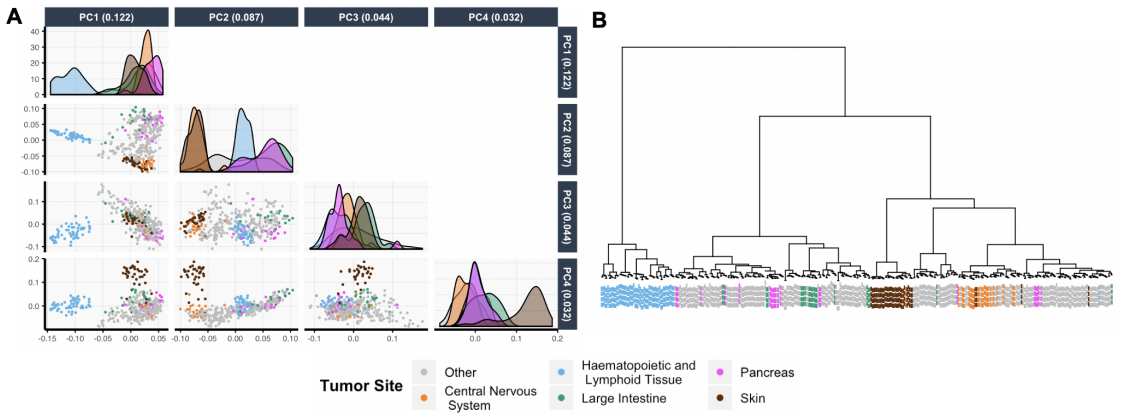


Figure 2.2: We apply (A) PCA and (B) hierarchical clustering (with Ward's linkage) to the log-transformed RNASeq data set and color the samples by their tumor site. For simplicity, we use color to distinguish between five prominent tumor sites and show the remaining tumor sites in grey. We also show the proportion of variance explained by each principal component in the subplot titles of (A). In both the PC plots and the hierarchical clustering dendrogram, we can see clusters of tumor sites, illustrating the inherent differences between tumor sites.

2.3 Prediction methods and evaluation metrics

Prediction Models

Like in data preprocessing, human judgment calls play a significant role in the modeling stage, including the decision of which methods to fit. Ideally, the chosen methods should have some justified connection to the biological problem at hand, but in our case, it is unclear which models or assumptions best fit the biological drug response mechanism a priori. Nevertheless, we have reasons to believe that the Lasso, elastic net, RF, and kernel ridge regression are particularly appealing fits for this problem.

First, the Lasso assumes a sparse linear model, meaning that the effect of each feature is additive and only a sparse number of the features contribute to the drug sensitivity. The simplicity and interpretability of the Lasso makes it a popular tool for bioinformatics prediction tasks, so we choose to use the Lasso as a baseline model for our analysis. The elastic net is perhaps even more popular than the Lasso in drug response prediction studies [60, 11]. Similar to the Lasso, the elastic net assumes linearity and some sparsity but is also able to better handle correlated features. Beyond linearity, kernel ridge regression with a Gaussian kernel allows for more flexible, but less interpretable, functional relationships that are not necessarily linear. Kernel methods have been applied in previous case studies with great success [32] and are hence promising candidates for our study as well. Lastly, random forest can be viewed as a collection of localized, non-linear thresholded decisions rules (like on-off switches), which are well-suited for many biological processes that match the combinatorial thresholding (or switch-like) behavior of decision trees [93]. Random forests are also invariant to the scale of the data. This is especially advantageous for integrating

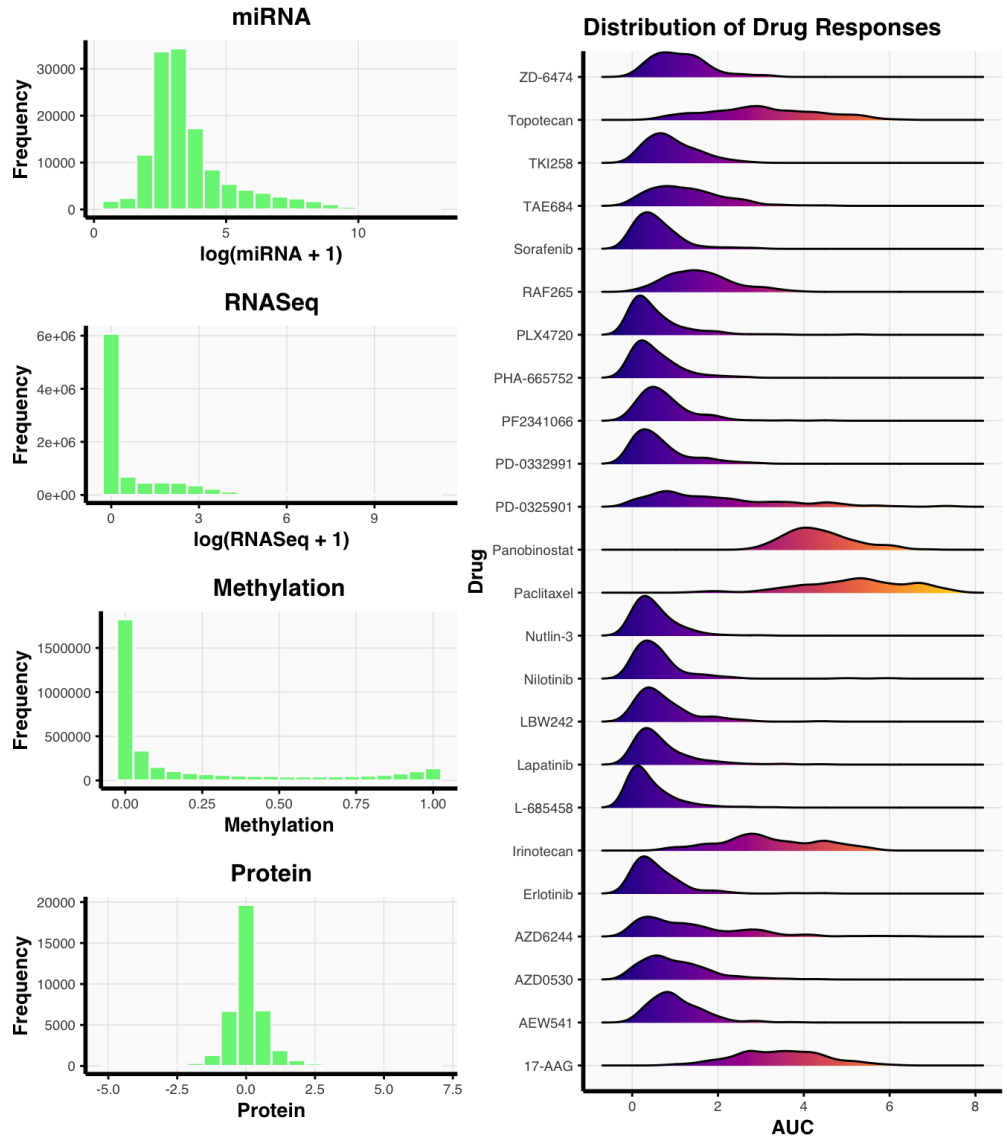


Figure 2.3: Left: Distribution of features in each of the four molecular profiles. Right: Distribution of the drug responses for each of the 24 drugs.

different data sets with varying scales and domain types (e.g., count-valued RNASeq expression, proportion-valued methylation data, continuous-valued protein expression).

In addition to fitting the aforementioned methods on each of the molecular profiles separately, we also tried fitting various data integration methods since incorporating multiple sources of -omics data can sometimes result in more accurate predictions than models built using only a single -omics sources [32, 48, 113]. The most natural integration idea is to concatenate the -omics data sets together and to fit a single model (e.g., the elastic net) on the concatenated data. When fitting models like the Lasso, elastic net, and kernel ridge regression which are not scale-invariant, the

molecular profiles are scaled to have columns with mean 0 and variance 1 to allow for fair comparisons between molecular types. We refer to this method as the concatenated data approach and use this as a baseline for evaluating data integration methods. More sophisticated methodology has also been proposed to integrate -omics data, including recent work using the X-VAE, a variational autoencoder for cancer data integration [113], and the BMTMKL, a Bayesian multitask multiple kernel learning method which won the NCI-DREAM 7 challenge [32].

Note that though an alternative approach would have been to develop new methodology, we instead leverage these existing machine learning methods that have been rigorously vetted and have been shown to work well in many related problems. In fact, by examining the stable properties across these existing methods, we obtain high-quality scientific findings, as made evident by the abundance of supporting literature (see Table 2.5).

Model hyperparameters

To select hyperparameters in each of these methods, we use 5-fold cross validation, where the folds are stratified by tumor type. We also investigate using the estimation stability cross validation (ESCV) metric for selecting the Lasso's hyperparameter. This ESCV metric combines a stability measure within the cross-validation framework to yield more stable estimation properties with minimal loss of accuracy when using the Lasso [74].

For the X-VAE model, we adapt an X-shaped network architecture to train a variation autoencoder that learns joint representation of the RNAseq and protein data. In particular, we take the 2,000 RNAseq features with highest variance, since the number of cell lines is too small compared with the original number of RNAseq features. In our experiment, both the encoder and the encoder have one hidden layer. There are 128 neurons corresponding to the RNAseq protein in the hidden layer of the encoder and the decoder, and 32 neurons corresponding to the protein features. The latent representation has a dimension of 32. The dimension of the hidden layers and the latent representation are based on the recommendation of [113], and are not tuned. We used ELU activation and employed batch normalization and a dropout component with rate 0.2, as recommended by [113]. The models were trained for 500 epochs using an Adam optimizer with a learning rate of 0.001.

Evaluation metrics

We primarily consider two evaluation metrics for prediction accuracy as each captures a different aspect of prediction - 1) R^2 value and 2) probabilistic concordance-index (PC-index). R^2 is defined as $1 - \frac{\text{MSE}(Y,\hat{Y})}{\text{Var}(Y)}$, where $\text{Var}(Y)$ denotes the variance of the observed responses, and $\text{MSE}(Y, \hat{Y})$ denotes the mean sum of squared errors between the predicted responses $\hat{Y}$ and observed responses Y. R^2 is a rescaling of the MSE that accounts for the amount of variation in the observed response and thus allows us to easily compare accuracies between drug response models with different amounts of variation in the observed response, but as with the MSE, R^2 can be heavily influenced by outliers. PC-index is a measure of how well the predicted rankings agree with the true responses. This metric takes into account the variance of the drug responses but it also assumes that the drug

responses follow a Gaussian distribution, which may not be true in some cases. We consider this metric because it is the primary method of evaluation in the NCI-DREAM 7 competition [32]. Given the large scale and breadth of this challenge, we compare our results to this work. For further details on the PC-index, we refer to Costello et al. [32].

In each of the evaluation metrics above, we receive a separate score for each of the 24 drug response models. It may also be beneficial to aggregate the 24 scores into a single number for concrete evaluation. In particular, Costello et al. [32] used a weighted average of the PC-indices to compare various models and referred to this evaluation metric as the weighted PC-index (WPC-index). To compare our results with the benchmark in Costello et al. [32], we also consider the WPC-index in evaluating our models.

2.4 Results on Prediction Accuracy

Building on the PCS framework, staDRIP first uses predictive accuracy as a reality check to filter out models that are poor fits for the observed data before turning to our primary goal of identifying important biomarkers for drug response prediction. For each of the available 24 anticancer drugs, we divide the data into a 50-25-25% training-validation-test split and use the training data to fit (1) an elastic net tuned with cross-validation (CV), which has been widely used and advocated by previous studies [11, 60], (2) Lasso tuned with CV, (3) Lasso tuned with ESCV, an alternative CV metric that incorporates stability to yield more stable estimation properties with minimal loss of accuracy [74], (4) Gaussian kernel ridge regression, and (5) random forest to predict the drug response given the miRNA, RNASeq, methylation, and protein expression profiles separately. We also fit several data integration methods including concatenated versions of the aforementioned methods, the recently proposed X-shaped Variational Autoencoder (X-VAE) [113], and the winner of the DREAM 7 challenge,[2] the Bayesian multi-task multiple kernel learning method (BMTMKL) [32].

For each of these fits, we report in Table 2.2 the average validation accuracy across all 24 drugs as measured by the R^2 value and the WPC-index, a weighted probabilistic concordance index, which has been used in previous studies and measures how well the predicted rankings agree with the true responses [32]. From Table 2.2, we see that kernel ridge regression trained only on the RNASeq data yields the best predictive performance. However, considering that our primary goal is not purely prediction, the differences between model prediction accuracies shown in Table 2.2 are relatively small from a practical viewpoint. In our inferential procedure discussed next, we will see that leveraging the stability across these methods with similar predictive accuracies is key to our staDRIP pipeline for identifying genes and proteins that are stable predictive features underlying the drug response models.

Nonetheless, for completeness, we report the test accuracy from the best model, the RNASeq-based kernel ridge regression, to have an R^2 ($\pm$1SD) of 0.204 ($\pm$ 0.038) and WPC-index of 0.620 ($\pm$ 0.0075) across the 24 drugs.

[2]The DREAM 7 challenge was a public competition where teams were tasked to integrate multiple –omics measurements and predict drug sensitivity in cancer cell lines.

In Tables 2.3 and 2.4, we provide additional insights into the drug response prediction accuracies at the individual drug level. In Table 2.3, we see that the best model depends on the particular drug, but the kernel ridge regression model works best on average. In Table 2.4, we show the test errors from the kernel ridge regression fit for each drug separately.

Table 2.2: Validation WPC-index and average R^2 across all 24 drug response models for various methods trained on each molecular profile separately and together. Higher values of R^2 and WPC-index indicate better fits.

	Validation Set WPC-Index					**Validation Set R^2**				
	Methyl.	miRNA	Protein	RNASeq	Integrated	Methyl.	miRNA	Protein	RNASeq	Integrated
Kernel Ridge	0.600	0.603	0.617	**0.631**	0.624	0.111	0.104	0.168	**0.231**	0.200
Elastic Net	0.602	0.606	0.608	0.626	0.625	0.102	0.124	0.126	0.183	0.162
Lasso	0.597	0.605	0.609	0.620	0.620	0.117	0.105	0.121	0.172	0.176
Lasso (ESCV)	0.600	0.601	0.609	0.623	0.618	0.114	0.113	0.129	0.195	0.141
RF	0.599	0.594	0.606	0.626	0.622	0.124	0.088	0.123	0.214	0.196
X-VAE	–	–	–	–	0.617	–	–	–	–	0.188
BMTMKL	–	–	–	–	0.613	–	–	–	–	0.179

Table 2.3: For each molecular profile (or the integrated profile) used for training, we count the number of drugs (out of 24) for which each method performed the best and gave the highest validation R^2 compared to its six other competitors.

	Methyl.	miRNA	Protein	RNASeq	Integrated
Kernel Ridge	7	5	16	12	6
RF	9	5	2	6	5
Elastic Net	1	8	1	1	2
Lasso (ESCV)	4	4	3	4	0
Lasso	3	2	2	1	5
X-VAE	–	–	–	–	2
BMTMKL	–	–	–	–	2

2.5 Identifying predictive -omic features with PCS inference

Beyond predictability, the PCS framework emphasizes stability throughout the data science life cycle so as to reduce reliance on particular human judgment calls. Accordingly, we leverage and quantify the stability of important features under numerous data and model perturbations in staDRIP as follows: for each of the 24 drugs separately,

1. **Use predictability as reality check:** select a set $\mathcal{M}$ of models with high predictive accuracy across a variety of metrics on the validation data.

Table 2.4: Test error for each drug using the RNASeq-based kernel ridge regression model

Drug	R^2	**PC-Index**
17-AAG	0.000	0.574
AEW541	0.034	0.558
AZD0530	0.037	0.560
AZD6244	0.425	0.675
Erlotinib	0.254	0.615
Irinotecan	0.307	0.644
L-685458	0.210	0.624
LBW242	-0.001	0.511
Lapatinib	0.208	0.607
Nilotinib	0.258	0.590
Nutlin-3	0.022	0.549
PD-0325901	0.543	0.701
PD-0332991	0.218	0.596
PF2341066	0.091	0.564
PHA-665752	0.115	0.559
PLX4720	0.305	0.585
Paclitaxel	0.369	0.670
Panobinostat	0.446	0.679
RAF265	0.215	0.625
Sorafenib	0.242	0.567
TAE684	0.024	0.576
TKI258	0.183	0.585
Topotecan	0.240	0.630
ZD-6474	0.155	0.591

2. **Compute stability of predictive features across data perturbations:** for each model $M \in \mathcal{M}$, refit the model M to B bootstrap replicates of the data, and compute the stability score of each feature as the proportion of B bootstrap samples where the feature is selected. Let F_M denote the subset of features with high stability scores (e.g., top 10).
3. **Select predictive features that are stable across model perturbations:** take the intersection $\cap_{M \in \mathcal{M}} F_M$ as the stable predictive -omic features across data and model perturbations.

In our work, we are primarily interested in identifying proteins and genes that are predictive of drug responses as many drugs are directly related to known proteins and genes. Hence, considering the five models trained on the RNAseq and protein data separately, we take $\mathcal{M} = \{\text{RF}, \text{Lasso (ESCV)}, \text{Elastic Net}\}$. Note that while kernel ridge has the highest accuracy, it is omitted from $\mathcal{M}$ since there is no straightforward, computationally efficient method to select features from kernel ridge to the best of our knowledge. We also omit the Lasso from $\mathcal{M}$ as it generally has the worst predictive accuracy. For each remaining model in $\mathcal{M}$, we then take F_M to be the

Table 2.5: Most stable protein associated with each drug, as identified by staDRIP, along with literature that supports the association between the protein and drug sensitivity.

Drug	Protein	Supporting Literature	Drug	Protein	Supporting Literature
17-AAG	Bax	He et al. [51]	PD-0332991	Bcl-2	Chen and Pan [30]
AEW541	Akt	Attias-Geva et al. [5]	PF2341066	c-Met	Camidge et al. [25]
AZD0530	p38	Yang et al. [133]	PHA-665752	MEK1	–
AZD6244	PI3K-p85	Balmanno et al. [7]	PLX4720	MEK1	Emery et al. [40]
Erlotinib	EGFR	McDermott et al. [85]	Paclitaxel	Src	Le and Bast [70]
Irinotecan	MDMX_MDM4	Ling et al. [75]	Panobinostat	VEGFR2	Strickler et al. [116]
L-685458	YAP	–	RAF265	PI3K-p85	Mordant et al. [87]
LBW242	ASNS	–	Sorafenib	Bcl-2	Tutusaus et al. [124]
Lapatinib	HER2	Esteva et al. [41]	TAE684	PTEN	–
Nilotinib	STAT5	Warsch et al. [127]	TKI258	CD49b	–
Nutlin.3	Bcl-2	Drakos et al. [37]	Topotecan	–	–
PD-0325901	MEK1	Henderson et al. [53]	ZD-6474	c-Kit	Yang et al. [134]

10 features with the highest stability scores [3] and list those genes and proteins in the top 10 most stable features across all three models in Table 2.6.

In Table 2.5, we provide our main evidence for the utility of staDRIP, listing the single most stable protein for each drug along with independent publications that support these findings. Specifically, of the 24 proteins identified as most stable by staDRIP, 18 have been associated with the drug sensitivity or identified as a known or possible drug target in prior preclinical studies.

Now in contrast to staDRIP, which finds stable predictive features across models with similar predictive accuracies, previous state-of-the-art methods [11, 60] use only an elastic net to identify predictive -omics features of drug responses. To compare staDRIP to this elastic net approach, we extract the proteins with the highest stability score for each drug when taking $\mathcal{M} = \{\text{Elastic Net}\}$. Repeating the same literature search procedure as we did for the proteins identified by staDRIP, we found only 14 of the 24 proteins identified by the elastic net are known from previous clinical studies (see Table 2.7). We now discuss in detail the disease signatures identified by staDRIP, and compare them with those identified by elastic net.

Discussion on the disease signatures identified by the staDRIP pipeline

In Table 2.6, we list the proteins and genes which we found to be stable and among the top 10 features for all three methods. Among these stable features, we list them in decreasing order by the sum of stability score rankings. Though we identify fewer stable genes, this is most likely due to two reasons. First, there are 5000 genes in the model, compared to only 214 proteins, so thresholding at the top 10 genes is extremely conservative. Secondly, the average correlation between genes is higher than that between proteins, adding to the instability.

[3] The stability scores are computed as follows. For each feature X_j from either the protein or RNAseq data set, let $\omega_j^{(b)}$ be defined in the following way: for the Lasso and elastic net, $\omega_j^{(b)} = 1$ if the coefficient of X_j is non-zero, and $\omega_j^{(b)} = 0$ otherwise; for the random forest, $\omega_j^{(b)}$ is the MDI feature importance of X_j. We then define the stability score $\text{sta}(X_j)$ of each feature X_j as $\text{sta}(X_j) = \frac{1}{B}\sum_{b=1}^{B} \omega_j^{(b)}$

Table 2.6: Stable protein and RNAseq signatures. A feature is included if it is among the top 10 most stable features under 3 different machine learning models (i.e., elastic net, Lasso (ESCV), and random forests). The stability of the features are computed from the PCS inference framework in staDRIP. Blank cells indicate that no features appeared among the top 10 most stable features for all three models.

Drug name	Protein signature	RNAseq Signature
17-AAG	Bax, p53, Caspase-7, eIF4E	CTD, AP2S1, BZW2
AEW541	Akt, Smad1, p27, PTEN, RAD51	B4GALT3, SEMA3B
AZD0530	p38, c-Kit	HPGD
AZD6244	PI3K-p85, TFRC, Bax	SPRY2, RP11, LYZ, DUSP6, PRSS57
Erlotinib	EGFR, Beclin, P-Cadherin	PIP4K2C, SEC61G
Irinotecan	MDMX_MDM4, Src	
L-685458	YAP, VEGFR2, Src	
LBW242	ASNS	MRPL24
Lapatinib	HER2, HER3, EGFR, Rab25, Heregulin	STARD3
Nilotinib	STAT5, c-Kit, SHP-2, Src, p27	
Nutlin.3	Bcl-2, Bax	
PD-0325901	MEK1, Bax, TFRC, PI3k	SPRY2, DUSP6, ETV4
PD-0332991	Bcl-2, MDMX_MDM4, Src	
PF2341066	c-Met	CAPZA2
PHA-665752	MEK1, c-Met	FMNL1
PLX4720	MEK1, Bax, PREX1, Beclin	FABP7
Paclitaxel	Src, beta-Catenin	ORMDL2
Panobinostat	VEGFR2, Src	
RAF265	PI3K-p85, FOXO3a, eEF2K	RETN
Sorafenib	Bcl-2, Src	
TAE684	PTEN, Akt, p70S6K, Bcl-2	H1FX
TKI258	CD49b, C-Raf	
Topotecan	c-Met	OSGIN1
ZD.6474	c-Kit, STAT5-alpha	

With regards to the identified protein signatures, we can roughly classify them into three categories. The first category contains those that are known targets of the corresponding target therapy drugs. For example, Erlotinib is a medication used to treat non-small cell lung cancer (NSCLC) and pancreatic cancer. It is an EGFR inhibitor and is specifically used for NSCLC patients with tumors positive for EGFR exon 19 deletions (del19) or exon 21 (L858R) substitution mutations. Correspondingly, EGFR is ranked in the top ten stable proteins in all three models. Other such examples include the drug Lapatinib and its target HER2, PD-0325901 and its target MEK, PHA-665752 and its target c-Met, and ZD-6474 and its target c-Kit.

The second category contains those that are not known to be direct targets of the drug but have been shown in preclinical studies to be potential therapeutic targets or are associated with drug resistance. For example, Ling et al. [75] identified a potential application of the drug Irinotecan as an MdmX inhibitor for targeted therapies, and in our pipeline, MdmX had the highest stability score for all three models. As another instance, we identified MEK1 as a top protein signature, ranked by stability score, for the drug PLX4720 while Emery et al. [40] showed that MEK1 mutations confer resistance to PLX4720.

The third category are proteins that do not belong to the two categories above. Still, these

Table 2.7: Most stable protein associated with each drug, as identified by the elastic net, along with preclinical evidence that supports the association between the listed protein and drug sensitivity.

Drug	Protein	Supporting Literature	Drug	Protein	Supporting Literature
17-AAG	p53	Naito et al. [92]	PD-0332991	Bcl-xL	Chen and Pan [30]
AEW541	Akt	Attias-Geva et al. [5]	PF2341066	PEA15	–
AZD0530	p38	Yang et al. [133]	PHA-665752	MEK1	–
AZD6244	CD20	–	PLX4720	MEK1	Emery et al. [40]
Erlotinib	P-Cadherin	–	Paclitaxel	Src	Le and Bast [70]
Irinotecan	RAD51	Shao et al. [111]	Panobinostat	Src	–
L-685458	VEGFR2	–	RAF265	PI3K-p85	Mordant et al. [87]
LBW242	Caspase-7	–	Sorafenib	14-3-3 epsilon	Wu et al. [131]
Lapatinib	HER2	Esteva et al. [41]	TAE684	Akt	–
Nilotinib	p27	Liu et al. [77]	TKI258	14-3-3 epsilon	–
Nutlin.3	Bcl-2	Drakos et al. [37]	Topotecan	14-3-3 epsilon	–
PD-0325901	MEK1	Henderson et al. [53]	ZD-6474	c-Kit	Nishioka et al. [96]

biomarkers are predictive of the drug response under various model and data perturbations. Given the evidence in the scientific literature that supports many of our identified features, the proteins in this category may be potential candidates for future preclinical investigation.

Among the list of overlapping stable features in Table 2.6, we list in Table 2.5 the one with the highest stability score ranking along with recent biomedical publications, supporting the association between the protein and the drug. The procedure of this literature search is as follows: we first searched for papers where the protein and the drug co-occurs. Then for each paper, we read the introduction section to understand their main conclusions. Each of the 18 papers listed in Table 2.5 includes sentences such as "our findings suggest that the over-expression of this protein will increase drug sensitivity/resistance" or "this protein is a potential (or known) therapeutic target for the drug". Out of the 24 predictive protein signatures that we identify as most stable, 18 of them have existing preclinical studies that confirm the effectiveness of our stability analysis.

In Table 2.7, we list the protein with the highest stability score when fitting an elastic net to 100 bootstrap samples of the training data for each of the 24 drugs. This approach of finding predictive -omics features was previously used in [11, 60]. Compared with staDRIP, which searches for stable features across different models, this approach only uses a single model (i.e, an elastic net) for feature selection. We repeat the same literature search procedure as we did for our findings and found that among the 24 most stable protein features identified by elastic net, only 14 are known from previous clinical studies. For 10 drugs, the most stable protein from elastic net and that from staDRIP is the same, and among these 10 proteins, 9 are implicated in the existing literature. For the other 14 drugs, 5 proteins identified by elastic net are implicated in the existing literature, while 9 protein features identified by staDRIP are implicated in the existing literature.

2.6 Conclusion

Rooted by the PCS framework, we emphasize the importance of predictability, (computability), and stability as minimum requirements for extracting scientific knowledge throughout the staDRIP

pipeline. We show that, guided by good prediction performance, incorporating a number of stability checks and extracting the stable parts of top-performing models can help to avoid the poor generalization exhibited by existing methods and can successfully identify candidate therapeutic targets for future preclinical research. We also acknowledge that while many stability considerations are built into staDRIP, there are inevitably human judgment calls that still impact our analysis. For example, we make a number of judgement calls in the data preprocessing stage. Additionally, many other reasonable models such as ridge regression and gradient boosting could be considered in the staDRIP pipeline. We thus provide transparent and extensive documentation here to justify these decisions using domain knowledge when possible.

Chapter 3

Interpretable deep learning for accurate molecular partner prediction in clathrin-mediated endocytosis

3.1 Introduction

The interaction between molecular partners, such as proteins which cooperate in a biological pathway, is fundamental to many biological processes. In studying these interactions, computational tools such as machine learning algorithms can complement wet-lab biochemical experiments, due to their ability to make fast and accurate predictions. In this paper, we propose a pipeline which uses a deep neural network to accurately predict the activity of one protein from its molecular partner, and test our pipeline extensively in the context of clathrin-mediated endocytosis (CME).

Clathrin-mediated endocytosis is the process by which a cell absorbs metabolites, hormones, proteins, or other materials through its cell membrane using clathrin-coated vesicles. It is fundamental to neurotransmission, signal transduction and the regulation of many plasma membrane activities and is thus essential to higher eukaryotic life [86]. Understanding clathrin-mediated endocytosis (CME) is a crucial question in cell biology [66], and many questions about this process remain unanswered [63]. One major challenge with analyzing CME is the ability to readily distinguish between abortive coats (ACs) and valid clathrin-coated pits (CCPs). Doing so enables better understanding of the mechanisms governing CCP dynamics and progressions. Previous approaches have largely relied on relatively simple thresholds of the clathrin fluorescence images, based on lifetime and intensity [1, 62]. However, recent studies have suggested that these features alone may be insufficient to discern CCPs [52, 126].

Here, we resolve ACs from CCPs by predicting the recruitment of two molecular partners of clathrin, auxilin and dynamin, from fluorescence imaging of CCP activity. It is known that CCPs lose their clathrin lattice within seconds of pinching off, through the action of the Hsc70 “uncoating ATPase”, recruited by auxilin 1 (Aux1) and auxilin 2 (GAK). Aux1 and GAK appear on coated vesicles immediately after dynamin-mediated membrane scission has released the vesicle from the

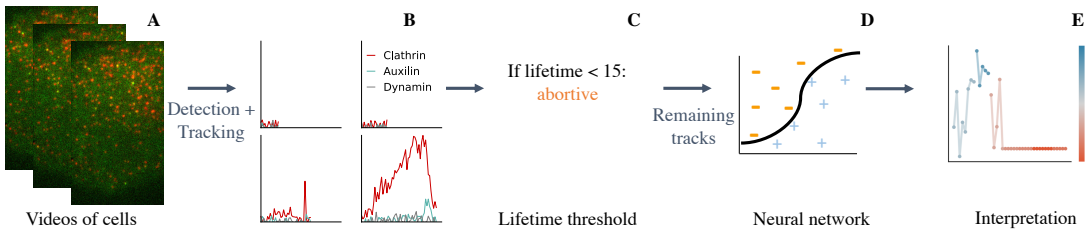


Figure 3.1: Pipeline for molecular partner prediction. **A** Cells tagged with markers for clathrin, auxilin, and dynamin are imaged. **B** Potential CME events are detected and tracked, yielding fitted amplitudes over the course of each event. **C** A data-selected threshold identifies many of the abortive events. **D** A neural network classifies the remaining events with longer lifetimes, which we refer to as the "difficult" region. **E** Interpretation techniques elucidate how the neural network makes its predictions.

.

plasma membrane [52]. As such, the detection of auxilin and dynamin can be used as validation of productive CCPs.

Our pipeline can be summarized in three steps. First, we use additional markers, such as auxilin 1 (AUX), auxilin 2 (GAK) and dynamin, to track the formation of CCPs. Using a combination of automatic rules and manual annotation, we curate a dataset of cells and whether their clathrin-coated structures (CCSs) are valid or abortive. Second, we develop a predictive pipeline (see Figure 3.1) to identify whether a CCS is valid, based only on imaging clathrin. In new data, this pipeline enables identifying CCPs without the need for additional markers, enabling simpler experimental setups. Third, we build an interpretation pipeline that aims at elucidating the elements in CCP progression used by the neural network in resolving ACs from CCPs.

The fitted model achieves a test accuracy of 92% on all events. Notably, our pipeline is accurate in predicting longer-lived events, where distinguishing ACs from CCPs is known to be more difficult. In particular, for events with lifetime greater than 15, our pipeline achieve an accuracy of 84.1%, which represents a 5% improvement over DASC [126], the current state-of-the-art technique. It also yields uncertainty measures which can be used to selectively predict only on high-confidence tracks. The fitted model is robust to reasonable perturbation of experiment conditions. For example, the accuracy on events with lifetime greater than 15 only reduces by 1% when the data sampling rate (Hz) is three times lower. We also show that the model transfers to new datasets collected at various settings.

In addition to developing an accurate neural network model for molecular partner prediction in CME, our paper also provides a comprehensive investigation of a wide ranging computational approaches on this problem. We examine the performance of an extensive list of pre-processing procedures, feature engineering techniques, classification algorithms, and deep learning architectures. We hope that this paper will help guide machine-learning applications in similar molecular-partner prediction problems in the future.

Related Work

One line of work aims to identify successful CCPs without the use of an additional marker. Initial attempts, before the maturation of detection/tracking used statistics to deconvolve and classify using lifetimes [78]. Later works [1, 62] used tracking + detection along with thresholds on lifetime and intensity. [56] trains a support vector machine to distinguish CCPs, although the features were still based mostly on lifetime/intensity thresholds. [126] use a thermodynamics-inspired method to resolve ACs from CCPs based on single channel fluorescent movies, which they term Disassembly asymmetry score classification (DASC). Alternatively, one can use the internalization of pH sensitive-cargo [122] to identify successful CCPs, although this requires more labeling and a more complicated experimental set up. A recent study introduces DEEPCLA [136], a method which combines convolutional neural networks and LSTMs to identify clathrin directly from videos.

However, clathrin alone is often insufficient to identify productive CCPs. More related to the work here are studies which use second markers to help identify abortive from productive events. Two studies [47, 39] study the recruitment of dynamin. Most related to the work here is one study, which analyzes the dynamics of auxilin 1 / GAK in clathrin-mediated traffic [52]. While previous works typically rely on a handful of manually engineered features or a particular prediction algorithm, our paper examines a wide array of engineered features and machine learning models. We further show that our proposed neural network model outperforms previous state-of-the-art method in identifying valid CCPs, and in doing so, demonstrate the capacity of machine learning in molecular partner prediction.

3.2 Results

Data and lifetimes

Figure 3.2 provides an introduction into the data collected here. In particular, Figure 3.2A shows 6 representative events, with varying lifetime and amplitude. More events are shown in Figure in the appendix. Figure 3.2B shows the raw videos of 2 events in Figure 3.2A, where the signals are circled out. Figure 3.2C shows the distribution of lifetime and max fitted amplitude for all the events in the dataset. While valid events tend to be longer-lived and have a higher max amplitude, it is also clear from this figure that using the lifetime and max fitted amplitude alone are not sufficient to separate abortive events from valid ones. This fact was noted by Wang et al. [126], who proposed a set of thermodynamics-inspired features of the tracks, known as the DASC. However, we find that various machine-learning models trained on the DASC features do not perform favorably to neural network models trained directly on the tracks, which we introduce in the sequel.

The data is collected in a variety of different conditions. Table 3.1 shows the datasets collected in a handful of various settings. In particular, the bolded dataset “clath_aux_dynamin” contains all 3 markers and has the most number of events, and will be used as the main dataset for training and evaluation. The other 5 supporting datasets only have clathrin and auxilin markers. ”Short” events refers to those with lifetime smaller than 15 (excluding the buffer frames). We find that over 90% of these events are abortive, consistent with previous literature [1].

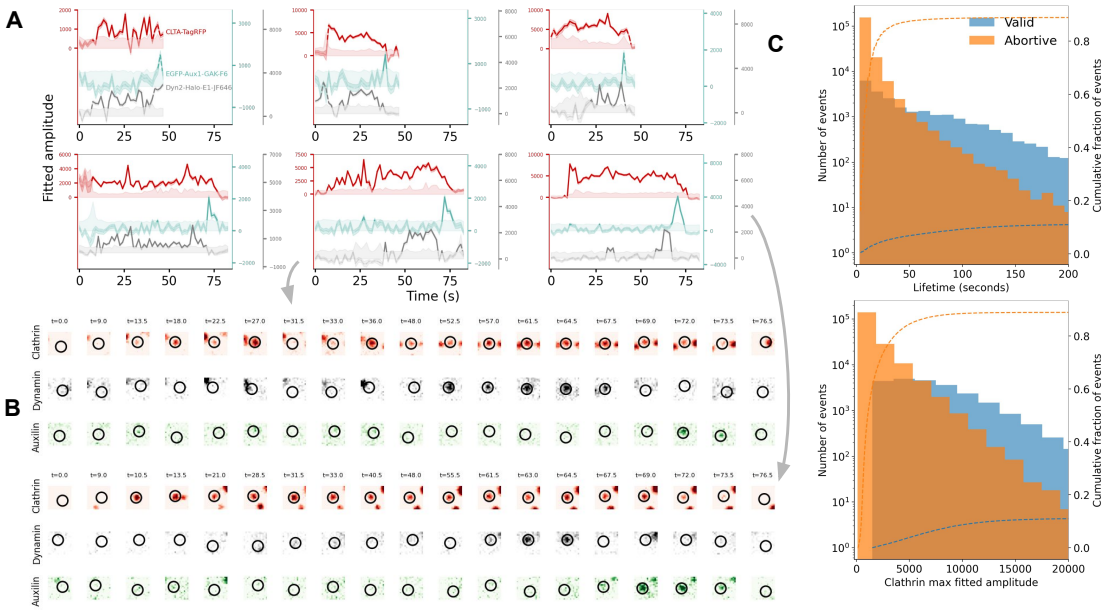


Figure 3.2: CME event examples and distribution. **A** Examples of amplitude traces for events that vary in lifetime and amplitude. Dark shaded areas around the trace are estimated uncertainties for detected intensity (s.d.), and light shaded areas are significance threshold above background (2 s.d.). Dashed lines show the 5 frames buffers before and after the tracked event. **B** Cropped frames from raw videos for two different events. For chosen time points in the course of the event, we plot a 15-by-15 square around the detected position of the event. Darker color indicates stronger signal. **C** Distribution of events based on lifetime and clathrin max amplitude: valid events tend to be longer and have a higher max amplitude. Total number of events is 210,587.

Dataset	markers	Partition	Total	Difficult	Difficult valid	Short	Short valid
clath_aux_dynamin	CLTA-TagRFP EGFP-Aux1-GAK-F6 Dyn2-Halo-E1-JF646	test	45061	8457	2920	36604	2205
		train	102712	20617	7050	82095	4977
clath_aux+gak	CLTA-TagRFP; EGFP-GAK-F6	test	1551	494	264	1057	96
		train	5783	1750	1054	4033	464
clath_aux+gak_a7d2	CLTA-TagRFP; EGFP-Aux1-A7D2 EGFP-GAK-F6	test	2448	1018	439	1430	73
		train	7245	2979	1523	4266	212
clath_aux+gak_a7d2_new	CLTA-TagRFP EGFP-Aux1-A7D2 EGFP-GAK-F6	test	16893	1495	978	15398	2157
		train	49743	7029	4900	42714	9591
clath_aux+gak_new	CLTA-TagRFP EGFP-Aux1 EGFP-GAK-F6	test	4040	633	211	3407	183
		train	16250	2922	1515	13328	1548
clath_gak	CLTA-TagRFP EGFP-GAK-A8	test	5189	1319	628	3870	378
		train	12938	3165	1288	9773	749

Table 3.1: Data summary. "Difficult" refers to events with lifetime > 15, whereas "Short" refers to tracks with lifetime ≤ 15.

Predictive modeling

The modeling pipeline begins with a threshold which classifies all events with lifetime below 15 as abortive (this threshold is chosen to maximize the total accuracy across the entire dataset). For the points with lifetimes above this threshold, we construct a model to predict whether events will be abortive or valid using the processed data. The best-performing model is a long short-term memory neural network network [54] which takes 40-dimensional input and has a hidden-state size of 20.

We also extensively consider an alternative approach which engineers predictive features from the tracks using domain knowledge and data-driven feature extraction. For example, the lifetime, the maximum clathrin intensity, derivatives of the track, information about the local maxima and minima, and the maximum rise and fall of the track. In addition, we add in a couple features about the motion of the pit (e.g. tracking the mean squared distance traveled by the center of the tracked pit, and the final distance traveled). We also consider many more engineered features, such as those obtained by coding the tracks using sparse coding or non-negative matrix factorization, but we omit these as they do not improve performance. We also try a variety of modeling techniques including linear models, random forests, multilayer perceptrons, convolutional neural networks, and support vector machines.

Figure 3.3 shows predictive results for our LSTM model and two other models on the difficult tracks, for cells imaged under 6 different conditions. We note that our LSTM in general has substantially higher prediction accuracy than the baseline DASC model, and also outperforms the gradient boosting model with hand-engineered features. This pattern is stable across different cells. In addition, the LSTM model has higher accuracy on cells with dynamin markers than on other cells. This is expected since there are more cells imaged under the same condition as the cells with dynamin markers in the training data. Detailed classification and regression accuracy of different methods are given in Table 3.2 in the appendix. These results confirm that the LSTM is the best-performing model across various metrics.

As the size of the training data gets smaller, the gradient-boosting model starts to outperform the LSTM model (Figure 3.3D).

We also conduct a series of analyses suggesting that the predictive performance here is limited by noise in the data (see Sec 3.5). shows closest matches using dynamic time warping [68, 106], with the constraint that the lifetimes are within seven frames of one another

We find that the LSTM model is stable to various judgment calls made during modeling. We downsample the tracks such that all the tracks have length 20, 40 and 100. In addition, we also tried padding zeros at the beginning instead of padding the zeros at the end, if the length of the original track is shorter than the targeted length. These perturbations lead to 6 different models. Fig 3.9 in the appendix shows the classification accuracy of these 6 models on different datasets. We find that although downsampling the tracks to length 20 leads to slightly degraded prediction performance, the LSTM model is quite stable to other perturbations.

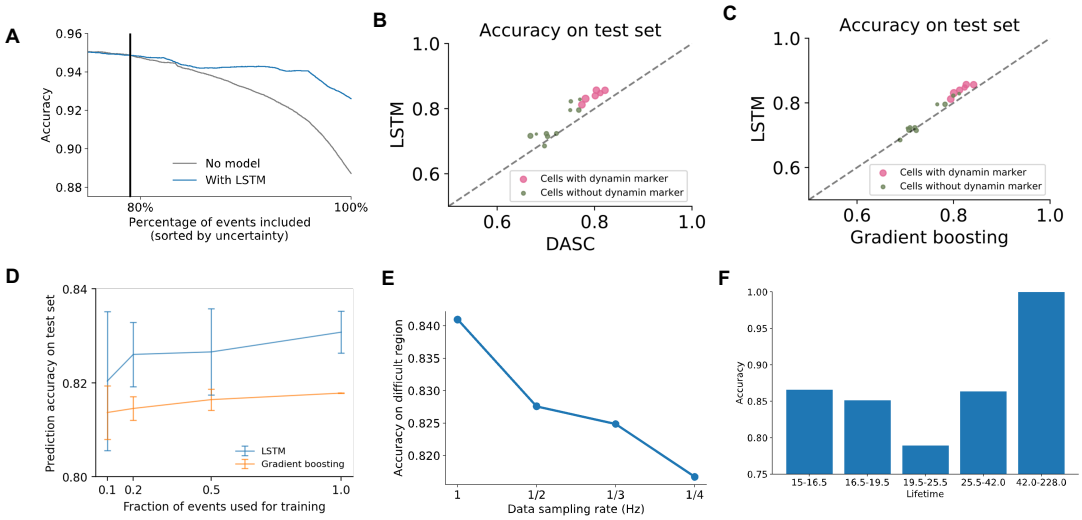


Figure 3.3: Performance of LSTM and competing models. **A** Cumulative accuracy with or without the (LSTM) model. Shows the overall performance for the model over entire dataset, including points with very short lifetimes. The model yields a predicted probability from 0-1, which we can use as its uncertainty. LSTM provides a significant improvement. The region to the right of the black vertical line corresponds to the "difficult region", and all other subparts of this figure examine the test accuracy of this region. **B** Comparison between classification accuracy of LSTM and DASC and **C** gradient boosting with hand-engineered features on the test set. In both figures, each point corresponds to a cell. Cells with dynamin markers are colored in pink, and are imaged under the same condition. Cells without dynamin markers are colored in green, and are imaged under 5 different conditions. The sizes of different points are proportional to the number of difficult clean tracks in the corresponding cell. For reference, we also plot $y = x$ as a dashed line. **D** Prediction accuracy of gradient-boosting and LSTM models when training on different number of samples. When data is limited, LSTM no longer has an advantage over gradient-boosting. When data downsampling rate is k, we randomly split all training data into k folds, and train on each of the k folds to obtain k different models. **E** When we sample the data at a different rate, performance degrades but not dramatically so. **F** Accuracy breakdown by lifetime.

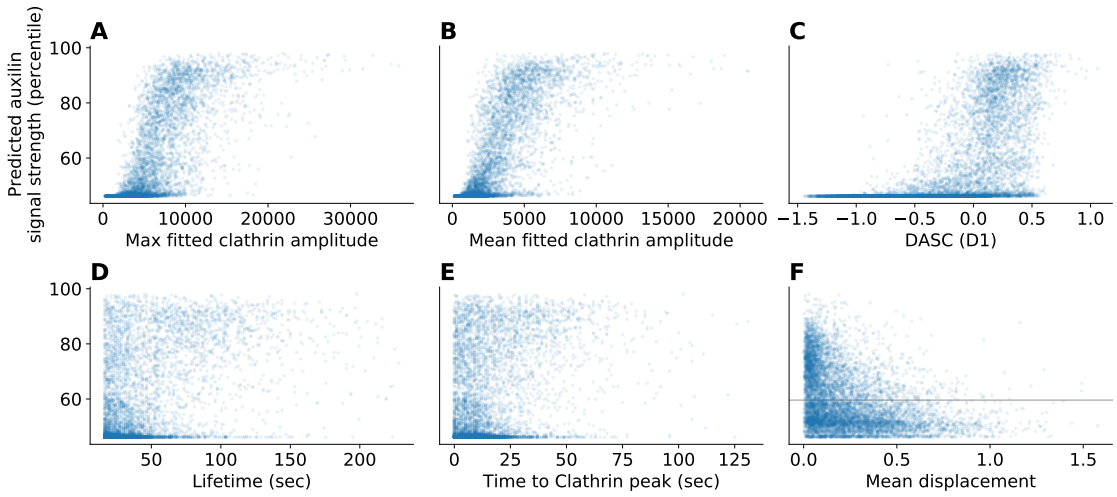


Figure 3.4: Scatter plots of model predictions of auxilin signal strength with different key features. Auxilin signal strength is defined as the mean fitted amplitude of significant fits to the auxilin signal. **A** Predictions for aux+ grow with both lifetime and maximum clathrin amplitude. **B** Predictions do not vary much with total displacement. **C** Predictions correlate with DASC features. **F** Note that mean total displacement is divided by lifetime, so larger total displacement is not the same as larger mean displacement.

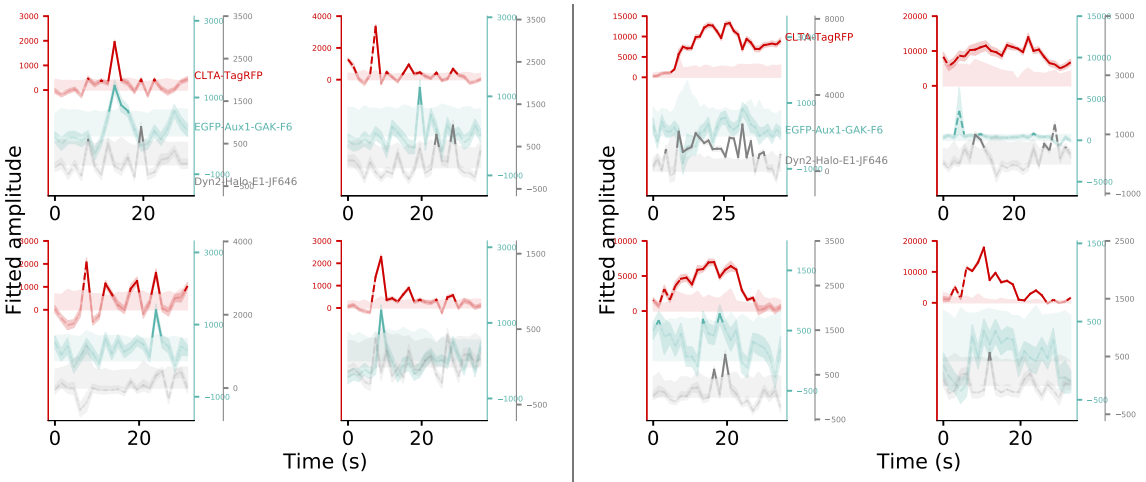
Interpretation

In this section, we interpret the model to audit and understand what it has learned, focusing only on the test set of one of the datasets (clath gak aux test). Figure 3.4 shows the predictions as a function of some key features. The model is trained to predict the peak amplitude of auxilin. This prediction is higher as an event's lifetime increases, it's max clathrin amplitude increases, or it's DASC feature increases.

The errors the model makes appear to be understandable (Figure 3.5). False negatives (Figure 3.5A) frequently occur for tracks which have short lifetimes, low clathrin amplitudes, and no clear spike at the errors the model makes; despite this lack of indicators, auxilin still occasionaly spikes, account for 7.3% of difficult tracks. On the other hand, false positive clathrin tracks (Figure 3.5B), display large clathrin amplitudes and have a distinctive amplitude drop but no corresponding auxilin spike.

Understanding the predictive model requires going beyond simply visualizing tracks based on their predictions. We would also like to identify the patterns a model *uses* to make its predictions. One approach to identifying such patterns is to assign feature importances to the inputs to the model. However, assigning feature importances to individual time points is relatively uninformative (e.g. see LIME [101] and SHAP [82] interpretations: Figure 3.10). This is because it is the trends/interactions between time points which drive the prediction, not time points by themselves.

In order to understand the interactions between time points, we use contextual decomposition [90, 114], a recent method which can score any interaction in a neural network prediction. We then devise an algorithm that succinctly summarizes the main interactions of an input sequence as



(A) False negative examples

(B) False positive examples

Figure 3.5: Model error examples. **A** False negative examples tend to have short lifetimes, low clathrin amplitudes, and no clear spike. **B** False positive examples display large clathrin amplitudes and have a distinctive amplitude drop but no corresponding auxilin spike.

a sequence of important segments. We also introduce a fast dynamic programming algorithm to quickly compute this segmentation (details in Sec 3.4).

3.3 Discussion

The particular model we develop is specific to CME and will not work if data is drastically different from the dataset here. Thus, judgement should be used when checking whether this model will work. However, this approach of molecular partner prediction can potentially be generalized to other scenarios.

One promising direction for future work would enable direct prediction from video rather than using a hand-engineered tracking step. This could potentially perform tracking in a more refined manner, better learning how to detect activity amid noise. We make a first pass at this, by using an architecture that predicts directly from video. The architecture is similar to that proposed in a previous work [136], but currently fails to outperform the combination of tracking and LSTM modeling. Future work could better enable video understanding through improved architectures.

3.4 Materials and methods

Processing

Tracking We use tracking code from previous work on 2D tracking [1] (which was later extended to 3D [2]). The tracking fits a Gaussian curve to the images (with standard deviation given by the

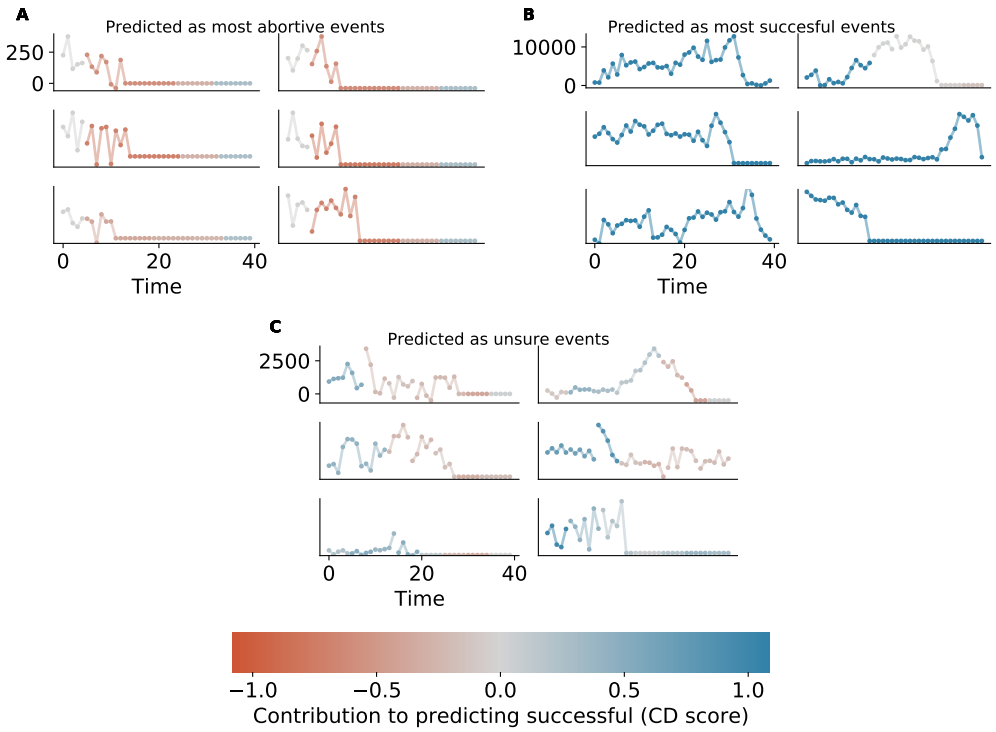


Figure 3.6: Importances of different parts of the clathrin track to predictions made by the neural network. **A** The most negative predictions see negative contributions due to their short lifetime (end of track) and lack of peaks near the end of the track. **B** The most positive predictions see their largest contributions at the highest amplitudes and after falling from a large peak. **C** Some neutral predictions made by the neural network. More examples in Figure 3.12.

camera parameters). When the fit to the first channel (i.e. clathrin) is significant[1], the track is recorded and a fit is forced to the second channel (i.e. auxilin). The amplitudes of each track are plotted over time in Figure 3.1B.

Preprocessing After running the tracking we exclude unclear tracks from our analysis. This includes tracks which start before the recording session / end after the recording session. We further exclude any tracks with a lifetime less than 3 frames.

We also identify "hotspots" - areas where there are multiple spikes and omit them. Since we are imaging a 3-dimensional volume in 2 dimensions, there are some pixels that correspond to multiple vesicles.

Defining the outcome For classification groundtruth, we manually defined the outcome using both the clathrin and auxilin. To begin with, we set all events as abortive. Then, we label an event as successful if its max auxilin amplitude is above a manually specified threshold of 973 or there

[1]Here, significant is defined to be p-value less than 0.05, but the results are not sensitive to this precise threshold.

are consecutive significant values in the auxilin fit. Then, curves are manually inspected to relabel. For regression, we take the mean of the significant auxilin observations.

Modeling

In the first step, the authors use a lifetime threshold, corresponding to a 95% threshold on classification accuracy. This results in all tracks with lifetime less than 16 being classified as abortive.

We then build a regression model to predict the mean amplitude of the significant auxilin observations. Once the regression model is fitted, we threshold the predicted outcome at zero to obtain the predicted class (i.e. abortive or valid).

We use a variety of machine-learning approaches, combined with feature-engineering approaches using appropriate domain knowledge. Features are normalized to have mean zero and standard deviation one before model fitting. Many machine-learning models are tried including Random Forests [21], SVMs, multi-layer perceptrons (i.e. fully connected neural networks), etc. More modeling approaches were tried but did not improve the results (see Sec 3.5). The model is also calibrated using some held-out data.

Interpretation

We introduce an algorithm to succinctly summarize interactions in the form of a segmentation. The algorithm works using scores for interactions (here contextual decomposition scores [90]) and aggregates them in order to identify key interactions (similar to the hierarchical procedure introduced by Singh, Murdoch, and Yu [114]). We assume we are given a sequence of inputs $x_1, x_2, ..., x_p$ and an interaction-scoring function $score(s, e)$, $s < e$ which returns the interaction score for a model using the inputs $x_s, x_{s+1}, ..., x_{e-1}$. Then the algorithm returns a segmentation (s_i, e_i) for $i = 0, ..., m$ where m is a variable number of segments. The segments are non-empty ($e_i > s_i$), non-overlapping ($s_i > s_{i-1}$) and cover the segment ($\forall j \; \exists i \; s.t. \; j \in [s_i, e_i)$). The heuristic objective to optimize is then

$$\max \sum_i |score(s_i, e_i)| \tag{3.1}$$

This objective aims to summarize interactions, i.e. when an interaction between two terms increases their score beyond the sum of the two terms, this objective will favor merging the terms into a segment.

Optimizing this difficult can be difficult (exponential in p), but using dynamic programming, we are able to compute it in only p^2 time. The dynamic programming algorithm starts at $t = 1$ and computes the values of the objective Eq (3.1). Then, t is increased from $t = 1$ to $t = p$, where each value of t uses the cached values of the objective for all $t' < t$.

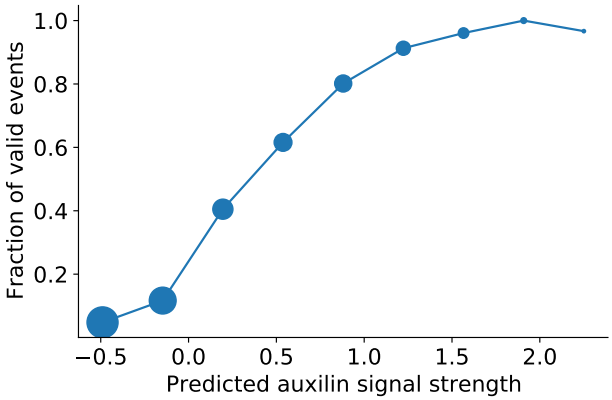


Figure 3.7: Model predictions are reasonably well-calibrated. Larger predictions by the model (of the fitted mean of the significant fitted amplitudes of the auxilin signal) correspond to true higher probabilities of valid events. Size of points correspond to number of points in each bin. Corresponds to the LSTM model's performance on the "difficult region" whose accuracy is given in Figure 3.3.

3.5 Supplementary tables and figures

Model performance and extended validation

In Figure 3.3, we show that the LSTM model is well-calibrated. In Table 3.2, we show the detailed classification and regression prediction accuracy of different methods, evaluated on all datasets. We see that the LSTM model is the overall best-performing model in terms of both classification and regression. Figure 3.8 shows the predicted and observed outcome for the regression task, where we predict the mean amplitude of the significant auxilin observations. In Figure 3.9, we show that the predictive accuracy of the model is robust to judgment calls made during modeling.

Interpretations continued

We compare the interpretation of the LSTM of different methods. Figure 3.10 and 3.11 show the pointwise of LIME and SHAP respectively. Figure 3.12 shows the interpretation of our pipeline for more tracks. We find that the interpretation made by our pipeline is more informative than those by LIME and SHAP.

Alternative modeling

Unsupervised learning We tried some approaches to learn features directly from tracks. These included sparse coding and non-negative matrix factorization, as shown by 3.13. However, we found that these features did not lead to higher prediction accuracy.

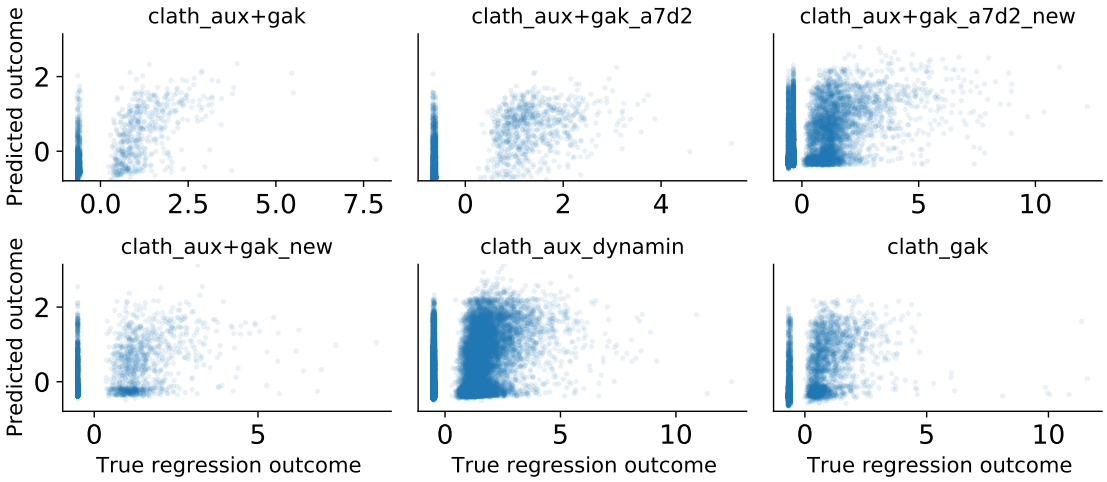


Figure 3.8: Prediction plots on test datasets for the LSTM model in FigureFigure 3.3 on 'difficult tracks' for all datasets.

Dataset	metric	GB	RF	SVM	DASC(SVM)	lstm
clath_aux+gak	accuracy	0.744	0.763	0.748	0.725	**0.769**
	roc_auc	0.840	0.831	0.832	0.789	**0.845**
	r2	0.408	0.396	0.387	0.271	**0.438**
	corr	0.643	0.632	0.628	0.536	**0.663**
clath_aux+gak_a7d2	accuracy	0.700	0.700	0.700	**0.709**	0.704
	roc_auc	0.788	0.788	**0.790**	0.773	0.785
	r2	**0.274**	0.265	0.240	0.248	0.228
	corr	**0.525**	0.520	0.513	0.506	0.507
clath_aux+gak_a7d2_new	accuracy	0.714	**0.727**	0.713	0.683	0.716
	roc_auc	0.830	0.825	0.830	0.771	**0.843**
	r2	0.415	0.403	0.395	0.287	**0.442**
	corr	0.648	0.638	0.639	0.559	**0.665**
clath_aux+gak_new	accuracy	0.804	0.785	0.804	0.758	**0.825**
	roc_auc	0.883	0.877	0.887	0.847	**0.889**
	r2	0.381	0.341	0.354	0.289	**0.388**
	corr	0.618	0.587	0.604	0.556	**0.625**
clath_aux_dynamin	accuracy	0.817	0.805	0.823	0.799	**0.841**
	roc_auc	0.896	0.886	0.891	0.878	**0.899**
	r2	0.383	0.366	0.347	0.308	**0.392**
	corr	0.622	0.605	0.615	0.579	**0.627**
clath_gak	accuracy	0.752	0.757	**0.767**	0.735	0.760
	roc_auc	0.840	0.837	**0.843**	0.808	0.841
	r2	**0.303**	0.291	0.289	0.215	0.290
	corr	0.552	0.542	0.555	0.484	**0.556**

Table 3.2: Classification and regression results for different datasets on "hard tracks". Best accuracy in each row is bolded. LSTM model generally outperforms all competing models. Extended version of Figure 3.3.

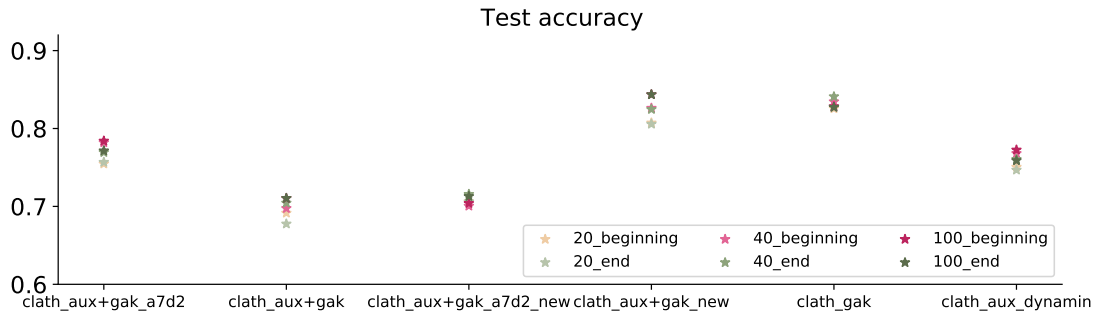


Figure 3.9: Investigating the robustness of LSTM prediction accuracy to various modeling judgement calls. Plotted are the test accuracy of 6 different models index by $(k, padding)$, for $k \in \{20, 40, 100\}$ and $padding \in \{\text{'beginning'}, \text{'end'}\}$. The parameter k is the length of tracks after downsampling, and the parameter $padding$ indicates whether we pad zeros at the beginning or end of the track.

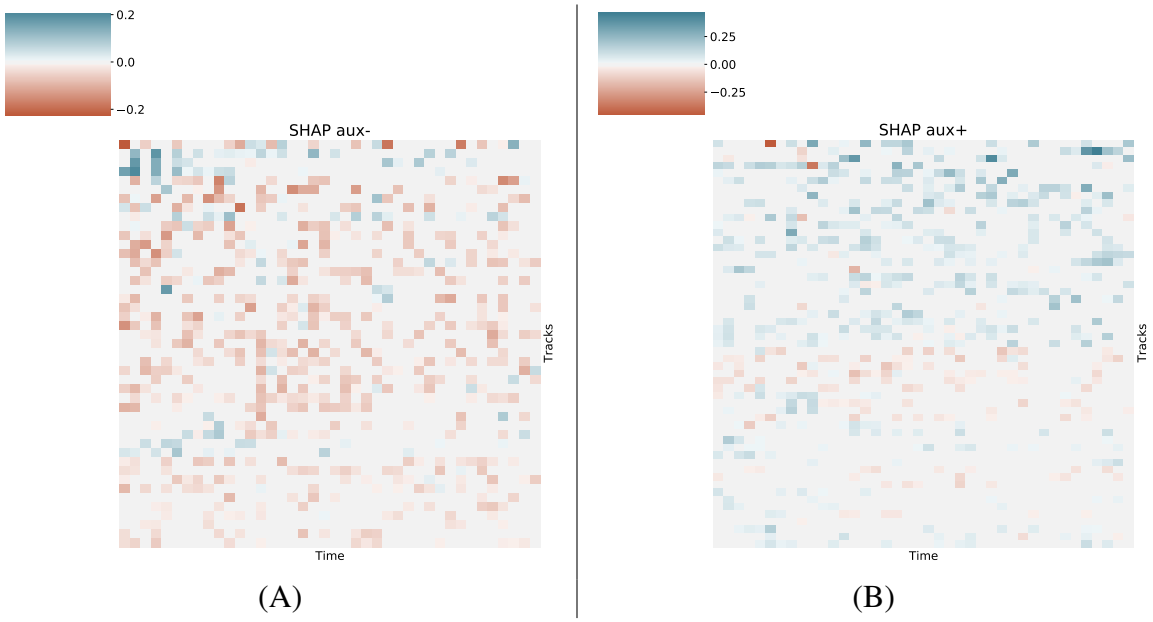
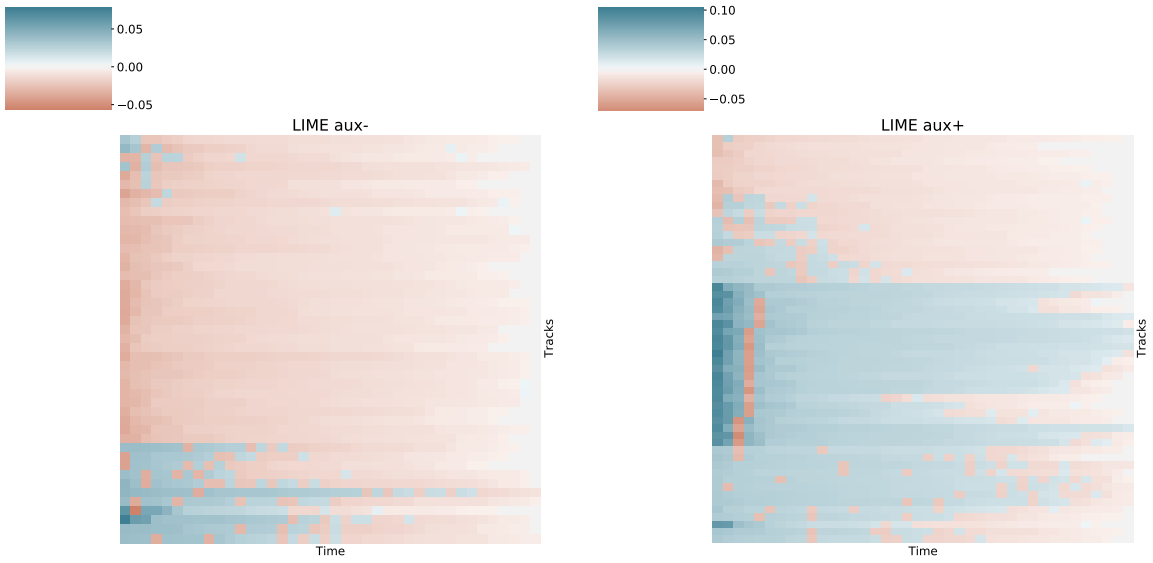


Figure 3.10: SHAP pointwise interpretations. Both for **A** aux- events and **B** events predictions, importances seem to be highest near the end of the track and have the same sign as the prediction. Corresponds to the interpretations of the LSTM model in Figure 3.6.

.



(A) (B)

Figure 3.11: LIME pointwise interpretations. Both for **A** aux- events and **B** events predictions, importances seem to be highest near the beginning of the track and have the opposite sign as the prediction. Corresponds to the interpretations of the LSTM model in Figure 3.6.

Neural networks We tried a handful of neural network architectures. In all cases, tracks were resampled to have a length of 40 points. We began with fully connected neural networks using the ReLU nonlinearity. We also tried convolutional neural networks, GRUs, and LSTMs [54], among other architectures. We also tried a linear model on top of learning single convolutions. Of all the architectures tests, the LSTM network has the lowest validation error and is therefore selected.

DASC baseline DASC does a reasonable job separating the classes in Figure 3.14.

Model limits

Dynamic time warping Figure 3.15 shows closest matches using dynamic time warping [68, 106], with the constraint that the lifetimes are within seven frames of one another.

Further data description

There is some distribution shift between different datasets. See Figure 3.16.

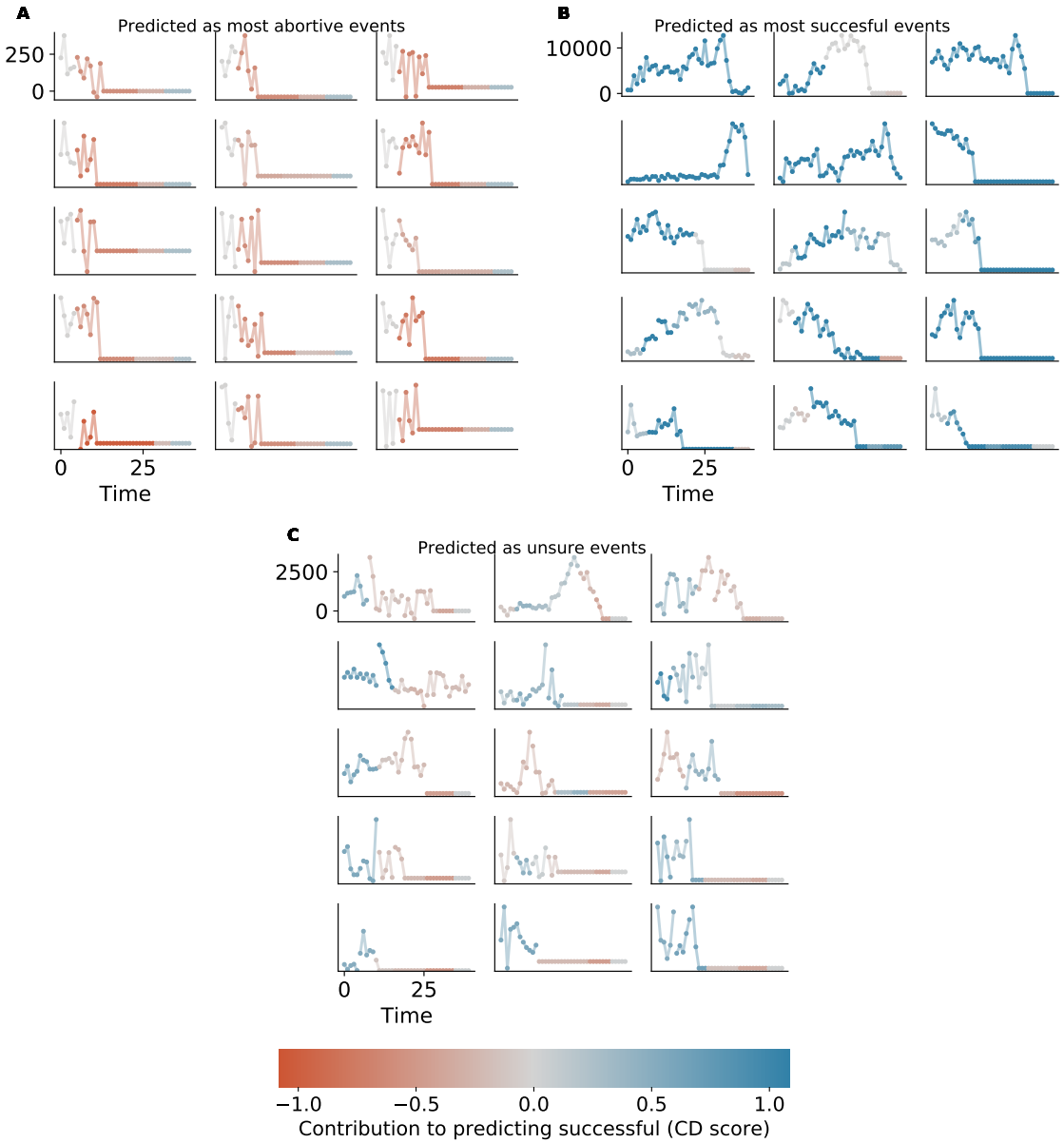


Figure 3.12: Importances of different parts of the clathrin track to predictions made by the neural network. **A** The most negative predictions see negative contributions due to their short lifetime (end of track) and lack of peaks near the end of the track. **B** The most positive predictions see their largest contributions at the highest amplitudes and after falling from a large peak.

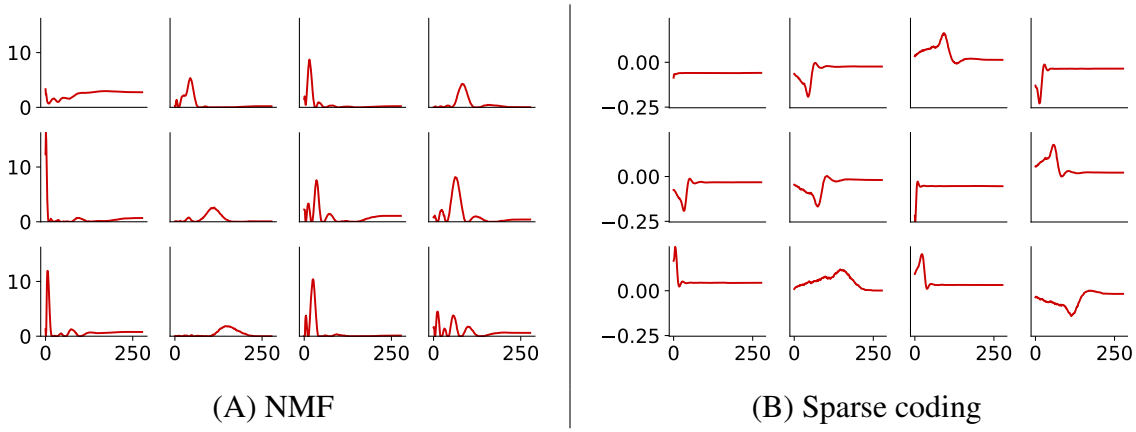


Figure 3.13: Features extracted via unsupervised learning. Left is NMF, right is sparse coding.

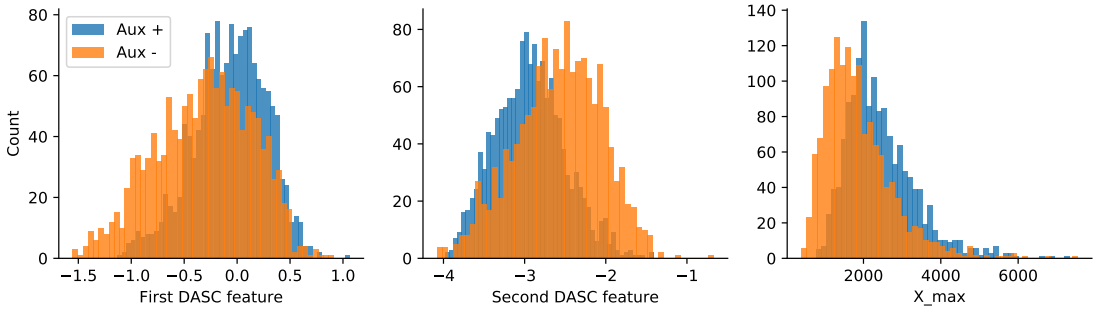


Figure 3.14: DASC features decently separate abortive and valid events.

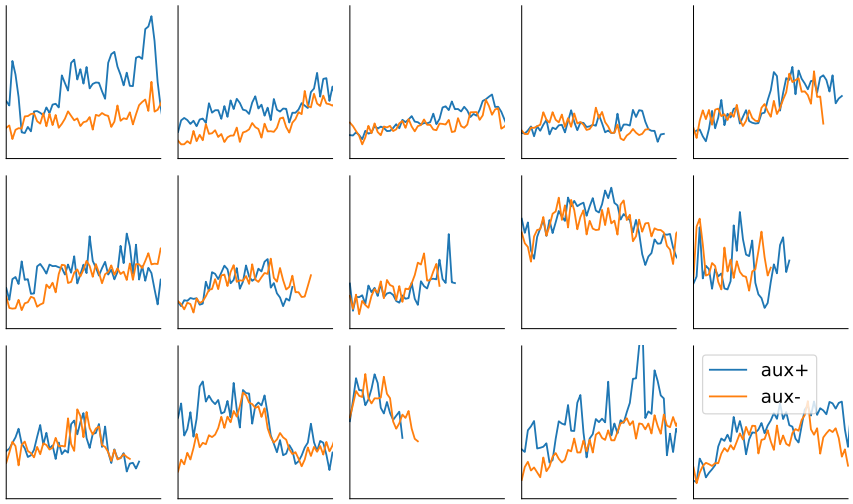


Figure 3.15: Some clathrin tracks from aux+ and aux- events are visually extremely similar, suggesting there is some level of stochasticity in the auxilin spike that cannot be predicted from the clathring track alone.

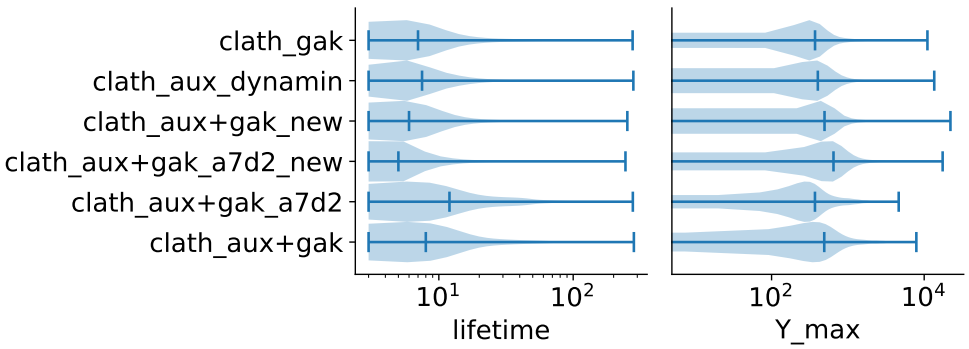


Figure 3.16: Distribution shift.

Chapter 4

A Debiased MDI Feature Importance Measure for Random Forests

4.1 Introduction

Understanding how a machine learning (ML) model makes predictions is important in many scientific and industrial problems [91]. Appropriate interpretations can help increase the predictive performance of a model and provide new domain insights. While a line of study focuses on interpreting any generic ML model [120, 101], there is a growing interest in developing specialized methods to understand specific models. In particular, interpreting Random Forests (RFs) [18] and its variants [79, 119, 117, 118, 13, 69] has become an important area of research due to the wide ranging applications of RFs in various scientific areas, such as genome-wide association studies (GWAS) [34], gene expression microarray [71, 103], and gene regulatory networks [58].

A key question in understanding RFs is how to assign feature importance. That is, which features does a RF rely on for prediction? One of the most widely used feature importance measures for RFs is mean decrease impurity (MDI) [20]. MDI computes the total reduction in loss or impurity contributed by all splits for a given feature. This method is computationally very efficient and has been widely used in a variety of applications [107, 58]. However, theoretical analysis of MDI has remained sparse in the literature [64]. Assuming there are an infinite number of samples, Louppe et al. [81] characterized MDI for totally randomized trees using mutual information between features and the response. They showed that noisy features, i.e., features independent of the outcome, have zero MDI importance. However, empirical studies have shown that MDI systematically assigns higher feature importance values to numerical features or categorical features with many categories [118]. In other words, high MDI values do not always correspond to the predictive associations between features and the outcome. We call this phenomenon MDI *feature selection bias*. Louppe [80] studied this issue and demonstrate via simulations that early stopping mechanisms (e.g., limited depth and larger leaf sizes) are effective to reduce the feature selection bias.

In this paper, using the original definition of MDI, we analyze the non-asymptotic behavior

of MDI and bridge the gap between the population case and the finite sample case. We find that under mild conditions, if the samples used for each tree are i.i.d, then the expected MDI feature importance of noisy features derived from any tree ensemble constructed on n samples with p features is upper bounded by $d_n \log(np)/m_n$, where m_n is the minimum leaf size and d_n is the maximum tree depth in the ensemble. In other words, deep trees with small leaves suffer more from feature selection bias. Our findings are particularly relevant for practical applications involving RFs, in which scenario deep trees are recommended [18] and used more often. To reduce the feature selection bias for RFs, especially when the trees are deep, we derive a new analytical expression for MDI and then use this new expression to propose a new feature importance measure that evaluates MDI using out-of-bag samples. We call this importance measure MDI-oob. For both regression and classification problems, we use simulated data and a genomic dataset to demonstrate that MDI-oob often achieves 5%–10% higher AUC scores compared to other feature importance measures used in several publicly available packages including `party` [27], `ranger` [130], and `scikit-learn` [100].

Related works

In addition to MDI [128, 83], some other feature importance measures have been studied in the literature and used in practice:

- Split count, namely, the number of times a feature is used to split [118], can be used as a feature importance measure. This method has been studied in [119, 13] and is available in XGBoost [31].
- Mean decrease in accuracy (MDA) measures a feature's importance by the reduction in the model's accuracy after randomly permuting the values of a feature. The motivation of MDA is that permuting an important feature would result in a large decrease in the accuracy while permuting an unimportant feature would have a negligible effect. Different permutation choices have been studied in [119, 61].

Recently, Lundberg et al. [83] show that for feature importance measures such as MDI and split counts, the importance of a feature does not always increase as the outcome becomes more dependent on that feature. To remedy this issue, they propose the tree SHAP feature importance, which focuses on giving consistent feature attributions to each sample. When individual feature importance is obtained, overall feature importance is straightforward to obtain by just averaging the individual feature importances across samples.

While our paper focuses on interpreting trees learned via the classic RF procedure, there is another line of work that focuses on modifying the tree construction procedure to obtain better feature importance measures. Hothorn et al. [57] introduced cforest in the R package `party` that grows classification trees based on a conditional inference framework. Strobl et al. [118] showed that cforest suffers less from the feature selection bias. Sandri and Zuccolotto [107] proposed to create a set of uninformative pseudo-covariates to evaluate the bias in Gini importance. Nembrini et al. [94] gave a modified algorithm that is faster than the original method proposed by Sandri

and Zuccolotto [107] with almost no overhead over the creation of the original RFs and available in the R package `ranger`. In a very recent paper, Zhou and Hooker [137] proposed to evaluate the decrease in impurity at each node using out-of-bag samples. However, our implementation is different from that in [137] and MDI-oob enjoys higher computational efficiency.

In Section 4.4, we will compare MDI-oob with all the aforementioned methods except the split count, for which we did not find a package that implements it for RFs.

Organization

The rest of this paper is organized as follows. In Section 4.2, we provide a non-asymptotic analysis to quantify the bias in the MDI importance when noisy features are independent of relevant features. In Section 4.3, we give a new characterization of MDI and propose a new MDI feature importance using out-of-bag samples, which we call MDI-oob. In Section 4.4, we compare our MDI-oob with other commonly used feature importance measures in terms of feature selection accuracy using the simulated data and a genomic ChIP dataset. We conclude our work and discuss possible future directions in Section 4.5.

4.2 Understanding the feature selection bias of MDI

In this section, we focus on understanding the finite sample properties of MDI importance and why it may have a significant bias in feature selection. We first briefly review the construction of RFs and introduce some important notations. Then, using the original definition of MDI, we give a tight upper bound to quantify the expected bias of MDI importance for a noisy feature. This upper bound is tight up to a $\log n$ factor where n is the number of i.i.d. samples.

Background and notations

Suppose that the data set $\mathcal{D}$ contains n i.i.d samples from a random vector $(X_1, \ldots, X_p, Y)$, where $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ are p input features and $Y \in \mathbb{R}$ is the response. The i^{th} sample is denoted by $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. We say that a feature X_k is a *noisy* feature if X_k and Y are independent, and a *relevant* feature otherwise. Note that this definition of noisy features has also been used in many previous papers such as [81, 109]. We denote $S \subset [p]$ as the set of indexes of relevant features. We are particularly interested in the case where the number of relevant features is small, namely, $|S|$ is much smaller than p. For any number $m \in \mathbb{N}$, $[m]$ denotes the set of integers $\{1, \ldots, m\}$. For any hyper-rectangle $R \subset \mathbb{R}^p$, let $\mathbb{1}(X \in R)$ be the indicator function taking value one when $X \in R$ and zero otherwise.

RFs are an ensemble of classification and regression trees, where each tree T defines a mapping from the feature space to the response. Trees are constructed independently of one another on a bootstrapped or subsampled data set $\mathcal{D}^{(T)}$ of the original data $\mathcal{D}$. Any node t in a tree T represents a subset (usually a hyper-rectangle) R_t of the feature space. A split of the node t is a pair (k, z) which divides the hyper-rectangle R_t into two hyper-rectangles $R_t \cap \mathbb{1}(X_k \leq z)$ and $R_t \cap \mathbb{1}(X_k > z)$,

corresponding to the left child t^{left} and right child t^{right} of node t, respectively. For a node t in a tree T, $N_n(t) = |\{i \in \mathcal{D}^{(T)} : \mathbf{x}_i \in R_t\}|$ denotes the number of samples falling into R_t and

$$\mu_n(t) := \frac{1}{N_n(t)} \sum_{i:\mathbf{x}_i \in R_t} y_i \tag{4.1}$$

denotes their average response.

Each tree T is grown using a recursive procedure which proceeds in two steps for each node t. First, a subset $\mathcal{M} \subset [p]$ of features is chosen uniformly at random. Then the optimal split $v(t) \in \mathcal{M}, z(t) \in \mathbb{R}$ is determined by maximizing:

$$\Delta_{\mathcal{I}}(t) := \text{Impurity}(t) - \frac{N_n(t^{\text{left}})}{N_n(t)}\text{Impurity}(t^{\text{left}}) - \frac{N_n(t^{\text{right}})}{N_n(t)}\text{Impurity}(t^{\text{right}}) \tag{4.2}$$

for some impurity measure $\text{Impurity}(t)$. The procedure terminates at a node t if two children contain too few samples, i.e., $\min\{N_n(t^{\text{left}}), N_n(t^{\text{right}})\} \leq m_n$, or if all responses are identical. The threshold m_n is called the *minimum leaf size*. If a node t does not have any children, it is called a leaf node; otherwise, it is called an inner node. We define the set of inner nodes of a tree T as $I(T)$. We say that T' is a sub-tree of T if T' can be obtained by pruning some nodes in T.

Some popular choices of the impurity measure $\text{Impurity}(t)$ include variance, Gini index, or entropy. For simplicity, we focus on the variance of the responses, i.e.,

$$\text{Impurity}(t) = \frac{1}{N_n(t)} \sum_{i:\mathbf{x}_i \in R_t} (y_i - \mu_n(t))^2, \tag{4.3}$$

throughout the paper unless stated otherwise. Later we show that this definition of impurity is equivalent to the Gini index of categorical variables with one hot encoding (see Remark in Section 4.3)

The Mean Decrease Impurity (MDI) feature importance of X_k, with respect to a single tree T (first proposed by Breiman et al. in [20]) and an ensemble of n_{tree} trees $T_1, \ldots, T_{n_{tree}}$, can be written as

$$\text{MDI}(k, T) = \sum_{t \in I(T), v(t)=k} \frac{N_n(t)}{n} \Delta_{\mathcal{I}}(t) \quad \text{and} \quad \text{MDI}(k) = \frac{1}{n_{tree}} \sum_{s=1}^{n_{tree}} MDI(k, T_s), \tag{4.4}$$

respectively. This expression is the best known formula for MDI and was analyzed in many papers such as Louppe et al. [81].

Finite sample bias of MDI importance for Random Forests

Given the set S of relevant features and a tree T, we denote

$$G_0(T) = \sum_{k \notin S} \text{MDI}(k, T) \tag{4.5}$$

as the sum of MDI importance of all noisy features. Ideally, $G_0(T)$ should be close to zero with high probability, to ensure that no noisy features get selected when using MDI importance for feature selection. In fact, Louppe et al. [81] show that $G_0(T)$ is indeed zero almost surely if we grow totally randomized trees with infinite samples. However, $G_0(T)$ is typically non-negligible in real data, and finite sample properties of $G_0(T)$ are not well understood. In order to bridge this gap, we conduct a non-asymptotic analysis of the expected value of $G_0(T)$. Our main result characterizes how the expected value of $G_0(T)$ depends on m_n, the minimum leaf size of T, and p, the dimension of the feature space. We start with the following simple but important fact.

Fact 1. *If T' is a sub-tree of T, then* $\text{MDI}(k, T') \leq \text{MDI}(k, T)$ *for any feature X_k.*

This fact naturally follows from the observation that by definition, $\Delta_{\mathcal{I}}(t) \geq 0$ for any node t. Since the impurity decrease at each node is guaranteed to be non-negative, $G_0(T)$ will never decrease as T grows deeper, in which case the minimum leaf size m_n will be smaller. Indeed, if T is grown to purity ($m_n = 1$), and all features are noisy ($S = \emptyset$), then $G_0(T)$ would simply be equal to the sample variance of the responses in the data $\mathcal{D}^{(T)}$. How fast does $G_0(T)$ increase as the minimum leaf size m_n becomes smaller? To quantify the relation between $G_0(T)$ and m_n, we need a few mild conditions which we now describe. Let

$$y_i = \phi(\mathbf{x}_{i,S}) + \epsilon_i, i = 1, \ldots, n \tag{4.6}$$

for some unknown function $\phi : \mathbb{R}^{|S|} \to \mathbb{R}$, where ϵ_i are i.i.d zero-mean Gaussian noise. We make the following assumptions.

(A1) $X_k \sim \text{Unif}[0, 1]$ for all $k \in [p]$. In addition, the noisy features $\{X_k, k \in [p] \backslash S\}$ are mutually independent, and independent of all relevant features. Here S denotes the set of relevant features.

(A2) ϕ is bounded: $\sup_{\mathbf{x} \in [0,1]^{|S|}} |\phi(\mathbf{x})| \leq M$ for some $M > 0$.

The Assumptions (A1) and (A2) are weaker than the assumptions usually made when studying the statistical properties of RF. The marginal uniform distribution condition in (A1) is common in the RF literature [109], and can be easily satisfied by transforming the features via its inverse CDF. Since we are interested in characterizing the MDI of noisy features, we do not require the relevant features to be independent of each other. We do require that noisy features are independent of relevant features, which is a limitation of Theorem 1 below. Correlated features are commonly encountered in practice and difficult for any feature selection method.

We now state our first main result which provides a non-asymptotic upper and lower bound for the expected value of the maximum of $G_0(T)$ over all tree T with minimum leaf size m_n.

Theorem 1. *Let $\mathcal{T}_n(m_n)$ denote the set of decision trees whose minimum leaf size is lower bounded by m_n, and $\mathcal{T}_n(m_n, d_n) \subset \mathcal{T}_n(m_n)$ denote the subset of $\mathcal{T}_n(m_n)$ whose depth is upper bounded by d_n. Under Assumptions (A1) and (A2), there exists a positive constant C such that,*

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n, d_n)} G_0(T) \leq C \frac{d_n \log(np)}{m_n}. \tag{4.7}$$

In addition, when $f = 0$ *and* $m_n \geq 36 \log p + 18 \log n$*,*

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n)} G_0(T) \geq \frac{\log p}{C m_n}. \tag{4.8}$$

We give the proof in the Appendix. To the best of our knowledge, Theorem 1 is the first non-asymptotic result on the expected MDI importance of tree ensembles. In particular, the upper bound can be directly applied to *any* tree ensembles with a minimum leaf size m_n and a maximum tree depth d_n, including Breiman's original RF procedure, if subsampling is used instead of bootstrapping.

Proof Sketch. Every node t in a tree $T \in \mathcal{T}_n(m_n, d_n)$ corresponds to an axis-aligned hyper-rectangle in $[0,1]^p$ which contains at least m_n samples and is formed by splitting on at most d_n dimensions consecutively. Therefore, bounding the supremum of impurity reduction for any potential node in $\mathcal{T}_n(m_n, d_n)$ boils down to controlling the complexity of all such hyper-rectangles. Two hyper-rectangles are considered equivalent if they contain the same subset of samples, since the impurity reductions of these two hyper-rectangles are always the same. Up to this equivalence, it can be proved that the number of unique hyper-rectangles of interest is upper bounded by $(np)^{d_n}$, which corresponds to the $d_n log(np)$ term in the upper bound. The final result is obtained via union bound. □

In the upper bound, each node t is obtained by splitting on at most d_n features. In practice, d_n is typically at most of order $\log n$. Indeed, if the decision tree is a balanced binary tree, then $d_n \leq \log_2 n$. Therefore, for balanced trees, the upper bound can be written as

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n, d_n)} G_0(T) \leq C \frac{d_n \log(np)}{m_n} \leq C \frac{(\log n)^2 + \log n \log p}{m_n}, \tag{4.9}$$

and the theorem shows that the sum of MDI importance of noisy features is of order $\frac{\log p}{m_n}$, i.e.,

$$\sup_{\phi : \|\phi\|_\infty \leq M} \mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n)} G_0(T) \sim \frac{\log p}{m_n}, \tag{4.10}$$

up to a $\log n$ term correction, which is typically small in the high dimensional $p \gg n$ setting. If all features X_j are categorical with a bounded number of categories, then the upper bound can be improved to

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n, d_n)} G_0(T) \leq C \frac{d_n \log p}{m_n}, \tag{4.11}$$

which shows that the MDI importance of noisy features can be better controlled if the noisy features are categorical rather than numerical. That is consistent with the previous empirical studies because the number of candidate split points for a numerical feature is larger than that for a categorical feature.

Theorem 1 shows that the supremum of MDI importance of noisy features over all trees with minimum leaf size m_n is, in expectation, roughly inversely proportional to m_n. In the Appendix Fig. 4.4, we show that the inversely proportional relationship is consistent with the empirical

$G_0(T)$ on a simulated dataset described in the first simulation study in Section 4.4. Therefore, to control the finite sample bias of MDI importance, one should either grow shallow trees, or use only the shallow nodes in a deep tree when computing the feature importance. In fact, since $G_0(T)$ depends on the dimension p only through a log factor $\log p$, we expect $G_0(T)$ to be very small even in a high-dimensional setting if m_n is larger than, say, $\sqrt{n}$. For a balanced binary tree grown to purity with depth $d_n = \log_2 n$, this corresponds to computing MDI only from the first $d_n/2 = (\log_2 n)/2$ levels of the tree, as the node size on the dth level of a balanced tree is $n/2^d$.

Fact 1 implies that the MDI importance of relevant features might also decrease as m_n increases, but we will show in simulation studies that they will decrease at a much slower rate, especially when the underlying model is sparse.

4.3 MDI using out-of-bag samples (MDI-oob)

As shown in the previous section, for balanced trees, the sum of MDI feature importance of all noisy features is of order $\frac{\log(p)}{m_n}$ if we ignore the $\log(n)$ terms. This means that the MDI feature selection bias becomes severe for trees with smaller leaf size m_n, which usually corresponds to a deeper tree. Fortunately, this bias can be corrected by evaluating MDI using out-of-bag samples. In this section, we first introduce a new analytical expression of MDI as the motivation of our new method, then we propose the MDI-oob as a new feature importance measure. For simplicity, in this section, we only focus on one tree T. However, all the results are directly applicable to the forest case.

A new characterization of MDI

Recall that the original definition of the MDI importance of any feature k is provided in Equation (4.4), that is, the sum of impurity decreases among all the inner nodes t such that $v(t) = k$. Although we can use this definition to analyze the feature selection bias of MDI in Theorem 1, this expression (4.4) gives us few intuitions on how we can get a new feature importance measure that reduces the MDI bias. Next, we derive a novel analytical expression of MDI, which shows that the MDI of any feature k can be viewed as the sample covariance between the response y_i and the function $f_{T,k}(\mathbf{x}_i)$ defined in Proposition 1. This new expression inspires us to propose a new MDI feature importance measure by using the out-of-bag samples.

Proposition 1. *Define the function $f_{T,k}(\cdot)$ to be*

$$f_{T,k}(X) = \sum_{t \in I(T): v(t)=k} \Big\{ \mu_n(t^{\mathit{left}}) \mathbb{1}(X \in R_{t^{\mathit{left}}}) + \mu_n(t^{\mathit{right}}) \mathbb{1}(X \in R_{t^{\mathit{right}}}) - \mu_n(t) \mathbb{1}(X \in R_t) \Big\}.$$

Then the MDI of the feature k in a tree T can be written as:

$$\frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i) \cdot y_i, \tag{4.12}$$

We give the proof in the Appendix. The proof is just a few lines but it requires a good understanding of MDI. Although we have not seen this analytical expression in the prior works, we found that the functions $f_{T,k}(\cdot)$ have been studied from a quite different perspective. Those functions were first proposed in Saabas [105] to interpret the RF predictions for each individual sample. According to this paper, $f_{T,k}$ can be viewed as the "contribution" made by the feature k in the tree T. For any tree, those functions $f_{T,k}$ can be easily computed using the python package *treeinterpreter*.

It can be shown that $\sum_{i\in\mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i) = 0$. That implies $\frac{1}{|\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i)\cdot y_i$ is essentially the sample covariance between $f_{T,k}(\mathbf{x}_i)$ and y_i on the bootstrapped dataset $\mathcal{D}^{(T)}$. This indicates a potential drawback of MDI: RFs use the training data $\mathcal{D}^{(T)}$ to construct the functions $f_{T,k}(\cdot)$, then MDI uses the same data to evaluate the covariance between y_i and $f_{T,k}(\mathbf{x}_i)$ in Equation (4.12).

Remark: So far we have only considered regression trees, and have defined the impurity at a node t using the sample variance of responses. For classification trees, one may use Gini index as the measure of impurity. We point out that these two definitions of impurity are actually equivalent when we use a one-hot vector to represent the categorical response. Therefore, our results are directly applicable to the classification case. Suppose that Y is a categorical variable which can take D values $c_1, c_2, \ldots, c_D$. Let $p_d = \mathbb{P}(Y = c_d)$. Then the Gini index of Y is $\text{Gini}(Y) = \sum_{d=1}^{D} p_d(1-p_d)$. We define the one-hot encoding of Y as a D-dimensional vector $\tilde{Y} = (\mathbb{1}(Y = c_1), \ldots, \mathbb{1}(Y = c_D))$. Then

$$\begin{aligned}
\text{Var}(\tilde{Y}) &= \|\tilde{Y} - \mathbb{E}\tilde{Y}\|_2^2 \\
&= \sum_{d=1}^{D}(\mathbb{E}\tilde{Y}_i^2 - (\mathbb{E}\tilde{Y}_i)^2) \\
&= \sum_{d=1}^{D}(\mathbb{E}\tilde{Y}_i - (\mathbb{E}\tilde{Y}_i)^2) \\
&= \sum_{d=1}^{D} p_d(1-p_d) = \text{Gini}(Y),
\end{aligned} \tag{4.13}$$

thereby showing that Gini index and variance are equivalent.

Evaluating MDI using out-of-bag samples

Proposition 1 suggests that we can calculate the covariance between y_i and $f_{T,k}(\mathbf{x}_i)$ in Equation (4.12) using the out-of-bag samples $\mathcal{D}\backslash\mathcal{D}^{(T)}$:

$$\text{MDI-oob of feature } k = \frac{1}{|\mathcal{D}\backslash\mathcal{D}^{(T)}|}\sum_{i\in\mathcal{D}\backslash\mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i)\cdot y_i. \tag{4.14}$$

In other words, for each tree, we calculate the $f_{T,k}(\mathbf{x}_i)$ for all the OOB samples $\mathbf{x}_i$ and then compute MDI-oob using (4.14). Although out-of-bag samples have been used for other feature importance measures such as MDA, to the best of the authors' knowledge, there are few results that use the out-of-bag samples to evaluate MDI feature importance. A naive way of using the out-of-bag samples to evaluate MDI is to directly compute the impurity decrease at each inner-node of a tree using OOB samples. However, this approach is not desirable since the impurity decrease at each node is still always positive unless the responses of all the OOB samples falling into a node are constant. In this case, an argument similar to the proof of Theorem 1 can show that the bias of directly computing impurity using OOB samples could still be large for deep trees. The idea of MDI-oob depends heavily on the new analytical MDI expression. Without the new expression, it is not clear how one can use out-of-bag samples to get a better estimate of MDI. One highlight of the MDI-oob is its low computation cost. The time complexity of evaluating MDI-oob for RFs is roughly the same as computing the RF predictions for $|\mathcal{D}\backslash\mathcal{D}^{(T)}|$ number of samples.

4.4 Simulation experiments

Simulated study on the effect of minimum leaf size and the tree depth

In this simulation, we investigate the empirical relationship between MDI importance and the minimum leaf size. To mimic the major experiment setting in the paper [118], we generate the data as follows. We sample $n = 200$ observations, each containing 5 features. The first feature is generated from standard Gaussian distribution. The second feature is generated from a Bernoulli distribution with $p = 0.5$. The third/fourth/fifth features have 4/10/20 categories respectively with equal probability of taking any states. The response label y is generated from a Bernoulli distribution such that $P(y_i = 1) = (1 + x_{i2})/3$. While keeping the number of trees to be 300, we vary the minimum leaf size of RF from 1 to 50 and record the MDI of every feature. The results are shown in Fig. 4.1. We can see from this figure that the MDI of noisy features, namely X1, X3, X4 and X5, drops significantly when the minimum leaf size increases from 1 to 50. This observation supports our theoretical result in Theorem 1. Besides the minimum leaf size, we also investigate the relationship between MDI and the tree depth. As tree depth increases, the minimum leaf size generally decreases exponentially. Therefore, we expect the MDI of noisy features to become larger for increasing tree depth. We vary the maximum depth from 1 to 20 and record the MDI of every feature. The results shown in Fig. 4.2 are consistent with our expectation. MDI importance of noisy features increase when the tree depth increases from 1 to 20. Fig. 4.3 shows the MDI-oob measure and it indeed reduces the bias of MDI in this simulation.

Noisy feature identification using the simulated data

In this experiment, we evaluate different feature importance measures in terms of their abilities to identify noisy features in a simulated data set. We compare our method with the following methods: MDA, cforest in the R package `party`, SHAP[83], default feature importance (MDI) in `scikit-learn`, the impurity corrected Gini importance in the R package `ranger`, UFI in

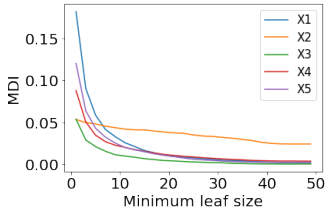


Figure 4.1: MDI against min leaf size.

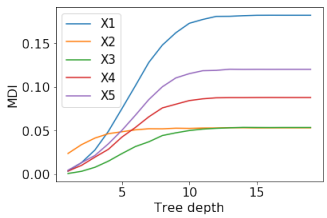


Figure 4.2: MDI against tree depth.

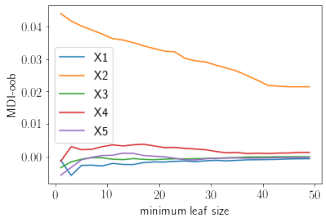


Figure 4.3: MDI-oob against min leaf size.

[137], and naive-oob, which refers to the naive method that evaluates impurity decrease at each node using out-of-bag samples directly. To evaluate feature importance measures, we generate the following simulated data. Inspired by the experiment settings in Strobl et al. [118], our setting involves discrete features with different number of distinct values, which poses a critical challenge for MDI. The data has 1000 samples with 50 features. All features are discrete, with the j^{th} feature containing $j+1$ distinct values $0, 1, \ldots, j$. We randomly select a set S of 5 features from the first ten as relevant features. The remaining features are noisy features. Choosing active features with fewer categories represents the most challenging case for MDI. All samples are i.i.d. and all features are independent. We generate the outcomes using the following rules:

- Classification: $P(Y=1|X) = \text{Logistic}(\frac{2}{5}\sum_{j\in S} X_j/j - 1)$.
- Regression: $Y = \frac{1}{5}\sum_{j\in S} X_j/j + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 100 \cdot \text{Var}(\frac{1}{5}\sum_{j\in S} X_j/j))$.

Treating the noisy features as label 0 and the relevant features as label 1, we can evaluate a feature importance measure in terms of its area under the receiver operating characteristic curve (AUC). Note that when a feature importance measure gives low importance to relevant features, its AUC score measure can be smaller than 0.5 or even 0. We grow 100 trees with the minimum leaf size set to either 100 (shallow tree case) or 1 (deep tree case). The number of candidate features m_{try} is set to be 10. We repeat the whole process 40 times and report the average AUC scores for each method in Table 4.1. The boxplots For this simulated setting, MDI-oob achieves the best AUC score under all cases.

Noisy feature identification using a genomic ChIP dataset

To evaluate our method MDI-oob in a more realistic setting, we consider a ChIP-chip and ChIP-seq dataset measuring the enrichment of 80 biomolecules at 3912 regions of the *Drosophila* genome [28, 84]. These data have previously been used in conjunction with RF-based methods, namely iterative random forests (iRF) [13], to predict functional labels associated with genomic regions. They provide a realistic representation of many issues encountered in practice, such as heterogeneity and dependencies among features, which make it especially challenging for feature selection problems. To evaluate feature selection in the ChIP data, we scale each feature X_j to be between 0 and 1. Second, we randomly select a set S of 5 features as relevant features and include the rest

as noisy features. We randomly permute values of any noisy features to break their dependencies with relevant features. By this means, we avoid the cases where RFs "think" some features are important not because they themselves are important but because they are highly correlated with other relevant features. Then we generate responses using the following rules:

- Classification: $P(Y = 1|X) = \text{Logistic}(\frac{2}{5}\sum_{j\in S} X_j - 1)$.
- Regression: $Y = \frac{1}{5}\sum_{j\in S} X_j + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 100 \cdot \text{Var}(\frac{1}{5}\sum_{j\in S} X_j))$.

All the other settings remain the same as the previous simulations. We report the average AUC scores for each method in Table 4.1. The standard errors and the beeswarm plots of all the methods are included in the Appendix. Naive-oob, namely, the method that directly computes MDI using the out-of-bag samples is hardly any better than the original gini importance. MDI-oob or UFI usually achieves the best AUC score in three out of four cases, except for shallow regression trees, when all methods appear to be equally good with AUC scores close to 1. Although UFI and MDI-oob use out-of-bag samples in different ways, their results are generally comparable. We also note that increasing the minimum leaf size consistently improves the AUC scores of all methods.

Another observation is that MDA behaves poorly in some simulations despite its use of a validation set. This could be due to the low signal-to-noise ratio in the simulation setting. After we train the RF model on the training set, we evaluated the model's accuracy on a test set. It turns out that the accuracy of the model is quite low. In that case, MDA struggles because the accuracy difference between permuting a relevant feature and permuting a noisy feature is small. We observe that the MDA gets better when we increase the signal-to-noise ratio.

The computation time of different methods is hard to compare due to a few factors. Because the packages including `scikit-learn` and `ranger` compute feature importance when constructing the tree, it is hard to disentangle the time taken to construct the trees and the time taken to get the feature importance. Furthermore, different packages are implemented in different programming languages so it is not clear if the time difference is because of the algorithm or because of the language. We implement MDI-oob in Python and for our first simulated classification setting, MDI-oob takes ~ 3.8 seconds for each run. To compare, `scikit-learn` which uses Cython (A C extension for Python) takes ~ 1.4 seconds to construct the RFs for each run. Thus, MDI-oob runs in a reasonable time frame and we expect it to be faster if it is implemented in C or C++.

4.5 Discussion and future directions

Mean Decrease Impurity (MDI) is widely used to assess feature importance and its bias in feature selection is well known. Based on the original definition of MDI, we show that its expected bias is upper bounded by an expression that is inversely proportional to the minimum leaf size under mild conditions, which means deep trees generally have a higher feature selection bias than shallow trees. To reduce the bias, we derive a new analytical expression for MDI and use the new expression to obtain MDI-oob. For the simulated data and a genomic ChIP dataset, MDI-oob has exhibited the state-of-the-art feature selection performance in terms of AUC scores.

Table 4.1: Average AUC scores for noisy feature identification

	Deep tree (min leaf size = 1)				Shallow tree(min leaf size = 100)			
	Simulated		ChIP		Simulated		ChIP	
	C	R	C	R	C	R	C	R
MDI-oob	**0.76**	0.52	0.87	0.98	**0.75**	**0.58**	**0.94**	0.98
UFI	0.72	**0.54**	**0.88**	**0.99**	**0.75**	0.56	**0.94**	0.98
naive-oob	0.18	0.10	0.67	0.71	0.60	0.39	0.89	0.97
SHAP	0.55	0.33	0.82	0.96	0.68	0.46	0.91	0.97
ranger	0.56	0.50	0.73	0.97	0.55	0.49	0.76	**0.99**
MDA	0.49	0.51	0.54	0.97	0.50	**0.58**	0.50	**0.99**
cforest	0.65	0.50	0.79	0.93	0.70	0.49	0.90	0.98
MDI	0.12	0.09	0.60	0.71	0.63	0.40	0.88	0.97

"C" stands for classification, "R" stands for regression. The column maximum is bolded.

Comparison to SHAP. SHAP originates from game theory and offers a novel perspective to analyze the existing methods. While it is desirable to have 'consistency, missingness and local accuracy', our analysis indicates that there are other theoretical properties that are also worth taking into account. As shown in our simulation, the feature selection bias of SHAP increases with the depth of the tree, and we believe SHAP can also use OOB samples to improve feature selection performance.

Relationship to honest estimation. Honest estimation is an important technique built on the core notion of sample splitting. It has been successfully used in causal inference and other areas to mitigate the concern of over-fitting in complex learners due to usage of same data in different stages of training. The proposed algorithm MDI-oob has important connections with "honest sampling" or "honest estimation". For example, in Breiman's 1984 book [20], he proposed to use a separate validation set for pruning and uncertainty estimation. In [125], each within-leaf prediction is estimated using a different sub-sample (such as OOB sample) than the one used to decide split points. Theoretical results of these papers and Proposition 1 of our paper convey the same message, that finite sample bias is caused by using the same data for growing trees and for estimation, and the bias can be reduced if we leverage OOB data. We believe the theoretical contributions of those papers can also help us analyze the statistical properties (such as variance) of the MDI-oob.

Future directions. Although the MDI-oob shows promising results for selecting relevant features, it also raises many interesting questions to be considered in the future. First of all, how can MDI-oob be extended to better accommodate correlated features? Going beyond feature selection, can importance measures also rank the relevant features in a reasonable order? Finally, can we use the new analytical expression of MDI to give a tighter theoretical bound for MDI's feature selection bias? We are exploring these interesting questions in our ongoing work.

4.6 Proofs

Proof of Theorem 1. To state the proof of the theorem, we need to define more notations. For a generic set $A \subset [0,1]^p$, with slight abuse of notations, let $N_n(A) = \sum_i \mathbb{1}(\mathbf{x}_i \in A)$ be the number of samples with input features in A, and

$$\mu_n(A) = \frac{\sum_{\mathbf{x}_i \in A} y_i}{N_n(A)}$$

be the average response of those samples. For any feature X_k and $z \in (0,1)$, let $\Delta_{\mathcal{I}}(A,(k,z))$ be the impurity decrease when splitting A into $A \cap \{X_k \leq z\}$ and $A \cap \{z < X_k\}$, and $\Delta_{\mathcal{I}}(A,k) = \sup_{0 \leq z \leq 1} \Delta_{\mathcal{I}}(A,(k,z))$.

The proof of the theorem proceeds in three parts. First, we prove a lemma which gives a tail bound for $\Delta_{\mathcal{I}}(A,k)$. Second, we use the lemma and union bound to derive the upper bound for the expectation of $G_0(T)$. Finally, we use a separate argument based on Gaussian comparison inequalities to obtain the lower bound.

Lemma 1. *For any axis-aligned hyper-rectangle $A \subset [0,1]^p$, $k \notin S$ and $\delta > 0$, we have*

$$\mathbb{P}_{X,\epsilon}(\Delta_{\mathcal{I}}(A,k) \geq \delta \big| N_n(A)) \leq 4N_n(A) e^{-\frac{\delta N_n(A)}{4(M+1)^2}}.$$

Proof of Lemma 1. We suppose without loss of generality that $\mathbf{x}_1, \ldots, \mathbf{x}_{N_n(A)} \in A$. For any $z \in [0,1]$, we let

$$A^{\text{left}} = A \cap \{0 \leq X_k \leq z\}, \quad A^{\text{right}} = A \cap \{z < X_k \leq 1\},$$

and introduce the shorthands

$$p^{\text{left}} = \frac{N_n(A^{\text{left}})}{N_n(A)}, \quad p^{\text{right}} = \frac{N_n(A^{\text{right}})}{N_n(A)}, \quad \mu^{\text{left}} = \mu_n(A^{\text{left}}), \quad \mu^{\text{right}} = \mu_n(A^{\text{right}}).$$

Then

$$\begin{aligned}
\Delta_{\mathcal{I}}(A,(k,z)) =& \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A}(y_i-\mu_n(A))^2 - \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A}(y_i-\mu_n(A^{\text{left}}))^2\,\mathbb{1}(x_{ik}\le z)\\
&- \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A}(y_i-\mu_n(A^{\text{right}}))^2\,\mathbb{1}(x_{ik}> z)\\
=& \frac{1}{N_n(A)}\sum_{\mathbf{x}_i\in A} y_i^2 - \mu_n(A)^2 - p^{\text{left}}\Big(\frac{1}{N_n(A)p^{\text{left}}}\sum_{\mathbf{x}_i\in A} y_i^2\,\mathbb{1}(x_{ik}\le z) - (\mu^{\text{left}})^2\Big)\\
&- p^{\text{right}}\Big(\frac{1}{N_n(A)p^{\text{right}}}\sum_{\mathbf{x}_i\in A} y_i^2\,\mathbb{1}(x_{ik}> z) - (\mu^{\text{right}})^2\Big)\\
=& p^{\text{left}}(\mu^{\text{left}})^2 + p^{\text{right}}(\mu^{\text{right}})^2 - \mu_n(A)^2\\
=& (p^{\text{left}}(\mu^{\text{left}})^2 + p^{\text{right}}(\mu^{\text{right}})^2)(p^{\text{left}}+p^{\text{right}}) - (p^{\text{left}}\mu^{\text{left}} + p^{\text{right}}\mu^{\text{right}})^2\\
=& p^{\text{left}}p^{\text{right}}(\mu^{\text{left}}-\mu^{\text{right}})^2\\
\le& 2p^{\text{left}}p^{\text{right}}[(\mu^{\text{left}}-\mu)^2 + (\mu^{\text{right}}-\mu)^2]\\
\le& 2p^{\text{left}}(\mu^{\text{left}}-\mu)^2 + 2p^{\text{right}}(\mu^{\text{right}}-\mu)^2,
\end{aligned}$$

where

$$\mu = \mathbb{E}[Y|X\in A] = \mathbb{E}[\phi(X)|X\in A].$$

Now suppose without loss of generality that $x_{1k} < x_{2k} < \cdots < x_{nk}$ (otherwise we can reorder the samples by X_k). Since $k\notin S$, X_k is independent of X_S and therefore independent of Y. Thus the distribution of $(y_1,\ldots,y_n)$ does not change after the reordering, i.e.,

$$y_i \overset{i.i.d}{\sim} (\phi(X)|X\in A) + \epsilon.$$

Note that

$$\sup_z\; p^{\text{left}}(\mu^{\text{left}}-\mu)^2 \le \sup_{1\le m\le N_n(A)} \frac{m}{N_n(A)}\left(\frac{1}{m}\sum_{i=1}^m y_i - \mu\right)^2.$$

Note that Y is sub-Gaussian with parameter $M+1$. Therefore, for each $1\le m\le N_n(A)$, by Hoeffding bound,

$$\mathbb{P}\left(\frac{m}{N_n(A)}\left(\frac{1}{m}\sum_{i=1}^m y_i-\mu\right)^2 \ge \delta \,\middle|\, N_n(A)\right) \le 2e^{-(M+1)^2\delta N_n(A)^2/m} \le 2e^{-\frac{\delta N_n(A)}{(M+1)^2}}.$$

Therefore

$$\mathbb{P}\left(\sup_z p^{\text{left}}(\mu^{\text{left}}-\mu)^2 \ge \delta \,\middle|\, N_n(A)\right) \le 2N_n(A)e^{-\frac{\delta N_n(A)}{(M+1)^2}}.$$

By symmetry, the same bound holds for $p^{\text{right}}(\mu^{\text{right}} - \mu)^2$. Therefore

$$\begin{aligned}
&\mathbb{P}(\Delta_{\mathcal{I}}(A,k) \geq \delta \big| N_n(A)) \\
\leq &\mathbb{P}\left(\sup_z p^{\text{left}}(\mu^{\text{left}} - \mu)^2 \geq \delta/2 \big| N_n(A)\right) + \mathbb{P}\left(\sup_z p^{\text{right}}(\mu^{\text{right}} - \mu)^2 \geq \delta/2 \big| N_n(A)\right) \\
\leq &4N_n(A)e^{-\frac{\delta N_n(A)}{4(M+1)^2}},
\end{aligned}$$

and the lemma is proved. □

Proof of the upper bound in Theorem 1

Without loss of generality, assume that when we split on feature k, the cut is always performed along the direction of k at some data point (and that data point falls into the right sub-tree). Suppose that ϵ_i has unit variance for all i. Let $C = 2\max\{256, 16(M+1)^2\}$. We also assume, without loss of generality, that $m_n \geq 8d_n$. Otherwise, since $G_0(T)$ is, by definition, upper bounded by the sample variance of y, we have

$$\mathbb{E}_{X,\epsilon} \sup_{T \in \mathcal{T}_n(m_n, d_n)} G_0(T) \leq \text{Var}(Y) \leq M^2 + 1 \leq 16(M+1)^2 \frac{d_n \log np}{m_n}.$$

To simplify notation, we define $\mathbf{x}_{n+1} = (0, \ldots, 0)$ and $\mathbf{x}_{n+2} = (1, \ldots, 1)$. For any $V \subset [p], \mathcal{L}, \mathcal{R} \in [n+2]^{|V|}$, let

$$A(V, \mathcal{L}, \mathcal{R}) = \{X = (X_1, \ldots, X_p) : x_{\mathcal{L}_i, V_i} \leq X_{V_i} < x_{\mathcal{R}_i, V_i}, 1 \leq i \leq |V|, 0 \leq X_k \leq 1, k \notin V\}$$

be the random axis-aligned hyper-rectangle obtained by splitting on features in V, where the left and right endpoints of the ith feature V_i are determined by $x_{\mathcal{L}_i, V_i}$ and $x_{\mathcal{R}_i, V_i}$. Note that in this definition, we treat $\mathbf{x}_i$ as random variables rather than fixed, and $A(V, \mathcal{L}, \mathcal{R})$ can be the empty set with non-zero probability. Let

$$A(V) = \{A(V, \mathcal{L}, \mathcal{R}) | \mathcal{L}, \mathcal{R} \in [n+2]^{|V|}\}$$

be all axis-aligned hyper-rectangles obtained by splitting on features in V. For any $d \leq d_n$, let

$$A_d = \cup_{|V|=d} A(V)$$

be the collection of all possible subsets of $[0,1]^p$ obtained by splitting on d features.

Fix $\delta > \frac{96M^2 d_n}{m_n}$. We will first show that

$$\begin{aligned}
&\mathbb{P}_{X,\epsilon}\left(\exists A \in A_d, k \notin S : \Delta_{\mathcal{I}}(A,k) \geq \frac{m_n \delta}{N_n(A)} \text{ and } N_n(A) \geq m_n\right) \\
\leq &5(np)^{d+1} \exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right).
\end{aligned} \tag{4.15}$$

Note that for any two events C_1 and C_2, the inequality $\mathbb{P}(C_1 \cap C_2) \leq \mathbb{P}(C_1|C_2)$ always holds. Therefore, for any hyper-rectangle A, we have

$$\begin{aligned}&\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)} \text{ and } N_n(A) \geq m_n\right)\\ \leq &\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)}\middle| N_n(A) \geq m_n\right)\end{aligned} \tag{4.16}$$

To simplify notation, we will drop the conditional event $N_n(A) \geq m_n$ in the remainder of the proof of the upper bound, unless stated otherwise.

Fix $V \subset [p], \mathcal{L}, \mathcal{R} \in [n+2]^{|V|}$, and $k \notin S$. Conditional on samples in $\mathcal{L}$ and $\mathcal{R}$, we would like to apply Lemma 1 to $A(V,\mathcal{L},\mathcal{R})$ and k. The only problem is that there are now samples on the boundary of $A(V,\mathcal{L},\mathcal{R})$, namely those in $\mathcal{L}$ and $\mathcal{R}$. Let $\mathbf{x}_{\mathcal{L}} = \{\mathbf{x}_i\}_{i\in\mathcal{L}}$ and $\mathbf{x}_{\mathcal{R}} = \{\mathbf{x}_i\}_{i\in\mathcal{R}}$. Conditional on $\mathbf{x}_{\mathcal{L}}$, $\mathbf{x}_{\mathcal{R}}$ and $N_n(A(V,\mathcal{L},\mathcal{R}))$, and on the random variable $X \in A(V,\mathcal{L},\mathcal{R})$, X is uniformly distributed in $A(V,\mathcal{L},\mathcal{R})$. For a set A, we let A° be the interior of A and let $\bar{A}$ be the boundary of A. Since $m_n \geq 8d_n$,

$$\frac{N_n(A^\circ(V,\mathcal{L},\mathcal{R}))}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n - 2d_n}{m_n} \geq \frac{3}{4}.$$

By Lemma 1, we have

$$\begin{aligned}&\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\middle|\mathbf{x}_{\mathcal{L}},\mathbf{x}_{\mathcal{R}},N_n(A(V,\mathcal{L},\mathcal{R}))\right)\\ \leq &4N_n(A^\circ(V,\mathcal{L},\mathcal{R}))\exp\left(-\frac{\delta m_n N_n(A^\circ(V,\mathcal{L},\mathcal{R}))}{12(M+1)^2 N_n(A(V,\mathcal{L},\mathcal{R}))}\right)\\ \leq &4n\exp\left(-\frac{\delta m_n}{16(M+1)^2}\right)\end{aligned} \tag{4.17}$$

for large n. Since the right hand side does not depend on $\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{R}}, N_n(A(V,\mathcal{L},\mathcal{R}))$, we can take expectation with respect to them, and obtain

$$\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right) \leq 4n\exp\left(-\frac{\delta m_n}{16(M+1)^2}\right) \tag{4.18}$$

On the other hand, we have the inequality

$$\begin{aligned}\Delta_{\mathcal{I}}(A(V,\mathcal{L},\mathcal{R}),k) &\leq \Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) + \frac{\sum_{i\in\mathcal{L},\mathcal{R}}(y_i - \mu_n(A(V,\mathcal{L},\mathcal{R})))^2}{N_n(A(V,\mathcal{L},\mathcal{R}))}\\ &\leq \Delta_{\mathcal{I}}(A^\circ(V,\mathcal{L},\mathcal{R}),k) + \frac{\sum_{i\in\mathcal{L},\mathcal{R}} 2(y_i^2 + \mu_n(A(V,\mathcal{L},\mathcal{R}))^2)}{N_n(A(V,\mathcal{L},\mathcal{R}))}.\end{aligned} \tag{4.19}$$

We have

$$\begin{aligned}
&\mathbb{P}_{X,\epsilon}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 2y_i^2}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 4(f^2(\mathbf{x}_i)+\epsilon_i^2)}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 4(f^2(\mathbf{x}_i)+\epsilon_i^2)}{m_n} \geq \frac{\delta}{3}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 4M^2+4\epsilon_i^2}{m_n} \geq \frac{\delta}{3}\right)\\
\leq&\mathbb{P}\left(\frac{\sum_{i=1}^{2d_n}(\epsilon_i^2-1)}{m_n} \geq \frac{\delta}{16N_n(A^\circ(V,\mathcal{L},\mathcal{R})}\right) \leq \exp(-\frac{\delta m_n}{256}),
\end{aligned} \tag{4.20}$$

for large n, where the fourth inequality holds because $\delta \geq 96M^2d_n/m_n$, and the last inequality follows from the well-known tail bound

$$\mathbb{P}\left(\left|\frac{1}{d}\chi_d^2 - 1\right| \geq \delta_0\right) \leq 2e^{-d\delta_0^2/8}$$

for χ_d^2 random variable and $\delta_0 < 1$. To upper bound $\mu_n(A(V,\mathcal{L},\mathcal{R}))$, note that

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sum_{i\in\mathcal{L},\mathcal{R}} 2\mu_n(A(V,\mathcal{L},\mathcal{R}))^2}{N_n(A(V,\mathcal{L},\mathcal{R}))} \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right)\\
\leq&\mathbb{P}\left(|\mu_n(A(V,\mathcal{L},\mathcal{R}))| \geq \sqrt{\frac{\delta m_n}{6d_n}}\right)\\
\leq&\mathbb{P}\left(\left|\frac{1}{N_n(A(V,\mathcal{L},\mathcal{R}))}\sum_{i=1}^{N_n(A(V,\mathcal{L},\mathcal{R}))}\epsilon_i\right| \geq \sqrt{\frac{\delta m_n}{6d_n}} - M\right)\\
\leq&2\exp\left(-\frac{1}{2}m_n(\sqrt{\frac{\delta m_n}{6d_n}} - M)^2\right)\\
\leq&2\exp\left(-\frac{\delta m_n}{4}\right),
\end{aligned} \tag{4.21}$$

where the last inequality follows from $m_n \geq 8d_n$ and $\delta \geq 96M^2d_n/m_n$. Combining Equations (4.18), (4.19), (4.21), we have

$$\mathbb{P}_{X,\epsilon}\left(\Delta_{\mathcal{I}}(A(V,\mathcal{L},\mathcal{R}),k) \geq \frac{m_n\delta}{3N_n(A(V,\mathcal{L},\mathcal{R}))}\right) \leq 5n\exp\left(-\frac{\delta m_n}{\max\{16(M+1)^2,256\}}\right) \tag{4.22}$$

for any $V \subset [p], |V| = d, \mathcal{L},\mathcal{R} \in [n+2]^{|V|}$, and $k \notin S$. Note that the set A_d has cardinality

$$|A_d| = \binom{p}{d}(2(n+2))^d \leq \left(\frac{pn}{d}\right)^d$$

for large n. Therefore by union bound,

$$\begin{aligned}\mathbb{P}\left(\exists A \in A_d, k \notin S : \Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)}\right) &\leq 5np|A_d|\exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right) \\ &\leq 5(np)^{d+1}\exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right).\end{aligned} \tag{4.23}$$

Suppose that $\Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)}$ for all $A \in \cup_{d\leq d_n} A_d$ and $k \notin S$, then for any $T \in T_n(m_n, d_n)$,

$$G_0(T) \leq \sum_{t:v(t)\notin S} \frac{N_n(t)}{n}\frac{m_n\delta}{N_n(t)} \leq \delta\frac{m_n|I(t)|}{n} \leq \delta,$$

where the last inequality follows since $|I(t)|+1$ is the total number of leaf nodes in T, and each leaf node contains at least m_n samples. Therefore

$$\begin{aligned}\mathbb{P}_{X,\epsilon}\left(\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T) \geq \delta\right) &\leq \sum_{d=1}^{d_n}\mathbb{P}\left(\exists A \in A_d, k \notin S : \Delta_{\mathcal{I}}(A,k) \geq \frac{m_n\delta}{N_n(A)}\right) \\ &\leq \sum_{d=1}^{d_n} 5(np)^{d+1}\exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right) \\ &\leq 10(np)^{d_n+1}\exp\left(-\frac{\delta m_n}{\max\{256, 16(M+1)^2\}}\right)\end{aligned} \tag{4.24}$$

for any $\delta > \frac{96M^2d_n}{m_n}$. Recall that $C = 2\max\{256, 16(M+1)^2\}$. Note that $\frac{Cd_n\log(np)}{m_n} \geq \frac{96M^2d_n}{m_n}$ for large n. Integrating over δ, we have

$$\begin{aligned}&\mathbb{E}_{X,\epsilon}\left[\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T)\right] \\ &\leq \frac{3d_n\log(np)}{2m_n} + \mathbb{E}_{X,\epsilon}\left[\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T)\mathbb{1}(\delta \geq \frac{3d_n\log(np)}{2m_n})\right] \\ &\leq \frac{3d_n\log(np)}{2m_n} + \int_{\frac{3d_n\log(np)}{2m_n}}^{\infty}\mathbb{P}_{X,\epsilon}\left(\sup_{T\in\mathcal{T}_n(m_n,d_n)} G_0(T) \geq \delta\right)d\delta \\ &\leq \frac{Cd_n\log(np)}{m_n}.\end{aligned} \tag{4.25}$$

This completes the proof of the upper bound.

Proof of the lower bound in Theorem 1

For the lower bound, let

$$d_n = \max\{d : 2^{d+1}m_n < n\}, \tag{4.26}$$

and consider a balanced, binary decision tree T constructed in the following way:

1. At each node on the first $d_n - 1$ levels of the tree, we split on feature X_1, at the mid-point of X_1's side of the rectangle corresponding to the node.

2. At each node on the d_nth level, we look at the remaining $p - 1$ features, and split on the feature that maximizes the decrease in impurity.

In the following proof, we will lower bound $G_0(T)$ by the sum of impurity reduction on the d_nth level alone. For $t = 1, \ldots, 2^{d_n-1}$, let

$$R_t = \left\{ \frac{t-1}{2^{d_n-1}} \leq X_1 < \frac{t}{2^{d_n-1}} \right\}.$$

be the hyper-rectangle corresponding to the tth node on the d_nth level. Applying Chernoff's inequality, we have

$$\mathbb{P}\left(\left| \frac{N_n(R_t)}{n} - \frac{1}{2^{d_n-1}} \right| \geq \frac{1}{3 \cdot 2^{d_n-1}} \right) \leq 2 \exp\left(- \frac{n}{27 \cdot 2^{d_n-1}} \right).$$

Let

$$B_1 = \left\{ \left| \frac{N_n(R_t)}{n} - \frac{1}{2^{d_n-1}} \right| \leq \frac{1}{3 \cdot 2^{d_n-1}} \text{ for all } t \right\}$$

be the event that each node on the d_nth level contains at least $\frac{2}{3}\frac{n}{2^{d_n-1}}$, but no more than $\frac{4}{3}\frac{n}{2^{d_n-1}}$ samples. Then

$$\mathbb{P}(B_1^c) \leq \sum_{t=1}^{2^{d_n-1}} \mathbb{P}\left(\left| \frac{N_n(R_t)}{n} - \frac{1}{2^{d_n-1}} \right| \geq \frac{1}{3 \cdot 2^{d_n-1}} \right) \leq 2^{d_n} \exp\left(- \frac{n}{27 \cdot 2^{d_n-1}} \right), \tag{4.27}$$

and conditional on B_1,

$$\frac{8}{3} m_n \leq \frac{2}{3}\frac{n}{2^{d_n-1}} \leq N_n(R_t) \leq \frac{4}{3}\frac{n}{2^{d_n-1}} \leq \frac{32}{3} m_n. \tag{4.28}$$

We define

$$R_t^l(k) = R_t \cap \left\{ 0 \leq X_k < \frac{1}{2} \right\}$$

and

$$R_t^r(k) = R_t \cap \left\{ \frac{1}{2} \leq X_k < 1 \right\}$$

and use R_t^l, R_t^r as shorthand when k is fixed. For each $t = 0, 1, \ldots, 2^d - 1$, by Equation

$$\Delta_{\mathcal{I}}(R_t, k) \geq \Delta_{\mathcal{I}}(R_t, (k, 1/2)) = \frac{N_n(R_t^l)}{N_n(R_t)} \frac{N_n(R_t^r)}{N_n(R_t)} (\mu_n(R_t^l) - \mu_n(R_t^r))^2$$

Let

$$\eta_k = \mu_n(R_t^l) - \mu_n(R_t^r)$$

Conditional on $N_n(R_t^l)$ and $N_n(R_t^r)$, $\eta = (\eta_2, \ldots, \eta_p)$ are jointly Gaussian with zero mean. To lower bound the impurity decrease at the tth node on the d_nth level, we use a Gaussian comparison argument to obtain a lower bound for $\sup_k |\eta_k|$, which requires us to calculate the covariance matrix of η. For any $2 \le k_1, k_2 \le p$, let us further define

$$R_t^{ll}(k_1,k_2) = R_t \cap \left\{0 \le X_{k_1} < \frac{1}{2}\right\} \cap \left\{0 \le X_{k_2} < \frac{1}{2}\right\};$$

$$R_t^{lr}(k_1,k_2) = R_t \cap \left\{0 \le X_{k_1} < \frac{1}{2}\right\} \cap \left\{\frac{1}{2} \le X_{k_2} < 1\right\};$$

$$R_t^{rl}(k_1,k_2) = R_t \cap \left\{\frac{1}{2} \le X_{k_1} < 1\right\} \cap \left\{0 \le X_{k_2} < \frac{1}{2}\right\};$$

$$R_t^{rr}(k_1,k_2) = R_t \cap \left\{\frac{1}{2} \le X_{k_1} < 1\right\} \cap \left\{\frac{1}{2} \le X_{k_2} < 1\right\}.$$

As before, we write $R_t^{ll}, R_t^{lr}, R_t^{rl}$ and R_t^{rr} as shorthand when k_1, k_2 are fixed. Conditional on $N_n(R_t)$, the samples falling into the hyper-rectangle R_t are uniformly distributed in R_t. Therefore we know from Chernoff's inequality that

$$\mathbb{P}\left(\left|\frac{N_n(R_t^{ll})}{N_n(R_t)} - \frac{1}{4}\right| \ge \frac{1}{16}\right) \le 2\exp\left(-\frac{N_n(R_t)}{48}\right)$$

for any k_1 and k_2, and that the same results hold for R_t^{lr}, R_t^{rl} and R_t^{ll} as well. Let

$$B_2 = \left\{\max_{\omega \in \{ll,lr,rl,rr\}} \left|\frac{N_n(R_t^{\omega}(k_1,k_2))}{N_n(R_t)} - \frac{1}{4}\right| \le \frac{1}{16}, \text{ for all } 1 \le t \le 2^{d_n-1}, 2 \le k_1 < k_2 \le p\right\}.$$

Then

$$\mathbb{P}(B_2^c) \le 2^{d_n} p^2 \exp\left(-\frac{N_n(R_t)}{48}\right), \tag{4.29}$$

and

$$\mathbb{P}(B_1 \cap B_2) \ge 1 - 2^{d_n+1} p^2 \exp\left(-\frac{N_n(R_t)}{48}\right) \ge 1 - 2^{d_n+1} p^2 \exp\left(-\frac{m_n}{18}\right) \ge \frac{8}{9} \tag{4.30}$$

for n large enough (under the condition that $m_n \ge 36\log p + 18\log n$). Conditional on the event B_2,

$$N_n(R_t^l) \ge N_n(R_t^{ll}) + N_n(R_t^{lr}) \ge \frac{3}{16}N_n(R_t) + \frac{3}{16}N_n(R_t) \ge \frac{3}{8}N_n(R_t),$$

for any $1 \le t \le 2^{d_n-1}$ and $2 \le k \le p$, and the same holds for $N_n(R_t^r)$. Therefore,

$$\text{Var}(\eta_k) = \frac{1}{N_n(R_t^l)} + \frac{1}{N_n(R_t^r)} \ge \frac{3}{4N_n(R_t)} \tag{4.31}$$

$$\text{Cov}(\eta_{k_1}, \eta_{k_2}) = \frac{1}{N_n(R_t^{ll})} + \frac{1}{N_n(R_t^{rr})} - \frac{1}{N_n(R_t^{lr})} - \frac{1}{N_n(R_t^{rl})} \leq \frac{1}{4N_n(R_t)}. \tag{4.32}$$

Consider $\tilde{\eta}_2, \ldots, \tilde{\eta}_p$ with

$$\mathbb{E}\tilde{\eta}_k = 0, \text{Var}(\tilde{\eta}_k) = \frac{3}{4N_n(R_t)}$$

and

$$\text{Cov}(\tilde{\eta}_{k_1}, \tilde{\eta}_{k_2}) = \frac{1}{4N_n(R_t)}.$$

Then conditional on $B_1 \cap B_2$, by Sudakov-Fernique lemma, we have

$$\mathbb{E}_\epsilon[\max_k \eta_k | B_1 \cap B_2] \geq \mathbb{E} \max_k \tilde{\eta}_k \geq \sqrt{\frac{\log p}{N_n(R_t)}} \geq \sqrt{\frac{3 \log p}{32 m_n}},$$

and the lower bound

$$\min\{N_n(R_t^l), N_n(R_t^r)\} \geq \frac{3}{8} N_n(R_t) \geq m_n,$$

for any k, t. where the last inequality follows from Equation (4.28). Therefore, conditional on $B_1 \cap B_2$ the minimum leaf size is lower bounded by m_n. Finally

$$\begin{aligned}
\mathbb{E}_{X,\epsilon}\left[\sup_{T \in \mathcal{T}_n(m_n)} G_0(T)\right] &\geq \mathbb{E}_{X,\epsilon}\left[\sup_{T \in \mathcal{T}_n(m_n)} G_0(T) \mathbb{1}_{B_1 \cap B_2}\right] \\
&\geq \mathbb{E}_X\left[\sum_t \frac{N_n(R_t)}{n} \mathbb{E}_\epsilon\left[\max_k \Delta_{\mathcal{I}}(R_t, k) \mathbb{1}_{B_1 \cap B_2}\right]\right] \\
&\geq \mathbb{E}_X \sum_t \frac{N_n(R_t)}{n} (\frac{3}{8})^2 (\mathbb{E}_\epsilon \max_k \eta_k^2 \mathbb{1}_{B_1 \cap B_2}) \\
&\geq \frac{9}{64} \frac{3 \log p}{32 m_n} \mathbb{P}(B_1 \cap B_2) \\
&\geq \frac{1}{80} \frac{\log p}{m_n}
\end{aligned} \tag{4.33}$$

when n is large enough, and the lower bound is proved. This concludes the whole proof.

□

Proof of Proposition 1. For simplicity, here we only present the proof for a single tree T. The case of multiple trees is straightforward. Recall that t^{left} and t^{right} are the left and right children of the node t. Based on (4.4), MDI at the node t is

$$\begin{aligned}
\frac{N_n(t)}{|\mathcal{D}^{(T)}|} \Delta_{\mathcal{I}}(t) = \frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} &[y_i - \mu_n(t)]^2 \mathbb{1}(\mathbf{x}_i \in R_t) \\
&- [y_i - \mu_n(t^{\text{left}})]^2 \mathbb{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) - [y_i - \mu_n(t^{\text{right}})]^2 \mathbb{1}(\mathbf{x}_i \in R_{t^{\text{right}}}).
\end{aligned} \tag{4.34}$$

Because $\mathbb{1}(\mathbf{x}_i \in R_t) = \mathbb{1}(\mathbf{x}_i \in R_{t^{\text{right}}}) + \mathbb{1}(\mathbf{x}_i \in R_{t^{\text{left}}})$, the above term becomes

$$\begin{aligned}
&\frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} \left((y_i - \mu_n(t))^2 - (y_i - \mu_n(t^{\text{left}}))^2\right) \mathbb{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) \\
&\qquad + \left((y_i - \mu_n(t))^2 - (y_i - \mu_n(t^{\text{right}}))^2\right) \mathbb{1}(\mathbf{x}_i \in R_{t^{\text{right}}}) \\
=&\frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} (\mu_n(t^{\text{left}}) - \mu_n(t))(2y_i - \mu_n(t) - \mu_n(t^{\text{left}}))\mathbb{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) \\
&\qquad + (\mu_n(t^{\text{right}}) - \mu_n(t))(2y_i - \mu_n(t) - \mu_n(t^{\text{right}}))\mathbb{1}(\mathbf{x}_i \in R_{t^{\text{right}}}). \qquad (4.35)
\end{aligned}$$

Since $\sum_{i \in \mathcal{D}^{(T)}} y_i \mathbb{1}(\mathbf{x}_i \in t^{\text{left}}) = N_n(t^{\text{left}})\mu_n(t^{\text{left}})$, we know

$$\sum_{i \in \mathcal{D}^{(T)}} (y_i - \mu_n(t^{\text{left}}))\mathbb{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) = 0.$$

Similar equations hold for the right child t^{right}, too. Then (4.35) reduces to

$$\begin{aligned}
&\frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} (\mu_n(t^{\text{left}}) - \mu_n(t))(y_i - \mu_n(t))\mathbb{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) \qquad (4.36) \\
&\qquad + (\mu_n(t^{\text{right}}) - \mu_n(t))(y_i - \mu_n(t))\mathbb{1}(\mathbf{x}_i \in R_{t^{\text{right}}}) \qquad (4.37)
\end{aligned}$$

Because of the definitions of $\mu_n(t^{\text{left}})$, $\mu_n(t^{\text{right}})$, and $\mu_n(t)$, we know

$$N_n(t^{\text{left}})\mu_n(t^{\text{left}}) + N_n(t^{\text{right}})\mu_n(t^{\text{right}}) = N_n(t)\mu_n(t). \qquad (4.38)$$

That implies

$$\sum_{i \in \mathcal{D}^{(T)}} (\mu_n(t^{\text{left}}) - \mu_n(t))\mathbb{1}(\mathbf{x}_i \in R_{t^{\text{left}}}) + (\mu_n(t^{\text{right}}) - \mu_n(t))\mathbb{1}(\mathbf{x}_i \in R_{t^{\text{right}}}) = 0.$$

Using this equation, (4.37) can be written as

$$\frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} (\mu_n(t^{\text{left}}) - \mu_n(t))y_i\mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{left}}}) + (\mu_n(t^{\text{right}}) - \mu_n(t))y_i\mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{right}}}). \qquad (4.39)$$

In summary, we have shown that:

$$\begin{aligned}
&\frac{N_n(t)}{|\mathcal{D}^{(T)}|}\Delta_{\mathcal{I}}(t) \qquad (4.40) \\
=&\frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} (\mu_n(t^{\text{left}}) - \mu_n(t))y_i\mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{left}}}) + (\mu_n(t^{\text{right}}) - \mu_n(t))y_i\mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{right}}}). \qquad (4.41)
\end{aligned}$$

Since the MDI of the feature k is the sum of $\frac{N_n(t)}{|\mathcal{D}^{(T)}|}\Delta_{\mathcal{I}}(t)$ across all inner nodes such that $v(t) = k$, we have

$$
\begin{aligned}
&\sum_{t \in I(T)} \frac{N_n(t)}{|\mathcal{D}^{(T)}|} \Delta_{\mathcal{I}}(t) \mathbb{1}(v(t) = k) \\
=& \sum_{t:v(t)=k} \frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} (\mu_n(t^{\text{left}}) - \mu_n(t)) y_i \mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{left}}}) + (\mu_n(t^{\text{right}}) - \mu_n(t)) y_i \mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{right}}}) \\
=& \frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} \Big[\sum_{t:v(t)=k} (\mu_n(t^{\text{left}}) - \mu_n(t)) \mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{left}}}) + (\mu_n(t^{\text{right}}) - \mu_n(t)) \mathbb{1}(\mathbf{x}_i \in \mathbb{R}_{t^{\text{right}}}) \Big] y_i \\
=& \frac{1}{|\mathcal{D}^{(T)}|} \sum_{i \in \mathcal{D}^{(T)}} f_{T,k}(\mathbf{x}_i) y_i.
\end{aligned}
$$

That completes the proof. □

4.7 Supplementary Table and Figures

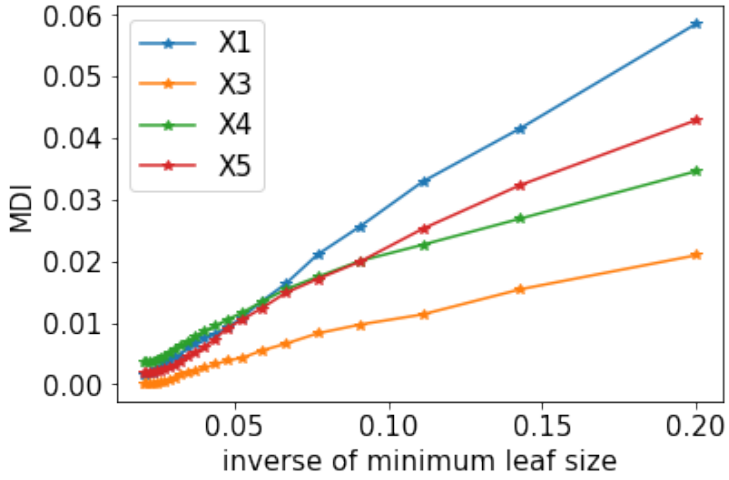


Figure 4.4: MDI against inverse min leaf size. This is coherent with our theoretical analysis as MDI is proportional to the inverse of minimum leaf size.

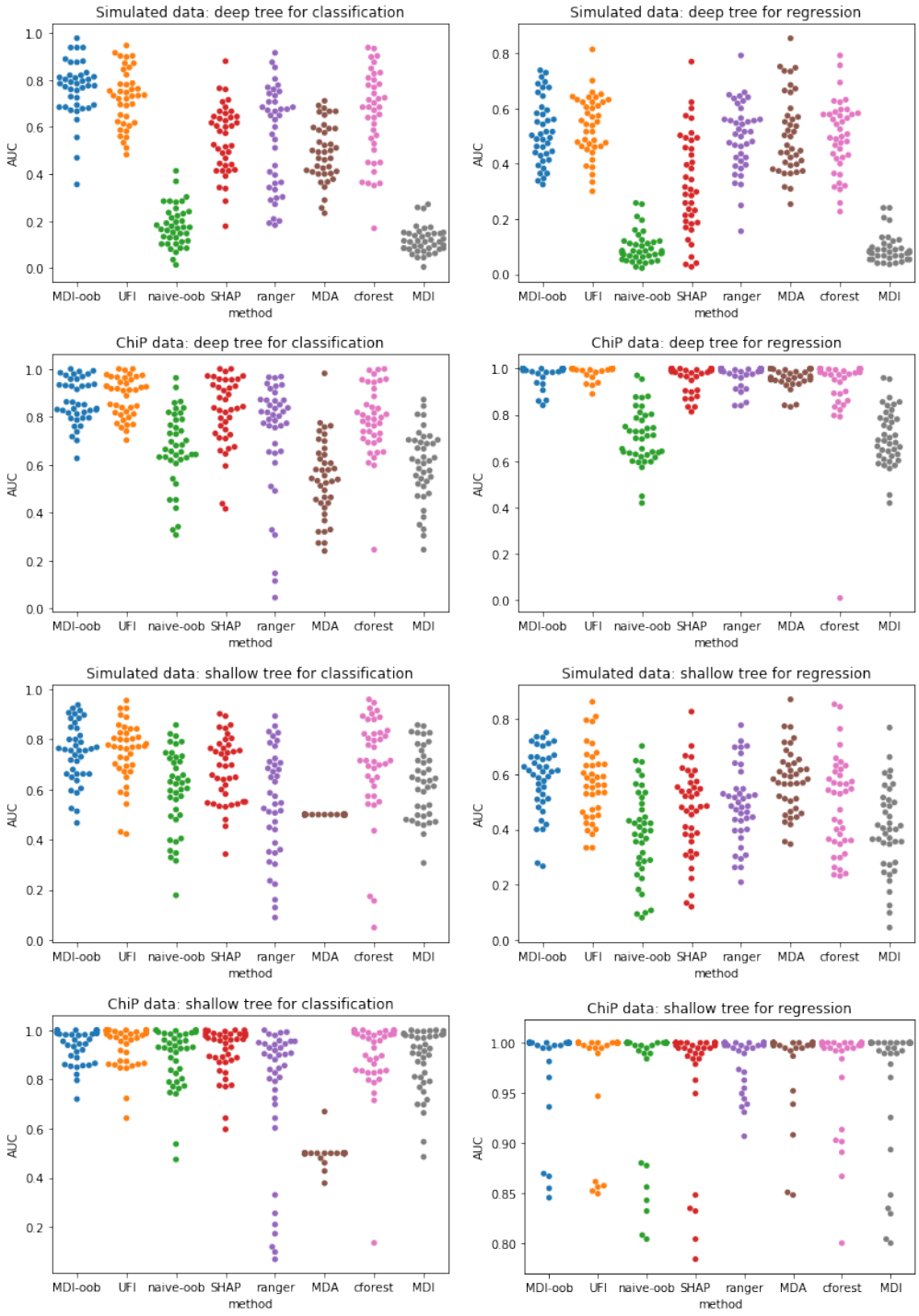


Figure 4.5: The beeswarm plots for different simulation settings.

	Deep tree (min leaf size = 1)				Shallow tree(min leaf size = 100)			
	Simulated		ChIP		Simulated		ChIP	
	C	R	C	R	C	R	C	R
MDI-oob	0.762(.019)	0.519(.018)	0.865(.015)	0.980(.006)	0.748(.019)	0.581(.019)	0.939(.011)	0.983(.007)
SHAP	0.548(.023)	0.325(.028)	0.821(.023)	0.963(.009)	0.677(.021)	0.462(.025)	0.912(.015)	0.972(.009)
ranger	0.555(.034)	0.496(.019)	0.726(.038)	0.974(.007)	0.549(.034)	0.487(.022)	0.755(.045)	0.985(.004)
MDA	0.493(.019)	0.507(.022)	0.542(.025)	0.966(.007)	0.500(.000)	0.577(.018)	0.498(.006)	0.986(.006)
cforest	0.649(.029)	0.499(.020)	0.788(.023)	0.929(.026)	0.701(.033)	0.488(.026)	0.900(.024)	0.979(.007)
MDI	0.118(.009)	0.092(.008)	0.597(.023)	0.706(.019)	0.632(.022)	0.397(.025)	0.877(.020)	0.971(.009)

"C" stands for classification, "R" stands for regression.

Table 4.2: Average AUC scores and standard deviations for noisy feature identification.

Chapter 5

Optimality of the max test for detecting sparse signals with Gaussian or heavier tail

5.1 Introduction

Sparse signal detection

Closely related to multiple testing is the problem of testing the *global null* or *intersection null*, which asserts that all of n univariate null hypotheses are true; this is sometimes called the *signal detection problem*, since it amounts to asking whether there is any signal at all. One strategy, popular among methodologists and practitioners alike for its simplicity, transparency, and robustness, is to reject when the largest univariate test statistic is above a critical threshold, or equivalently when the smallest univariate p-value is below an appropriately corrected significance level. This method, called the *max test*, is closely associated with the multiple testing procedure that rejects individual hypotheses with p-values below the same threshold, which is $1-(1-\alpha)^{1/n}$ if the p-values are independent (called the *Šidák correction*), or α/n if the dependence structure is completely unknown (the *Bonferroni correction*), and may be obtained by simulation in other cases [112]. Because the associated multiple testing procedure controls the familywise error rate (FWER), the max test can be tacked on as a logical deduction about the global null, incurring no additional FWER.

However, the adequacy of the max test for signal detection has been placed in doubt because it does not always achieve an optimal detection boundary in the *Gaussian sequence model* where we observe $X \sim N_n(\mu, I_n)$, a canonical testing ground for high-dimensional statistical methods. In certain sparse asymptotic regimes of this model, the max test is outperformed by more sophisticated special-purpose tests of the global null $H_0:\ \mu_i = 0$ for all i, against $H_1:\ \mu_i \neq 0$ for some i.

Most notably, Donoho and Jin [35, 36] compared the max test to the *higher criticism* (HC) test, which rejects the global null for large values of Tukey's higher criticism statistic

$$HC_n = \sup_{1 \le i \le n/2} \frac{\sqrt{n}(i/n - p_{(i)})}{\sqrt{p_{(i)}(1-p_{(i)})}} = \sup_{0 \le t \le 1/2} \frac{\sqrt{n}(\widehat{F_n}(t) - t)}{\sqrt{t(1-t)}},$$

where $p_{(1)} \leq \cdots \leq p_{(n)}$ are the ordered p-values and $\widehat{F_n}(t)$ is their empirical distribution function. They also studied two related tests: the *modified higher criticism* test, which rejects for large values of

$$mHC_n = \sup_{1/n \leq t \leq 1/2} \frac{\sqrt{n}(\widehat{F_n}(t) - t)}{\sqrt{t(1-t)}},$$

and the *Berk-Jones test*, which rejects for large values of

$$BJ_n = \max_{1 \leq k \leq n/2} (2n)^{1/2} \left\{ \frac{k}{n} \log\left(\frac{k}{np_{(k)}}\right) + \left(1 - \frac{k}{n}\right) \log\left(\frac{n-k}{n(1-p_{(k)})}\right) \right\}^{1/2}.$$

They showed, in a model where all nonzero μ_i take the same value, that the higher criticism, modified higher criticism, and Berk-Jones tests all achieve the optimal detection boundary in the sparse asymptotic regime where the number n_1 of nonzero signals grows more slowly than $n^{1/2}$ (for denser signals, the χ^2 test is typically much more powerful than all tests under comparison here). By contrast, the max test falls short unless $n_1 = O(n^{1/4})$. In light of these results, it has been widely accepted as a stylized fact that these special-purpose tests dominate the max test for sparse signal detection.

While [35] provide a remarkably detailed and complete picture of global testing in the asymptotic regime they study, it is natural to ask how the story might change if we relax the rather restrictive assumption that all of the nonzero signals have identical strength, since in real applications we would expect these to vary in magnitude. This article considers a more general setting where the non-null signals are instead drawn from a distribution G_n:

$$\{\mu_i\}_{i=1}^n \overset{i.i.d}{\sim} (1-\pi_n)\delta_0(\cdot) + \pi_n G_n(\cdot), \quad \pi_n = n^{-\beta}, \quad 0 < \beta < 1. \tag{5.1}$$

This model was previously studied by Cai and Wu [24], who showed under certain regularity conditions in the sparse regime $\beta > 1/2$ that the higher criticism test achieves the optimal detection boundary in the signal sparsity parameter β. In particular we will be interested in the case where all G_n come from a common scale family with scale parameter σ_n. The regime of [35] is a special case where $G_n = \delta_{\sigma_n}$ for $\sigma_n = \sqrt{2r \log n}$.

Interestingly, we find that relaxing the assumption of identical non-null signals shows the max test in a considerably better light. Our main results are summarized in the last three rows of Table 5.1. Essentially, if the tails of G_n are at least as heavy as Gaussian, the max test achieves optimal performance throughout the sparse regime, i.e. $\beta > 1/2$. Furthermore, if G_n has polynomial tails, we find that the max test asymptotically dominates the modified higher criticism test; the higher criticism and Berk-Jones tests remain competitive but only because of their similarity to the max test. We give explicit formulae for the detection threshold when G_n has Gaussian, exponential, and polynomial tails and confirm our results with numerical experiments. We find empirically that a hybrid test combining the max test and χ^2 test is a practical choice with high power across all sparsity levels.

We hope that our results will help to rehabilitate the max test, which enjoys many practical advantages over its special-purpose competitors in settings where asymptotic results are equivocal:

	Asymptotic parameters		Achieves optimal asymptotic behavior		
Alternative distribution	σ_n	β	Max test	Higher criticism	Modified HC
Point mass	$r\sqrt{\log n}$	$(1/2, 3/4)$	✗	✓	✓
		$(3/4, 1)$	✓	✓	✓
Gaussian	r	$(1/2, 1)$	✓	✓	✓
Exponential	$\frac{r}{\sqrt{2\log n}}$	$(1/2, 1)$	✓	✓	✓
Student's t_ν	$\frac{r\sqrt{2\log n}}{n^{(1-\beta)/\nu}}$	$(1/2, 1)$	✓	✓	✗

Table 5.1: Optimality of different tests for special cases of our asymptotic regime, where σ_n is calibrated so that the problem is barely solvable. For the point mass, Gaussian and exponential distribution, a checkmark ✓indicates that the test achieves the optimal "detection boundary" for the parameter r. For Student's t_ν, there exists no sharp "detection boundary" for r, and a checkmark ✓indicates that the test has full asymptotic power as $r \to \infty$. These results are proved in Theorems 2–3 and Corollary 1.

First, its Type I error control is fairly robust to incorrect specification of the dependence between p-values; by contrast, the higher criticism test can be highly anticonservative even with very slight correlations between p-values. Second, when the max test rejects, the logical and mathematical basis for rejection is extremely simple and transparent: namely, that one $|X_i|$ value was too large. This simplicity confers a form of scientific robustness, allowing non-expert users to more easily interrogate how modeling assumptions contribute to the scientific conclusion. Third, beyond the multiple testing interpretation giving rise to the max test, we can also easily invert it to obtain a simple rectangular confidence region for $\mu \in \mathbb{R}^n$ giving simultaneous confidence intervals for every μ_i; the totality of these inferences is much more informative than a binary accept/reject decision about the global null. By contrast, for the other tests, there is a more complex relationship between rejecting the global null and making inferences about individual μ_i values. Fourth, the modified higher criticism test cannot reject unless the fifth-largest $|X_i|$ is quite large; as a result, it is essentially powerless in the sparsest setting, where there are one or two extremely large signals. Finally, the max test is computationally cheap while the others require lengthy simulations.

Related work

Some recent theoretical work on global testing has relaxed the assumption of identical non-null means. Cai, Jessie Jeng, and Jin [22] considered the case where the non-null means are sampled from a Gaussian distribution $N(A_n, \sigma^2)$ where the variance σ^2 is fixed and $A_n = \sqrt{2r\log n}$ for some $r \in (0, 1)$. Under this model, they showed that the higher criticism test achieves optimal asymptotic behaviour for $\beta \in (0, 1)$. Although different from a point mass, the model resembles the one in Donoho and Jin [35]: since σ^2 is fixed as $n \to \infty$, the non-null means still concentrate around $\sqrt{2r\log n}$, leading to qualitatively similar limiting behavior as a point mass. Cai and Wu [24] expanded this analysis to the more general model (5.1), proving optimality in certain conditions for the higher criticism test but not discussing the power of the commonly used max

test.

The higher criticism's favorable theoretical performance has led to many efforts to generalize it beyond the model with independent errors studied here. One line of theoretical work has focused on studying the properties of higher criticism type tests when observations are correlated. Hall and Jin [50] gave a detailed discussion of related issues. They showed that the null distribution of higher criticism changes dramatically under weak dependence. In contrast, the max test is more robust to dependence, and the Type-I error can be controlled under arbitrary dependence. Hall, Jin, et al. [49] later proposed the innovated higher criticism to deal with the case of known covariance matrix with polynomially decaying off-diagonal elements. However, the innovated higher criticism can only be used if the covariance matrix of observations can be estimated reasonably well. Statisticians have also proposed various extensions of higher criticism type tests for more general settings, such as ANOVA [3], time-frequency analysis [23], genetic association studies [10], multi-sample analysis [29], and polynomial tailed noise distributions [4], etc. It is an interesting question for future work whether the max test or generalizations thereof might perform equally well.

There has also been a lot of work that studies higher criticism type tests from a computational perspective. In practice, the cutoff and p-values of higher criticism type statistics is often obtained by Monte Carlo simulation. An alternative approach for small sample size via numerical recursion was given by Noé [97], Owen [98] and further developed by Moscovich, Nadler, and Spiegelman [89], Moscovich and Nadler [88] and Li and Siegmund [73]. Li and Siegmund [73] showed that their approximations for the p-value of higher criticism type statistics are reasonably accurate, even for small p-values and large samples.

Most papers on the global testing problem focus on the performance of the higher criticism or related statistics. Our contributions differ from these in that we show the max test enjoys many of the same theoretical advantages despite its simple form, and has similar finite sample power as the higher criticism test in a wide range of settings.

5.2 Main results

The critical sparsity level

We consider the following sequence of alternatives

$$H_1^n : \mu_i \overset{i.i.d}{\sim} (1-\pi_n)\delta_0(\cdot) + \pi_n G_n(\cdot), \tag{5.2}$$

where the expected proportion of nonzero means is

$$\pi_n = n^{-\beta}, 0 < \beta < 1$$

and $G_n(\mu)$ is the distribution of the nonzero means. With slight abuse of notation, we will also use G_n to denote the cumulative distribution function of the distribution. The alternative hypothesis in Donoho and Jin [35] is a special case of this model taking G_n as the point mass at $\sqrt{2r\log n}$, for $0 < r \leq 1$. Following most previous literature on this topic, we restrict our attention to the sparse

regime with $\beta < 1/2$; otherwise the χ^2 test is potentially much more powerful than other tests. For simplicity, we drop the superscript on H_1 when the dimension n is clear.

The *total variation* (TV) distance between two probability measures Q_1 and Q_2 is defined as $d_{TV}(Q_1, Q_2) = \sup_A |Q_1(A) - Q_2(A)|$. For any test that tries to distinguish H_1^n from H_0^n, the sum of its Type I and Type II error is lower bounded by

$$1 - d_{\text{TV}}(H_0^n, H_1^n),$$

where we write $d_{\text{TV}}(H_0^n, H_1^n)$ as a shorthand for

$$d_{\text{TV}}\left(H_0^n, H_1^n\right) = d_{\text{TV}}(\Phi^n, ((1-\pi_n)\Phi + \pi_n(G_n * \Phi))^n)\,.$$

By the Neyman-Pearson lemma [95], the likelihood ratio test is uniformly most powerful for testing H_1^n against H_0^n. Indeed, the above lower bound is achieved if we reject H_0^n when the likelihood ratio is greater than 1. Therefore, the TV distance $d_{\text{TV}}(H_0^n, H_1^n)$ tightly characterizes the hardness of the testing problem.

For any sequence G_n, the TV distance $d_{\text{TV}}(H_0^n, H_1^n)$ is non-increasing in β for each n, with larger values of β making the testing problem harder. Following [24], we introduce the concept of the *critical sparsity level*, which is a value β^* that demarcates a sharp transition from asymptotic consistency to asymptotic powerlessness:

Definition 1. *Fixing the sequence $\{G_n\}$, we define*

$$\underline{\beta}^* = \sup\left\{\beta \geq 0 : \lim_n d_{TV}(H_0^n, H_1^n) = 1\right\};\ \ \textit{and}\ \ \bar{\beta}^* = \inf\left\{\beta \leq 1 : \lim_n d_{TV}(H_0^n, H_1^n) = 0\right\}.$$

When $\underline{\beta}^* = \bar{\beta}^*$, we denote the common value as β^*, and call it the critical sparsity level corresponding to $\{G_n\}$.

If a critical sparsity level β^* exists for a sequence $\{G_n\}$ (i.e., if $\underline{\beta}^* = \bar{\beta}^*$), it follows from Definition 1 that

- If $\beta > \beta^*$, then $\lim_{n\to\infty} \mathbb{E}_{H_1^n}[\phi_n(X)] \to \alpha$ for any sequence of level-α tests ϕ_n, and
- If $\beta < \beta^*$, then $\lim_{n\to\infty} \mathbb{E}_{H_1^n}[\phi_{LRT}(X)] \to 1$ for the level-α likelihood ratio test ϕ_{LRT}.

We say that a sequence of level-α tests ϕ_n is *asymptotically consistent* on the sequence $\{H_1^n\}$ if $\lim_{n\to\infty} \mathbb{E}_{H_1^n}[\phi_n(X)] \to 1$ for any α, and *asymptotically powerless* on the sequence $\{H_1^n\}$ if $\lim_{n\to\infty} \mathbb{E}_{H_1^n}[\phi_n(X)] \to \alpha$ for any α. We say that the sequence *achieves the optimal critical sparsity level* for the sequence $\{G_n\}$ if it has full asymptotic power whenever $\beta < \beta^*$.

It will often be natural to parameterize the tail of Gaussian distribution as $\sqrt{2\delta \log n} \approx z_{n^{-\delta}}$, the upper $n^{-\delta}$ quantile of the standard normal distribution. If we define

$$\tau_n(\delta) = \log_n \mathbb{P}_{\mu\sim G_n}(X > \sqrt{2\delta \log n}) \tag{5.3}$$

as the tail probability of a single non-null observation, [24] proved sufficient conditions for optimality of the higher criticism test in the sparse regime:

Proposition 2. *Suppose that* $\{\tau_n(\delta)\}_{n=1}^{\infty}$ *converges uniformly for all* $\delta \in [0,1]$. *Then the sequence of alternatives in* (5.2) *has a critical sparsity level* β^*, *and if* $\beta^* > 1/2$ *then the level-*α *higher criticism test has full asymptotic power whenever* $\beta < \beta^*$.

While [24] only explicitly proved this for the higher criticism test, one can slightly modify their proof to show that this proposition holds for the modified higher criticism test and the Berk-Jones test as well (the proof is deferred to the supplementary material). In this paper, we are interested in the following question: for which distributions G_n does the max test achieve the same critical sparsity level β^*? [35] showed that this is true when G_n is a point mass and $\beta^* \geq 3/4$, which is by far the best-known result for this problem. We will show that under a mild regularity condition, when $\beta^* > 1/2$, the max test also achieves the optimal critical sparsity level.

Optimality of the max test

To formally state our main result, we first need to introduce *regularly varying functions*. Following [17], we say that a function $Q : (0,\infty) \to (0,\infty)$ is a regularly varying function if the limit

$$g_Q(t) = \lim_{x\to\infty} \frac{Q(tx)}{Q(x)}$$

is finite and nonzero for all $t > 0$. For any regularly varying function Q, it was shown in [45] that the limit $g_Q(t)$ has the form

$$g_Q(t) = t^{\gamma}$$

for some value $\gamma \in (-\infty,\infty)$, which is called the *index of regular variation* of Q.

Among distributions with unbounded support, we consider those for which

$$-\max\{\log(1-G(\theta)), \log G(-\theta)\} \text{ is a regularly varying function.}$$

As noted by [4], this class of distributions extended the definition of generalized Gaussian models, which are commonly used as benchmarks in this line of work. It covers the cases where $\log(1-G(\theta)) = \log G(-\theta) \sim -\theta^a(\log\theta)^b, a > 0, b \in \mathbb{R}$. The index γ corresponds to the tail of the distribution Q, with smaller γ indicating heavier tails. In particular, $\gamma = 2$ corresponds to a Gaussian tail, and $\gamma = 1$ to an exponential tail.

Our main result shows essentially that the max test achieves the optimal detection boundary as long as $\beta \geq 3/4$ *or* the tail of G is no lighter than Gaussian:

Theorem 2. *Under the assumptions of Proposition 2, suppose that either*

(A1) $\beta^* > 3/4$, *or*

(A2) G_n *is a scale family with* $G_n(\mu) = G(\mu/\sigma_n)$ *for some sequence* σ_n, *where* $-\max\{\log(1-G(\theta)), \log G(-\theta)\}$ *is a regularly varying function with index of regular variation* $\gamma \leq 2$.

Then if $\beta^* > 1/2$, *the level-*α *max test* ϕ_{Max} *has full asymptotic power whenever* $\beta < \beta^*$.

We will provide intuition for Theorem 2 and a partial proof in Section 5.2, deferring a key technical lemma to the supplementary material. The regularly varying assumption cannot be removed for $\beta^* < 3/4$; see the supplementary material for a counterexample where G is stochastically larger than an exponential distribution but the max test is not optimal. Finally, note that neither Proposition 2 nor Theorem 2 characterizes what occurs at the boundary where $\beta = \beta^*$; we discuss this boundary regime in the polynomial tail case in Section 5.2.

Viewing the results of Donoho and Jin [35] in light of Theorem 2, we see that the suboptimality of the max test in their asymptotic regime is a result of the assumption that all nonzero μ_i are identical. As a direct corollary of Theorem 2, we can derive explicit formulae for the critical sparsity levels of densities with polynomial tails, exponential tails, and Gaussian tails respectively:

Corollary 1. *Suppose that G_n belong to a scale family with $G_n(\mu) = G(\mu/\sigma_n)$, for some distribution G with density function $g(\theta)$.*

(1) If $g(\theta) = \Theta(\theta^{-\nu-1})$ for some $\nu > 0$, and $\sigma_n \sim n^\rho$ with $\rho > -(2\nu)^{-1}$, then the critical sparsity level is

$$\beta^*(\rho) = \nu\rho + 1.$$

(2) If $g(\theta) = \Theta(e^{-\theta})$ and $\sigma_n = r(2\log n)^{-1/2}$ with $r > \sqrt{2}/(\sqrt{2}-1)$, then

$$\beta^*(r) = \left(1 - \frac{1}{r}\right)^2.$$

(3) If $g(\theta) = \Theta\left(e^{-\frac{(\theta-\gamma)^2}{2}}\right)$ for some $\gamma \in \mathbb{R}$, and $\sigma_n = r$ with $r > 1$, then

$$\beta^*(r) = \frac{r^2}{r^2+1}.$$

[24] derived the critical sparsity level when the alternative means follow the generalized Gaussian distribution, and Part (2) and (3) of this corollary are special cases of such distribution. In these two scenarios, Theorem 2 shows that the likelihood ratio test, the max test and the higher criticism test are asymptotically consistent when $\beta < \beta^*(r)$, and asymptotically powerless when $\beta > \beta^*(r)$. As such, the critical sparsity level produces a sharp detection boundary for the scale parameter r. For example, if $g(\theta) = \Theta(e^{-\theta})$ and $\sigma_n = r(2\log n)^{-1/2}$, then all three tests are asymptotically consistent if $r > r^*(\beta) := (1-\sqrt{\beta})^{-1}$, and asymptotically powerless if $r < (1-\sqrt{\beta})^{-1}$. Thus, the desired sharp detection boundary is $r^*(\beta) = (1-\sqrt{\beta})^{-1}$. Part (1) of this corollary exhibits a different regime: when the alternative means follow a t distribution, there does not exist a sharp detection boundary for a scale parameter r. Instead, there is a sharp detection boundary in the growth rate ρ if we set $\sigma_n = n^\rho$. We explore the boundary regime of the polynomial tail case further in Section 5.2.

Proving Theorem 2 using excess tail values

In this section, we will explain the mathematical intuition behind Theorem 2, and provide a sketch of its proof. We begin by introducing a useful transformation of the empirical distribution of X_i

values, in terms of the tail parameter δ. Defining $N(\delta) = \#\left\{i : \ |X_i| > \sqrt{2\delta \log n}\right\}$, the higher criticism statistic may be rewritten as

$$\text{HC}_n = \sup_{\delta \geq 0} \frac{N(\delta) - \mathbb{E}_0 N(\delta)}{\sqrt{\text{Var}_0 N(\delta)}} \approx \sup_{\delta > 0} \frac{N(\delta) - n^{1-\delta}}{n^{(1-\delta)/2}},$$

where the approximation holds for large n, if the supremum is not achieved too close to $\delta = 0$. Roughly speaking, then, the higher criticism test will have high power when the number of excess tail values is much larger than $n^{(1-\delta)/2}$, for some $\delta > 0$. By contrast, the max test rejects roughly when $N(1) > 0$.

Under the alternative, the most likely source of these excess tail values is the $n\pi_n$ non-null observations. We quantify their contribution as $N_1(\delta) = \#\left\{i : \ \mu_i \neq 0, |X_i| > \sqrt{2\delta \log n}\right\}$, and define

$$\lambda_n(\delta) = \log_n \mathbb{E}_1 N_1(\delta) = 1 - \beta + \tau_n(\delta), \tag{5.4}$$

where τ_n is defined in Equation 5.3. Continuing our intuition from above, we expect that the higher criticism test will have high power when $\lambda_n(\delta) > \frac{1-\delta}{2}$ for any $\delta \in (0, 1]$, while the max test will have high power when $\lambda_n(1) > 0$ in the limit.

Suppose that $\{\tau_n(\delta)\}_{n=1}^{\infty}$ converges uniformly for all $\delta \in [0, 1]$. This is the same condition as Proposition 2 and Theorem 2. Under this condition, we denote

$$\tau^*(\delta) = \lim_{n \to \infty} \tau_n(\delta), \quad \text{and} \quad \lambda^*(\delta) = \lim_{n \to \infty} \lambda_n(\delta).$$

We can formalize the above heuristic characterization in Proposition 3:

Proposition 3. *Suppose that $\beta > 1/2$. Then*

(a) If

$$\sup_{\delta \in (0,1]} \left[\lambda^*(\delta) - \frac{1-\delta}{2}\right] < 0,$$

then $d_{TV}(H_0^n, H_1^n) \to 0$.

(b) If

$$\sup_{\delta \in (0,1]} \left[\lambda^*(\delta) - \frac{1-\delta}{2}\right] > 0,$$

then the likelihood ratio test, the higher criticism test, modified higher criticism test, and Berk-Jones tests all enjoy full asymptotic power.

(c) The max test is asymptotically powerless if $\lambda^*(1) < 0$, and enjoys full asymptotic power if $\lambda^*(1) > 0$.

The proof of part (a) of Proposition 3 is given in [24]. Cai and Wu [24] also proved that the higher criticism test enjoys full asymptotic power under the condition of Part (b). For the modified

higher criticism test and Berk-Jones tests, the proof is similar and is given in the supplementary material for completeness. Part (c) follows directly from the first and second Borel-Cantelli Lemma.

Proposition 3 leaves open the question of what happens in the boundary regime where the supremum converges to 0. Indeed, Section 5.2 studies a natural regime with polynomial tails where $\lambda_n(1) \to 0$ and the modified higher criticism test is powerless in the limit even while the other tests enjoy full asymptotic power.

Note further that the sufficient condition in Theorem 3 for the max test to have full asymptotic power is more restrictive than the sufficient condition for the other three. This analysis suggests a disadvantage for the max test, which we illustrate in Figure 5.1 showing four different λ curves plotted against $\frac{1-\delta}{2}$. The black curve takes G_n as a point mass, and shows a bad case for the max test: it rises above $\frac{1-\delta}{2}$ for a range of δ values that exclude 1. The other three curves, however (taking G_n as Gaussian, exponential, and Cauchy), all show cases where the supremum is achieved at $\delta = 1$, so that all of the tests enjoy high power.

Roughly speaking, max test achieves the optimal critical sparsity level if the supremum of $\lambda^*(\delta) - \dfrac{1-\delta}{2}$ is achieved at $\delta = 1$. The following technical lemma connects this supremum with the tail property of G_n, and is essential in the proof of Theorem 2.

Lemma 2. *(a) For any $\beta > 1/2$ and sequence $\{G_n\}$,*

$$\sup_{\delta\in(0,1]} \left[\lambda^*(\delta) - \frac{1-\delta}{2}\right] \;\le\; \max\left\{\lambda^*(1),\; \frac{3}{4} - \beta\right\}$$

(b) Under Assumption (A2) of Theorem 2,

$$\sup_{\delta\in(0,1]} \left[\lambda^*(\delta) - \frac{1-\delta}{2}\right] \;\le\; \max\left\{\lambda^*(1),\; \frac{1}{2} - \beta\right\}.$$

The proof of the lemma is given in Section 5.5. Theorem 2 is then a direct result of Lemma 2 and Proposition 3.

Proof of Theorem 2. First, if $\beta < \beta^*$, then by definition of β^*,

$$\sup_{\delta\in(0,1]} \left[\lambda^*(\delta) - \frac{1-\delta}{2}\right] \ge \beta^* - \beta > 0. \tag{5.5}$$

Suppose that $\beta^* > 3/4$ and $\beta < \beta^*$. Since the power of the max test is non-increasing in β, we can assume without loss of generality that $\beta > 3/4$. Since $3/4 - \beta < 0$, we can combine (5.5) with part 1 of Lemma 2 to conclude that $\lambda^*(1) > 0$, implying that the max test has full asymptotic power. If Assumption (A2) of Theorem 2 holds, then we can repeat the same argument replacing 3/4 with 1/2 and applying part 2 of Lemma 2 instead of part 1.

□

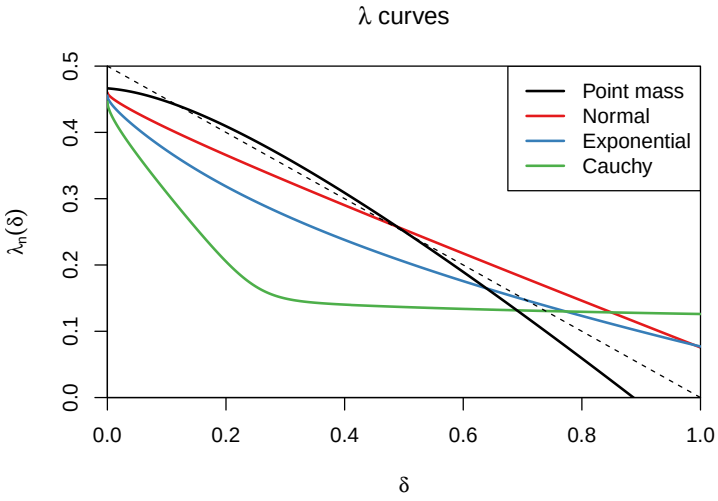


Figure 5.1: $\lambda_n(\delta)$ curves plotted against $\frac{1-\delta}{2}$ for four different tests, for $\lambda_n(\delta)$ as defined in (5.4). If a curve rises above $\frac{1-\delta}{2}$ for some values of δ that exclude 1, then the likelihood ratio test has full asymptotic power, while the max test does not. The black curve, which takes G_n as a point mass, shows this scenario. The other curves show case where the supremum $\lambda_n(\delta) - \frac{1-\delta}{2}$ of is achieved at $\delta = 1$, so that the max test, the higher criticism test and the likelihood ratio test all enjoy full asymptotic power.

Power analysis for polynomial tails

Theorem 2 does not characterize the power of different tests in the boundary regime. We now study a natural regime with polynomial tails, with $\beta = \beta^*$. The boundary regime with polynomial tails is more interesting because we have shown in Corollary 1 that there is not a sharp detection boundary for a scale parameter, but rather in the growth rate ρ where $\sigma_n = n^\rho$.

In this section we explore a sequence of alternative distributions growing at the critical rate ρ, and parameterized by a scale parameter r. Under this sequence of alternatives, we will show that the asymptotic power of level-α max test is a smooth function of $r \in (0, \infty)$, and converges to 1 as $r \to \infty$. In addition, we will show that the modified higher criticism test is asymptotically powerless no matter what r is.

Suppose that $G_n(\mu) = G(\mu/\sigma_n)$, where G is the t distribution with ν degrees of freedom. Then the density function $g(\theta) = \Theta(\theta^{-\nu-1})$. Recall from Corollary 1 that if $\liminf_n \log_n \sigma_n > (\beta-1)/\nu$, then the max test and higher criticism test both have full asymptotic power. If $\limsup_n \log_n \sigma_n < (\beta-1)/\nu$, then both tests are powerless. Therefore, to study the boundary regime, we are interested in the case where $\lim_n \log_n \sigma_n = (\beta-1)/\nu$. Fix $\beta \in (1/2, 1)$, and let

$$\sigma_n = \frac{r\sqrt{2\log n}}{n^{(1-\beta)/\nu}}, r \in (0, \infty).$$

Then it can be verified that the power of the max test has smooth transition from α to 1 as r goes from 0 to ∞. The higher criticism test also shares this smooth transition behavior, as the rejection

threshold for $p_{(1)}$ in the higher criticism statistic is very close to α/n. Perhaps surprisingly, the modified higher criticism test is asymptotically powerless in this case, as detailed by the following theorem.

Theorem 3. *Suppose that G satisfies $\lim_{\mu\to\infty}(1-G(\mu))\mu^\nu = \lim_{\mu\to\infty} G(-\mu)\mu^\nu = C$ with tail index $\nu > 0$, and $\sigma_n = \dfrac{r\sqrt{2\log n}}{n^{(1-\beta)/\nu}}$ for some $\beta \in (1/2, 1)$. Then $\beta^* = \beta$, and*

1. *the asymptotic power of the level-α max test, is*

$$\lim_{n\to\infty} \mathbb{P}_{H_1}(\textit{reject } H_0) = 1 - e^{-2Cr^\nu + \log(1-\alpha)}.$$

   In particular, the power tends to 1 as $r \to \infty$.

2. *for any $r \in (0, \infty)$, the modified higher criticism is asymptotically powerless.*

We note that for fixed r, the power of the max test as n goes to infinity does not depend on the sparsity parameter β. This is because σ_n is a decreasing function of the sparsity level β, thereby implicitly adjusting for the sparsity level.

Compared to the original higher criticism test, the modified higher criticism test was designed to ignore $p-$values smaller than $1/n$. These small p-values cause the original higher criticism statistics to have a heavy right tail under the null distribution, and the modified higher criticism test is considered in [35] as a refined test with potentially better finite sample performance. However, this modification also makes the modified higher criticism test powerless in situations where the smallest p-values provide the best evidence against the null. Recall that $\lambda_n(\delta)$ is defined as the log of the expected number of non-null observations that are greater than $\sqrt{2\delta\log n}$. In Theorem 3's setting, the proof of Corollary 1 shows that $\lambda^*(\delta) = 0$ for all $\delta \in (0, 1]$; as a result $\lambda^*(\delta) < (1-\delta)/2$ for all $\delta < 1$. In other words, evidence against the null is only present in the number of tail values exceeding $\sqrt{2\log n}$, which is roughly the Bonferroni threshold. Because the p values of these observations are smaller than $1/n$, they are effectively truncated by the modified higher criticism test, making it asymptotically powerless. The original higher criticism test, however, is still powerful because, like the max test, it can reject on the strength of the largest p-value alone. A full proof of Theorem 3 is given in Section 5.5.

5.3 Numerical results

We now provide simulation results showing that the max test has similar power as the higher criticism test when the distribution of non-null signals has Gaussian or heavier tails. We generate data under the following alternative:

$$\begin{aligned} X_i \overset{ind}{\sim} N(\mu_i, 1), \mu_i \overset{i.i.d.}{\sim} G_n, &\quad \text{for } i = 1, \ldots, n_1 \\ X_i \overset{ind}{\sim} N(0, 1), &\quad \text{for } i = n_1 + 1, \ldots, n. \end{aligned}$$

In this section, we consider the case where G_n has either exponential or Cauchy tail. In the supplementary material, we provide additional simulation results for other distributions G_n, including the Gaussian distribution. We take $n = 50,000$ and $n_1 = \lfloor n^{1-\beta} \rfloor$, where the sparsity parameter β ranges from 0.1 to 0.9. We compare the power of the following 6 tests: the max test, the higher criticism test, the modified higher criticism test, the Berk-Jones test, the χ^2 test and a hybrid test which combines the max test and the χ^2 test. The rejection region of the level α hybrid test has the form

$$\left\{\max_i |X_i| > m(n, \alpha/2)\right\} \cup \left\{\sum_i X_i^2 > c(n, \alpha/2)\right\},$$

where $m(n, \alpha/2)$ and $c(n, \alpha/2)$ are the $1 - \alpha/2$ quantiles of $\max_i |X_i|$ and $\sum_i X_i^2$ under the null. For all 6 tests, we control Type-I error at $\alpha = .05$. For the first five tests, we use the empirical 95% percentile of the test statistics under the null distribution as the cutoff value; for the hybrid test, we use the empirical 97.5% percentile of $\max_i |X_i|$ and $\sum_i X_i^2$ to estimate the threshold $m(n, \alpha/2)$ and $c(n, \alpha/2)$. Our results are summarized below.

When G_n has exponential tail In particular, we choose $G_n = \text{Laplace}(0, r)$. The power of all six tests are shown in Figure 5.2. First, we found that when $\beta \leq 0.3$, the χ^2 test (yellow curve) outperforms all five others, and the max test is least powerful due to relatively dense signals. Second, the modified higher criticism test has very low power when $\beta > 0.5$. Since the modified higher criticism test does not use the p-values smaller than $1/n$, it performs subpar in the sparse regime where the max test and the higher criticism test reject the null based on those p-values. Third, when $\beta > 0.5$ the power of the max test, the higher criticism test and the Berk-Jones test are very similar. This finding agrees with our Theorem 1, which states that the max test achieves the optimal critical sparsity level for exponentially distributed alternatives when $\beta > 0.5$. Finally, the hybrid test, which combines the max test and the χ^2 test, performs on par if not better than the higher criticism under all sparsity regimes.

When G_n has polynomial tail In particular, we choose $G_n = \text{Cauchy}\left(0, \frac{r\sqrt{2\log n}}{n^{(1-\beta)}}\right)$. Recall that according to Theorem 3, under this setting the max test and the higher criticism should have very high power when r is big, while modified higher criticism should have little power. Indeed, the max test, the higher criticism test, the Berk-Jones test and the hybrid test have almost identical power for all combinations of (β, r), and the modified higher criticism performs worst among all tests. All of these findings are consistent with our Theorem 2. We also notice that for fixed r value, the power of max test, higher criticism and Berk-Jones are almost constant for different parameter β. This finding also agrees with the asymptotic power of max test in Theorem 3.

Section 5.6 in the supplementary material gives analogous simulations for Gaussian, logistic, $\chi^2(1)$, t_3, and t_5 distributions, with qualitatively similar results. Overall, our simulation confirms that the higher criticism test does not have better finite sample power than the max test when the max test achieves the optimal critical sparsity level. On the other hand, when the higher criticism does have better power over the max test, the non-null signals are likely dense enough such that the χ^2 test is even more powerful.

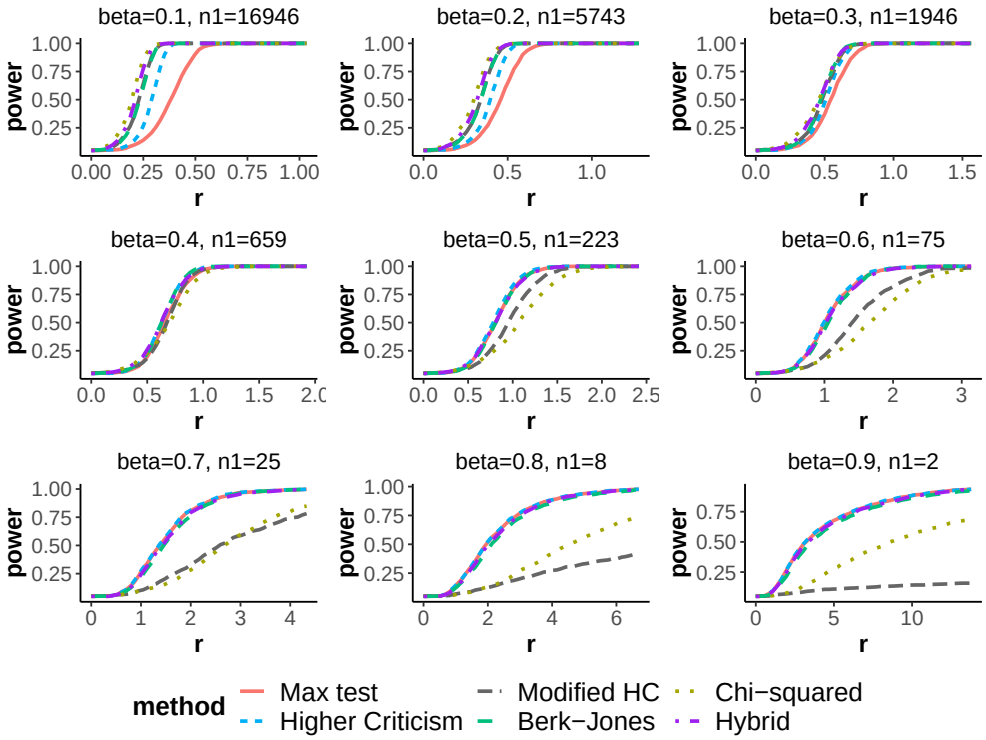


Figure 5.2: Comparison of power for different tests: the max test (red curve), the higher criticism test (light blue curve), the modified higher criticism test (grey curve), the Berk-Jones test (green curve) ,the χ^2 test (yellow curve) and the hybrid test (purple curve). Here $n = 50,000$ with $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from Laplace$(0, r)$. The horizontal axis shows the value of r while the vertical axis shows power.

5.4 Discussion

We have shown, theoretically and numerically, that the max test has optimal asymptotic behavior in the sparse regime, provided that the distribution of non-null signals has a tail no lighter than Gaussian. In addition, the max test dominates the modified higher criticism test when the distribution of nonzero signals has polynomial tails. We believe our results complicate the conventional wisdom that the max test is a substandard test for the purpose of signal detection and suggest that in many applied settings practitioners will not suffer low performance by using the max test. In these settings, the max test can be derived as a "free" (incurring no additional FWER) deduction from simultaneous confidence intervals for the coordinates of μ_i.

The higher criticism has been generalized to many interesting statistics problems beyond the signal detection problem studied here. It is an interesting question for future work whether in many of these cases it may be possible to find an analogous generalization of the max test whose performance matches the higher-criticism-type test.

Like other papers in this line of research, our paper did not address the "weak, dense" regime, where the sparsity parameter β is smaller than $1/2$. It is well known that in the dense regime, the

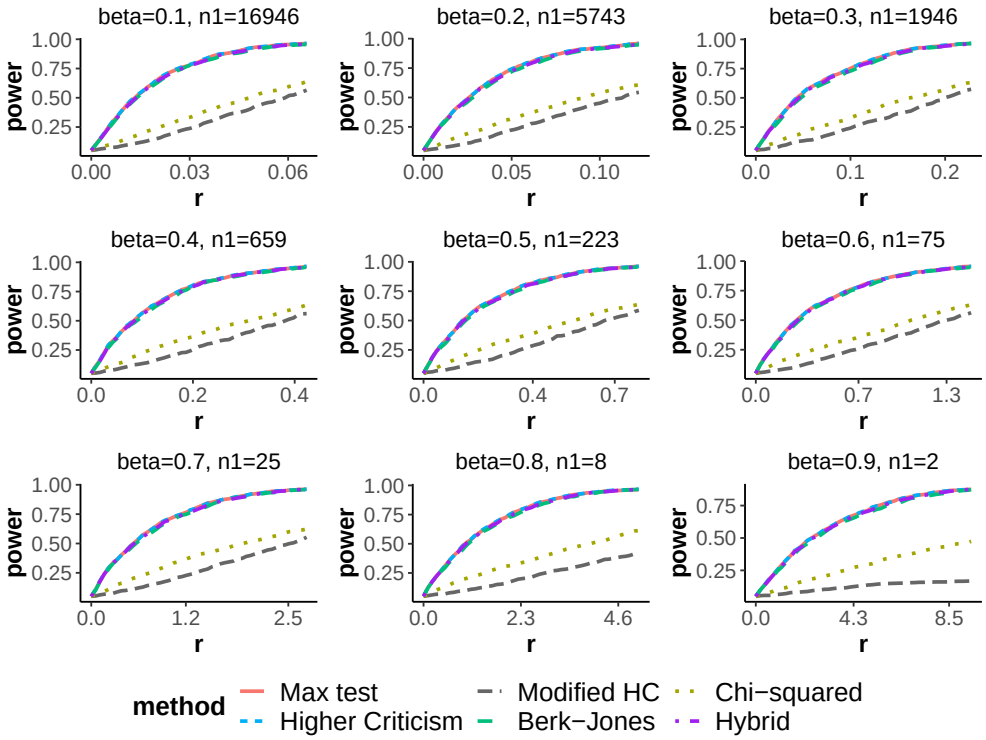


Figure 5.3: Comparison of power for different tests, where $n = 50,000$ with $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from Cauchy$(0, r\sqrt{2 \log n} n^{-(1-\beta)})$. The horizontal axis shows the value of r while the vertical axis shows power.

χ^2 test has higher power than the higher criticism and max test when the distribution of non-null means is a point mass. We have suggested a hybrid test based on combining the p-values of the χ^2 and the max test, and shown numerical evidence that it performs well throughout the sparse and dense regimes. By inverting this hybrid test, we can obtain a joint confidence region for $\mu \in \mathbb{R}^n$ that is the union of an ℓ_2 ball and an ℓ_∞ ball around the observed X, simultaneously giving short intervals for coordinates of μ_i and reasonable intervals for all linear combinations of μ. Finding a test that achieves the optimal critical sparsity level under the general model in this regime is a interesting direction for future research.

5.5 Proofs of main results

We begin by proving the following result on the tail probability of $X \sim N(\mu, 1)$, where μ is generated from some distribution G_n. This is a standard result and is repeated here for completeness of the proof.

Lemma 3. *Let $\bar{G}_n(\theta) = 1 - G_n(\theta)$. Under the alternative hypothesis* (5.2)*, we have*

$$\tau_n(\delta) = \sup_{0\le t\le 1} -\frac{Q_n(t\sqrt{2\delta\log n})}{\log n} - \delta(1-t)^2 + O\left(\frac{\log\log n}{\log n}\right),$$

where $Q_n(\theta) = -\max\{\log \bar{G}_n(\theta), \log G_n(-\theta)\}$ and the $O\left(\frac{\log\log n}{\log n}\right)$ term is uniform over all $\delta \in [0,1]$.

Proof. For any $0 \le t, \delta \le 1$, we have

$$\begin{aligned}
&\mathbb{P}_{\mu\sim G_n}\left(X \ge \sqrt{2\delta\log n}\right)\\
&\ge \mathbb{P}_{\mu\sim G_n}\left(X \ge \sqrt{2\delta\log n}, \mu \ge t\sqrt{2\delta\log n}\right)\\
&= \mathbb{P}_{\mu\sim G_n}\left(\mu \ge t\sqrt{2\delta\log n}\right)\mathbb{P}_{\mu\sim G_n}\left(X \ge \sqrt{2\delta\log n} \;\middle|\; \mu \ge t\sqrt{2\delta\log n}\right)\\
&\ge \left(1 - G_n(t\sqrt{2\delta\log n})\right)\left(1 - \Phi((1-t)\sqrt{2\delta\log n})\right)\\
&\ge \frac{1}{6\sqrt{2\log n}}\exp\left\{\log\bar{G}_n\left(t\sqrt{2\delta\log n}\right) - (1-t)^2\delta\log n\right\},
\end{aligned}$$

where the last inequality follows from the fact that $1 - \Phi(x) \ge \frac{1}{3(x+1)}e^{-x^2/2}$ for any $x > 0$. Taking the supremum over $t \in [0,1]$, we have

$$\mathbb{P}_{\mu\sim G_n}\left(X \ge \sqrt{2\delta\log n}\right) \ge \frac{1}{6\sqrt{2\log n}}\exp\left\{\sup_{0\le t\le 1}\log\bar{G}_n\left(t\sqrt{2\delta\log n}\right) - (1-t)^2\delta\log n\right\}.$$

On the other hand, Fubini's theorem yields

$$\begin{aligned}
&\mathbb{P}_{\mu\sim G_n}\left(X \ge \sqrt{2\delta\log n}\right)\\
=& -\int_{-\infty}^{\infty}\bar{\Phi}(\sqrt{2\delta\log n} - \mu)d\bar{G}_n(\mu)\\
=& -\int_{-\infty}^{0}\bar{\Phi}(\sqrt{2\delta\log n} - \mu)d\bar{G}_n(\mu) - \int_{0}^{\sqrt{2\delta\log n}}\bar{\Phi}(\sqrt{2\delta\log n} - \mu)d\bar{G}_n(\mu)\\
&-\int_{\sqrt{2\delta\log n}}^{\infty}\bar{\Phi}(\sqrt{2\delta\log n} - \mu)d\bar{G}_n(\mu) \qquad (5.6)
\end{aligned}$$

For the first and third terms of Equation 5.6, we have

$$-\int_{-\infty}^{0}\bar{\Phi}(\sqrt{2\delta\log n} - \mu)d\bar{G}_n(\mu) \le n^{-\delta}\bar{G}_n(0) \qquad (5.7)$$

and

$$-\int_{\sqrt{2\delta\log n}}^{\infty}\bar{\Phi}(\sqrt{2\delta\log n} - \mu)d\bar{G}_n(\mu) \le \bar{G}_n(\sqrt{2\delta\log n}). \qquad (5.8)$$

For the second term, we have

$$
\begin{aligned}
& -\int_0^{\sqrt{2\delta\log n}} \bar{\Phi}(\sqrt{2\delta\log n}-\mu)d\bar{G}_n(\mu) \\
\leq\ & -\int_0^{\sqrt{2\delta\log n}} e^{-\frac{1}{2}(\sqrt{2\delta\log n}-\mu)^2}d\bar{G}_n && (5.9)\\
=\ & -\bar{G}_n(\mu)e^{-\frac{1}{2}(\sqrt{2\delta\log n}-\mu)^2}\Big|_{\mu=0}^{\mu=\sqrt{2\delta\log n}} + \int_0^{\sqrt{2\delta\log n}} (\sqrt{2\delta\log n}-y)\bar{G}_n(y)e^{-\frac{1}{2}(\sqrt{2\delta\log n}-y)^2}dy && (5.10)\\
=\ & n^{-\delta}\bar{G}_n(0)-\bar{G}_n(\sqrt{2\delta\log n})+\int_0^1 (2\delta\log n)(1-t)\bar{G}_n(t\sqrt{2\delta\log n})e^{-\frac{1}{2}(\sqrt{2\delta\log n}(1-t))^2}dt && (5.11)\\
\leq\ & n^{-\delta}\bar{G}_n(0)-\bar{G}_n(\sqrt{2\delta\log n})+(2\log n)\exp\left\{\sup_{0\leq t\leq 1}\log\bar{G}_n(t\sqrt{2\delta\log n})-(1-t)^2\delta\log n\right\}. && (5.12)
\end{aligned}
$$

where Equation 5.9 is obtained by Gaussian tail bounds, Equation 5.10 by integration by parts, Equation 5.11 by changing of variables, and Equation 5.12 by taking the supremum of the integrand over $t\in[0,1]$. Combining Equations 5.6, 5.7, 5.8, and 5.12, we have

$$
\mathbb{P}_{\mu\sim G_n}\left(X\geq\sqrt{2\delta\log n}\right)\leq(2\log n+2)\exp\left\{\sup_{0\leq t\leq 1}\log\bar{G}_n(t\sqrt{2\delta\log n})-(1-t)^2\delta\log n\right\}.
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{6\sqrt{2\log n}}\exp\left\{\sup_{0\leq t\leq 1}\log\bar{G}_n\left(t\sqrt{2\delta\log n}\right)-(1-t)^2\delta\log n\right\}\\
&\leq\ \mathbb{P}_{\mu\sim G_n}\left(X\geq\sqrt{2\delta\log n}\right) && (5.13)\\
&\leq(2\log n+2)\exp\left\{\sup_{0\leq t\leq 1}\log\bar{G}_n\left(t\sqrt{2\delta\log n}\right)-(1-t)^2\delta\log n\right\}.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\frac{1}{6\sqrt{2\log n}}\exp\left\{\sup_{0\leq t\leq 1}\log G_n\left(-t\sqrt{2\delta\log n}\right)-(1-t)^2\delta\log n\right\}\\
&\leq\ \mathbb{P}_{\mu\sim G_n}\left(X\leq-\sqrt{2\delta\log n}\right) && (5.14)\\
&\leq(2\log n+2)\exp\left\{\sup_{0\leq t\leq 1}\log G_n\left(-t\sqrt{2\delta\log n}\right)-(1-t)^2\delta\log n\right\}.
\end{aligned}
$$

Combining the two equations above, we have

$$\begin{aligned}
&\frac{1}{3\sqrt{2\log n}}\exp\left\{\sup_{0\le t\le 1} -Q_n\left(t\sqrt{2\delta\log n}\right)-(1-t)^2\delta\log n\right\}\\
&\le\ \mathbb{P}_{\mu\sim G_n}\left(|X|\ge\sqrt{2\delta\log n}\right)\\
&\le(4\log n+4)\exp\left\{\sup_{0\le t\le 1} -Q_n\left(t\sqrt{2\delta\log n}\right)-(1-t)^2\delta\log n\right\}.
\end{aligned}\tag{5.15}$$

Taking $\log_n$ on both sides, we have

$$-\frac{\log(3\sqrt{2\log n})}{\log n}\le\tau_n(\delta)-\left[\sup_{0\le t\le 1}-\frac{Q_n(t\sqrt{2\delta\log n})}{\log n}-\delta(1-t)^2\right]\le\frac{\log(4\log n+4)}{\log n}.$$

We conclude that

$$\tau_n(\delta)=-\sup_{0\le t\le 1}\frac{Q_n(t\sqrt{2\delta\log n})}{\log n}-\delta(1-t)^2+O\left(\frac{\log\log n}{\log n}\right),$$

where the $O\left(\frac{\log\log n}{\log n}\right)$ term is uniform over all $\delta\in[0,1]$ □

We are now ready to restate and prove Lemma 2:

Lemma 2. *(a) For any $\beta>1/2$ and sequence $\{G_n\}$,*

$$\sup_{\delta\in(0,1]}\left[\lambda^*(\delta)-\frac{1-\delta}{2}\right]\ \le\ \max\left\{\lambda^*(1),\ \frac{3}{4}-\beta\right\}$$

(b) Under Assumption (A2) of Theorem 2,

$$\sup_{\delta\in(0,1]}\left[\lambda^*(\delta)-\frac{1-\delta}{2}\right]\ \le\ \max\left\{\lambda^*(1),\ \frac{1}{2}-\beta\right\}.$$

Proof. Define

$$g_n(\delta,t)=-\frac{Q_n(t\sqrt{2\delta\log n})}{\log n}+\delta\left[\frac{1}{2}-(1-t)^2\right],\quad\text{and}\quad h_n(\delta)=\sup_{0\le t\le 1}g_n(\delta,t).$$

Applying Lemma 3, we have

$$\begin{aligned}
h_n(\delta)&=\tau_n(\delta)+\frac{\delta}{2}+O\left(\frac{\log\log n}{\log n}\right)\\
&=\lambda_n(\delta)-\frac{1-\delta}{2}+\beta-\frac{1}{2}+O\left(\frac{\log\log n}{\log n}\right).
\end{aligned}$$

To prove part (a), it suffices to show that

$$h_n(\delta) \le \max\{h_n(1), 1/4\}, \quad \text{for all } \delta \in (0,1).$$

We prove this claim by supposing that $h_n(\delta) > \max\{h_n(1), 1/4\}$ for some $\delta < 1$, and deriving a contradiction.

Let δ_n^* and t_n^* be values that jointly maximize $g_n(\delta, t)$ over $0 \le \delta, t \le 1$. By assumption,

$$\frac{1}{4} < g_n(\delta_n^*, t_n^*) \le \delta_n^* \left[\frac{1}{2} - (1-t_n^*)^2\right], \tag{5.16}$$

so we must have $\delta_n^* > 1/2$ and $t_n^* > 1/2$, and also

$$\frac{1}{4\delta_n^*} < \frac{1}{2} - (1-t_n^*)^2.$$

Further, because $g_n(\delta_n^*, t_n^*) > h_n(1)$, we also have

$$\begin{aligned} 0 &< g_n(\delta_n^*, t_n^*) - g_n(1, t_n^*\sqrt{\delta_n^*}) \\ &< \delta_n^* \left(\frac{1}{2} - (1-t_n^*)^2\right) - \left(\frac{1}{2} - \left(1 - t_n^*\sqrt{\delta_n^*}\right)^2\right), \end{aligned}$$

which leads to

$$\begin{aligned} & \delta_n^* \left(\frac{1}{2} - (1-t_n^*)^2\right) - \left(\frac{1}{2} - \left(1 - t_n^*\sqrt{\delta_n^*}\right)^2\right) > 0 \\ \iff & \frac{1}{2}\delta_n^* - \delta_n^* + 2t_n^*\delta_n^* - \delta_n^*(t_n^*)^2 - \frac{1}{2} + 1 + \delta_n^*(t_n^*)^2 - 2t_n^*\sqrt{\delta_n^*} > 0 \\ \iff & \frac{1}{2} - \frac{1}{2}\delta_n^* + 2t_n^*\delta_n^* - 2t_n^*\sqrt{\delta_n^*} > 0 \\ \iff & 2t_n^*\sqrt{\delta_n^*}(\sqrt{\delta_n^*} - 1) > \frac{1}{2}(\sqrt{\delta_n^*} - 1)(\sqrt{\delta_n^*} + 1) \\ \iff & 2t_n^*\sqrt{\delta_n^*} < \frac{1}{2}(\sqrt{\delta_n^*} + 1) \\ \iff & \frac{1}{4\delta_n^*} > \left(2t_n^* - \frac{1}{2}\right)^2. \end{aligned}$$

Combining the two equations above, we have

$$\frac{1}{2} - (1-t_n^*)^2 > \left(2t_n^* - \frac{1}{2}\right)^2,$$

a contradiction for $t_n^* > 1/2$.

Turning to part (b), suppose that $Q_n(\theta) = Q(\theta/\sigma_n)$ for some sequence σ_n, where $Q(\theta)$ is a regularly varying function with $g_Q(a) \le a^2$. We consider the following two scenarios.

(i) $\limsup \sqrt{2\log n}\sigma_n^{-1} < \infty$.

Since the distribution G has unbounded support, and

$\limsup \sqrt{2\log n}\sigma_n^{-1} < \infty$,

$$Q(\sqrt{2\log n}\sigma_n^{-1})$$

is bounded. Therefore

$$\lim_n \frac{Q(\sqrt{2\log n}\sigma_n^{-1})}{\log n} = 0,$$

and

$$\lambda^*(\delta) - \frac{1-\delta}{2} = \lim_n h_n(\delta) + \frac{1}{2} - \beta = \sup_{0\le t\le 1} \delta\left[\frac{1}{2} - (1-t)^2\right] + \frac{1}{2} - \beta = \frac{\delta}{2} + \frac{1}{2} - \beta.$$

Hence the supremum of $\lambda^*(\delta) - \frac{1-\delta}{2}$ on $\delta \in [0,1]$ is attained at $\delta = 1$.

(ii) $\limsup \sqrt{2\log n}\sigma_n^{-1} = \infty$.

Note that the limit $\lambda^*(\delta)$ exists for any δ. Therefore, by considering the sub-sequence of σ_n with $\sqrt{2\log n}\sigma_n^{-1} \to \infty$, we can assume without loss of generality that $\sqrt{2\log n}\sigma_n^{-1} \to \infty$. To prove the desired inequality, it suffices to show that, for any $\epsilon > 0$, there exists $\bar{n}(\epsilon) \in \mathbb{N}$ such that

$$h_n(\delta) \le \max\{h_n(1), 0\} + \epsilon, \quad \text{for all } \delta \in (0,1),\ n > \bar{n}(\epsilon).$$

Fix $\epsilon > 0$. Like part (a), we will prove this by supposing that $h_n(\delta) > \max\{h_n(1), 0\} + \epsilon$ for all n and some δ, and deriving a contradiction. Suppose that for any $N > 0$, there exists $n > N$ and $(\delta_n^*, t_n^*) \in [0,1) \times [0,1]$ such that

$$g_n(\delta_n^*, t_n^*) > \max\{\epsilon, h_n(1) + \epsilon\}.$$

To make use of the regularly varying property, we need to first obtain upper and lower bound for $t_n^*\sqrt{\delta_n^*}$. Since

$$g_n(\delta_n^*, t_n^*) = \delta_n^*\left[\frac{1}{2} - (1-t_n^*)^2\right] - \frac{Q(t_n^*\sqrt{2\delta_n^*\log n}\sigma_n^{-1})}{\log n} > \epsilon \tag{5.17}$$

and Q is non-negative, we have $\delta_n^* > 2\epsilon$ and $t_n^* > 1 - \sqrt{2}/2 > 1/4$. Therefore

$$\frac{1}{2t_n^*\sqrt{\delta_n^*}} < \sqrt{\frac{2}{\epsilon}}.$$

On the other hand, we have $g_n(\delta_n^*, t_n^*) > h_n(1) + \epsilon \ge g_n(1, 1/2) + \epsilon$, that is,

$$\delta_n^*\left[\frac{1}{2} - (1-t_n^*)^2\right] - \frac{Q(t_n^*\sqrt{2\delta_n^*\log n}\sigma_n^{-1})}{\log n} > \frac{1}{4} + \epsilon - \frac{Q(\frac{1}{2}\sqrt{2\log n}\sigma_n^{-1})}{\log n}. \tag{5.18}$$

Following the first claim (see Equation 5.16), we have

$$\delta_n^* \left[\frac{1}{2} - (1-t_n^*)^2\right] \leq \frac{1}{4} < \frac{1}{4} + \epsilon.$$

Comparing the two equations above and noting that Q is non-decreasing, we have

$$\frac{Q(t_n^*\sqrt{2\delta_n^* \log n}\sigma_n^{-1})}{\log n} \leq \frac{Q(\frac{1}{2}\sqrt{2\log n}\sigma_n^{-1})}{\log n}.$$

Therefore

$$t_n^*\sqrt{\delta_n^*} \leq \frac{1}{2}.$$

Using properties of regularly varying functions [17], we know that

$$\lim_{s\to\infty} \sup_{a\in\Gamma} \left|\frac{Q(as)}{Q(s)} - a^2\right| \to 0$$

for any compact set Γ. Therefore, for any $c_0 > 0$ there exists $S > 0$ such that

$$\frac{Q(as)}{Q(s)} \leq a^2 + c_0$$

for any $s > S$ and $a \in \left[1, \sqrt{\frac{2}{\epsilon}}\right]$. Take $s = t_n^*\sqrt{2\delta_n^* \log n}\sigma_n^{-1}$. Since $\sqrt{2\log n}\sigma_n^{-1} \to \infty$, we know that for large enough n,

$$\frac{Q(\frac{1}{2}\sqrt{2\log n}\sigma_n^{-1})}{Q(t_n^*\sqrt{2\delta_n^* \log n}\sigma_n^{-1})} \leq \frac{1}{4t_n^{*2}\delta_n^*} + c_0. \tag{5.19}$$

Combining Equations 5.17, 5.18 and 5.19, we have

$$\left(\frac{1}{4t_n^{*2}\delta_n^*} + c_0\right) \delta_n^* \left[\frac{1}{2} - (1-t_n^*)^2\right] > \frac{1}{4} + \epsilon.$$

Since c_0 is arbitrary, we can take $c_0 < \epsilon$. It follows that

$$\epsilon > c_0 t_n^{*2}\delta_n^* > c_0\delta_n^* \left[\frac{1}{2} - (1-t_n^*)^2\right],$$

and the above equation yields

$$\frac{1}{t_n^{*2}} \left[\frac{1}{2} - (1-t_n^*)^2\right] > 1,$$

a contradiction, and the second claim is proved.

□

Proof of Corollary 1

Proof. Recall the definition of λ_n. For the first part, it suffices to notice that

$$\begin{aligned}\lim_n \lambda_n(1) &= \lim_n \sup_{0\le t\le 1} -\frac{\nu\log(t\sqrt{2\log n}\sigma_n^{-1})}{\log n} - (1-t)^2 + 1 - \beta \\ &= \lim_n \sup_{0\le t\le 1} -\frac{\nu\log(\sigma_n^{-1})}{\log n} - (1-t)^2 + 1 - \beta \\ &= \nu\rho + 1 - \beta.\end{aligned}$$

Therefore $\beta^*(\rho) = \nu\rho + 1$.

For the second part, note that

$$\lim_n \lambda_n(1) = \sup_{0\le t\le 1} -\frac{2at}{r} - (1-t)^2 = \sup_{0\le t\le 1} -[t-(1-\frac{a}{r})]^2 + (1-\frac{a}{r})^2 - \beta.$$

Since $0 < 1 - \frac{a}{r} < 1$, it follows that

$$\lim_n \lambda_n(1) = (1-\frac{a}{r})^2 - \beta > 0 \quad \Longleftrightarrow \quad r > \frac{a}{(1-\sqrt{\beta})}.$$

Therefore

$$\beta^*(r) = (1-\frac{a}{r})^2.$$

For the third part, by Lemma 3,

$$\lim_n \lambda_n(1) = \lim_{n\to\infty} \sup_{0\le t\le 1} \frac{\log[1-\Phi(t\sqrt{2\log n}r^{-1}-\mu)]}{2\log n} - (1-t)^2 + 1 - \beta.$$

Using properties of Gaussian tail probability, it can be easily verified that

$$\lim_{n\to\infty} \frac{\log\left(1-\Phi(t\sqrt{2\log n}r^{-1}-\mu)\right)}{2\log n} = -\frac{t^2}{\sigma^2 r^2} \quad \text{uniformly on } t\in[0,1].$$

Therefore

$$\lim_{n\to\infty} \lambda_n(1) > 0 \quad \Longleftrightarrow \quad \sup_{0\le t\le 1} -\left[\frac{t^2}{r^2} + (1-t)^2\right] > 1-\beta,$$

and $\beta^*(r) = \frac{r^2}{r^2+1}$. □

Proof of Theorem 2

Next, we restate and prove Theorem 3:

Theorem 3. *Suppose that G satisfies $\lim_{\mu\to\infty}(1-G(\mu))\mu^\nu = \lim_{\mu\to\infty} G(-\mu)\mu^\nu = C$ with tail index $\nu > 0$, and $\sigma_n = \frac{r\sqrt{2\log n}}{n^{(1-\beta)/\nu}}$ for some $\beta \in (1/2, 1)$. Then $\beta^* = \beta$, and*

1. *the asymptotic power of the level-α max test, is*

$$\lim_{n\to\infty} \mathbb{P}_{H_1}(\textit{reject } H_0) = 1 - e^{-2Cr^\nu + \log(1-\alpha)}.$$

In particular, the power tends to 1 as $r \to \infty$.

2. *for any $r \in (0, \infty)$, the modified higher criticism is asymptotically powerless.*

Proof. We first improve on Lemma 3 and derive a tighter bound on the tail probabilities of the alternative distribution. For any $0 < \delta \leq 1$, we have

$$\begin{aligned}\mathbb{P}_{\mu_n \sim G_n}(X \geq \sqrt{2\delta \log n}) &= \int_{-\infty}^{\infty} \left(1 - G\left(\frac{\sqrt{2\delta\log n} - z}{\sigma_n}\right)\right)\phi(z)dz. \\ &= \int_{-\infty}^{\sqrt{2\delta\log n}-1} \left(1 - G\left(\frac{\sqrt{2\delta\log n} - z}{\sigma_n}\right)\right)\phi(z)dz \\ &\quad + \int_{\sqrt{2\delta\log n}-1}^{\infty} \left(1 - G\left(\frac{\sqrt{2\delta\log n} - z}{\sigma_n}\right)\right)\phi(z)dz \end{aligned} \tag{5.20}$$

Because $\sigma_n \to 0$, the tail approximation for $1 - G(\mu)$ holds uniformly for $\mu > 1/\sigma_n$. Thus, we can approximate the first term in (5.20) as

$$\int_{-\infty}^{\delta_n} \left(1 - G\left(\frac{\sqrt{2\delta\log n} - z}{\sigma_n}\right)\right)\phi(z)dz \Big/ \int_{-\infty}^{\delta_n} C\left(\frac{\sigma_n}{\sqrt{2\delta\log n} - z}\right)^\nu \phi(z)dz \to 1,$$

as $n \to \infty$, where $\delta_n = \sqrt{2\delta \log n} - 1$. It is also straightforward to show that, as $n \to \infty$,

$$\begin{aligned}&\int_{-\infty}^{-(2\delta\log n)^{1/4}} \left(\frac{\sqrt{2\delta\log n}}{\sqrt{2\delta\log n} - z}\right)^\nu \phi(z)dz \\ \leq\ &\left(\frac{\sqrt{2\delta\log n}}{\sqrt{2\delta\log n} + (2\delta\log n)^{1/4}}\right)^\nu \Phi(-(2\delta\log n)^{1/4}) \ \to\ 0, \\ &\int_{-(2\delta\log n)^{1/4}}^{(2\delta\log n)^{1/4}} \left(\frac{\sqrt{2\delta\log n}}{\sqrt{2\delta\log n} - z}\right)^\nu \phi(z)dz \ \to\ 1, \qquad \text{and} \\ \int_{(2\delta\log n)^{1/4}}^{\sqrt{2\delta\log n}-1} &\left(\frac{\sqrt{2\delta\log n}}{\sqrt{2\delta\log n} - z}\right)^\nu \phi(z)dz \ \leq\ \left(\sqrt{2\delta\log n}\right)^\nu \left(1 - \Phi((2\delta\log n)^{1/4})\right) \ \to\ 0.\end{aligned}$$

As a result, we have

$$\int_{-\infty}^{\sqrt{2\delta\log n}-1} C\left(\frac{\sigma_n}{\sqrt{2\delta\log n} - z}\right)^\nu \phi(z)dz \Big/ \left(C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}}\right) \to 1. \tag{5.21}$$

Let $\epsilon_0 = \min\{\beta/2 - 1/4, 1/2 - \beta/2\}$. Turning to the second term in (5.20), we have

$$0 \leq \int_{\sqrt{2\delta\log n}-1}^{\infty} \left(1 - G\left(\frac{\sqrt{2\delta\log n} - z}{\sigma_n}\right)\right)\phi(z)dz \leq 1 - \Phi\left(\sqrt{2\delta\log n} - 1\right) \leq \frac{1}{n^{\delta-\epsilon_0}}. \tag{5.22}$$

Combining (5.20)–(5.22) and recalling the definition of σ_n, we have

$$\begin{aligned}(1+o(1))C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}} &\le \mathbb{P}_{\mu_n\sim G_n}(X \ge \sqrt{2\delta\log n}) \\ &\le (1+o(1))C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}} + \frac{1}{n^{\delta-\epsilon_0}}\end{aligned} \tag{5.23}$$

For $\delta > 1-\beta+\epsilon_0$ we have

$$\left(\frac{1}{n^{\delta-\epsilon_0}}\right) \Bigg/ \left(C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}}\right) = O\left(n^{1-\beta-\delta+\epsilon_0}\right) \to 0.$$

Therefore for $\delta > 1-\beta+\epsilon_0$,

$$\mathbb{P}_{\mu_n\sim G_n}(X_n \ge \sqrt{2\delta\log n}) \Bigg/ \left(C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}}\right) \to 1.$$

Similarly,

$$\mathbb{P}_{\mu_n\sim G_n}(X_n \le -\sqrt{2\delta\log n}) \Bigg/ \left(C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}}\right) \to 1.$$

Therefore for $\delta > 1-\beta+\epsilon_0$,

$$\mathbb{P}_{\mu_n\sim G_n}(|X_n| \ge \sqrt{2\delta\log n}) \Bigg/ \left(2C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}}\right) \to 1. \tag{5.24}$$

Suppose that the $1-\alpha$ quantile of $\max_i |X_i|$ under the null is $m(n,\alpha)$. Then the level-α max test rejects the null when $\max_i |X_i| > m(n,\alpha)$. Since $m(n,\alpha)/\sqrt{2\log n} \to 1$, we have

$$\mathbb{P}_{\mu_n\sim G_n}(|X_n| \ge m(n,\alpha)) \Bigg/ \left(2C\frac{(r/\sqrt{\delta})^\nu}{n^{1-\beta}}\right) \to 1$$

and

$$n^{1-\beta}\mathbb{P}_{\mu_n\sim G_n}(|X_n| \ge m(n,\alpha)) \to 2Cr^\nu.$$

Hence the level-α max test satisfies

$$\begin{aligned}&\mathbb{P}_{H_1}(\text{reject } H_0) \\ =&1-\left(1-(1-n^{-\beta})\mathbb{P}(|N(0,1)|\ge m(n,\alpha)) - n^{-\beta}\mathbb{P}_{\mu_n\sim G_n}(|X_n|\ge m(n,\alpha))\right)^n \\ =&1-\left(1-(1-n^{-\beta})(1-(1-\alpha)^{1/n}) - n^{-\beta}\mathbb{P}_{\mu_n\sim G_n}(|X_n|\ge m(n,\alpha))\right)^n \\ \to&1-e^{-2Cr^\nu+\log(1-\alpha)}, \quad \text{as } n\to\infty,\end{aligned} \tag{5.25}$$

and the first part of the proposition is proved.

Next we show that modified higher criticism is asymptotically powerless. For modified higher criticism, the critical value of the test $b(n,\alpha) \sim \sqrt{2\log\log n}$. Let $p_i = \mathbb{P}(|N(0,1)| \geq |X_i|), i = 1,\ldots,n$ be the p-values. Suppose that under H_1, the p-values are $i.i.d$ with distribution function F_n. Let

$$\widehat{F_n}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(p_i \leq t)},$$

be the empirical distribution of $\{p_i\}, i = 1,\ldots,n$. Let $\tilde{p}_i = F_n(p_i)$, and

$$\widetilde{F_n}(t) = \widehat{F_n}(F_n^{-1}(t)) = \frac{1}{n}\sum_{i=1}^{n} 1_{(F_n(p_i)\leq t)} = \frac{1}{n}\sum_{i=1}^{n} 1_{(\tilde{p}_i \leq t)},$$

Then $\tilde{p}_i \overset{i.i.d.}{\sim} \text{Unif}[0,1]$, and $\{\widetilde{F_n}(t), 0 \leq t \leq 1)\}$ follows the same distribution as the empirical distribution of $\{p_i\}, i = 1,\ldots,n$ under the null. Note that the higher criticism statistics can be decomposed as

$$\begin{aligned}
&\sup_{1/n<t<1/2} \frac{\sqrt{n}(\widehat{F_n}(t) - t)}{\sqrt{t(1-t)}} \\
=& \sup_{F_n(1/n)<t<F_n(1/2)} \frac{\sqrt{n}(\widehat{F_n}(F_n^{-1}(t)) - F_n^{-1}(t))}{\sqrt{F_n^{-1}(t)(1-F_n^{-1}(t))}} \\
=& \sup_{F_n(1/n)<t<F_n(1/2)} \frac{\sqrt{n}(\widetilde{F_n}(t) - F^{-1}(t))}{\sqrt{F_n^{-1}(t)(1-F_n^{-1}(t))}} \\
=& \sup_{F_n(1/n)<t<F_n(1/2)} \left(\sqrt{\frac{t(1-t)}{F_n^{-1}(t)(1-F_n^{-1}(t))}} \frac{\sqrt{n}(\widetilde{F_n}(t) - t)}{\sqrt{t(1-t)}} + \frac{\sqrt{n}(t - F_n^{-1}(t))}{\sqrt{F^{-1}(t)(1-F_n^{-1}(t))}} \right).
\end{aligned}$$

We denote

$$A_n(t) := \sqrt{\frac{t(1-t)}{F_n^{-1}(t)(1-F_n^{-1}(t))}},$$

$$B_n(t) := \frac{\sqrt{n}(t - F_n^{-1}(t))}{\sqrt{F_n^{-1}(t)(1-F_n^{-1}(t))}}$$

and

$$W_n(t) := \frac{\sqrt{n}(\widetilde{F_n}(t) - t)}{\sqrt{t(1-t)}}.$$

Note that by Taylor expansion,

$$A_n(F_n(t)) - 1 \leq \frac{1}{2}\frac{F_n(t) - t}{t}, \quad \text{for any } t > 0.$$

Let $D(t) = F_n(t) - t = n^{-\beta}\left(\mathbb{P}_{\mu_n \sim G_n}(|X_n| \geq \Phi^{-1}(1-t/2)) - t\right)$ and $q_n = (\log n)^3/2n$. Let $\delta_0 = 1 - \epsilon$ for $\epsilon > 0$ small enough. Then for large enough n, by Equation 5.24

$$n \sup_{1/n \leq t \leq q_n} D(t) \leq n^{1-\beta}\left(\mathbb{P}_{\mu_n \sim G_n}\left(|X| \geq \sqrt{2\delta_0 \log n}\right)\right) \to 2Cr^\nu \delta_0^{-\nu/2} \leq 4Cr^\nu.$$

For large enough n, we have

$$\sup_{F_n(1/n) < t < F_n(q_n)} A_n(t) \leq 1 + \frac{n}{2} \sup_{1/n \leq t \leq q_n} D(t) \leq 2Cr^\nu$$

and

$$\sup_{F_n(1/n) < t < F_n(q_n)} B_n(t) \leq n \sup_{1/n \leq t \leq q_n} D(t) \leq 4Cr^\nu.$$

Note that $1/n \leq F_n(1/n)$ and

$$F_n(q_n) \leq q_n + D(q_n) \leq q_n + 4Cr^\nu/n \leq q_n + (\log n)^3/2n = (\log n)^3/n.$$

Lemma 3 and 4 in Jaeschke [59] implies that

$$\sup_{F_n(1/n) < t < F_n(q_n)} W_n(t)/\sqrt{2\log\log n} \leq \sup_{1/n < t < (\log n)^3/n} W_n(t)/\sqrt{2\log\log n} \xrightarrow{p} 0.$$

Therefore

$$\mathbb{P}\left(\sup_{F_n(1/n) < t < F_n(q_n)} A_n(t)W_n(t) + B_n(t) > b(n,\alpha)\right) = 0.$$

Write $t = 2(1 - \Phi(\sqrt{2\delta \log n}))$ for $0 < \delta < 1$. Then $t \sim n^{-\delta}$ up to $\log n$ factors. Recall that $\epsilon_0 = \min\{\beta/2 - 1/4, 1/2 - \beta/2\}$. It can be easily verified from Equation 5.23 that

$$F_n(t) - t \leq \begin{cases} (1+o(1))C(\frac{r}{\sqrt{\delta}})^\nu n^{-1} & \text{for } 1 - \beta + \epsilon_0 \leq \delta < 1, \\ n^{-(\beta+\delta-\epsilon_0)} & \text{for } \epsilon_0 < \delta < 1 - \beta + \epsilon_0, \\ n^{-\beta} & \text{for } \delta < \epsilon_0, \end{cases}$$

Let $q_n^* = 2(1 - \Phi(\sqrt{2(1-\beta+\epsilon_0)\log n}))$. Then $t \geq n^{-\delta-\epsilon_0}$, and it follows that for some constant C_0,

$$\begin{aligned} &\sup_{F(q_n) < t < F(1/2)} (A_n(t) - 1) \\ \leq &\max\left\{\sup_{q_n < t < q_n^*} A_n(F_n(t)) - 1, \sup_{q_n^* < t < 1/2} A_n(F_n(t)) - 1\right\} \\ \leq &\max\left\{\frac{C_0}{(\log n)^3}, n^{-\beta+2\epsilon_0}\right\} = O\left((\log n)^{-3}\right). \end{aligned}$$

Similarly we have

$$\begin{aligned}
&\sup_{F_n(q_n)<t<F_n(1/2)} B_n(t) \\
\leq &\max\left\{\sup_{q_n<t<q_n^*} B_n(F_n(t)), \sup_{q_n^*<t<1/2} B_n(F_n(t))\right\} \\
\leq &\max\left\{\frac{C_0}{(\log n)^{3/2}}, n^{-\frac{1}{2}\beta+\frac{1}{4}}\right\} = O\left((\log n)^{-3/2}\right).
\end{aligned}$$

Therefore by Theorem 1 in Jaeschke [59], for large enough n we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{F_n(q_n)<t<F_n(1/2)} A_n(t)W_n(t) + B_n(t) > b(n,\alpha)\right) \\
\leq &\mathbb{P}\left(\sup_{F_n(q_n)<t<F_n(1/2)} W_n(t) > b(n,\alpha) - \frac{C_0(b(n,\alpha)+1)}{(\log n)^{3/2}}\right) \\
\leq &\mathbb{P}\left(\sup_{0<t<F_n(1/2)} W_n(t) > b(n,\alpha) - \frac{1}{\log n}\right) \to \alpha,
\end{aligned}$$

and the proof is complete.

□

5.6 Supplementary results

Counterexample showing that the condition in Theorem 2 is almost necessary

Suppose that $\sigma_n = r/\sqrt{2\log n}, r > 0$ and $G(\theta)$ is the distribution with $\mathbb{P}(\theta = 3^m) = e^{-3^m}, m = 1, 2, \ldots,$ and $\mathbb{P}(\theta = 0) = 1 - \sum_{m=1}^{\infty} e^{-3^m}$. Let $\beta = 0.52$ and $n_k = e^{5\cdot 3^k}, k = 1, 2, \ldots.$ Then we have

$$\mathbb{P}\left(\mu_{n_k} = \sqrt{2\cdot(0.2\cdot 3^m r)^2 \log n_k}\right) = n_k^{-(0.52+0.2\cdot 3^m)}, m = -k, \ldots, -1, 0, 1, \ldots.$$

For $m \geq 1$, the probability is less than $n_k^{-1.1}$, and the corresponding signal can not be used for detection. For $m \leq 0$, we have $0.52 + 0.2\cdot 3^m \leq 0.72 < 0.75$. Therefore for max test to have full asymptotic power, we need

$$(0.2\cdot 3^m r)^2 > \left(1 - \sqrt{1-(0.52+0.2\cdot 3^m)}\right)^2 \quad \text{for some integer } m \leq 0 \Rightarrow r > 2.354.$$

For the higher criticism to have full power [35], we need

$$(0.2\cdot 3^m r)^2 > 0.52 + 0.2\cdot 3^m - 0.5 \quad \text{for some integer } m \leq 0 \Rightarrow r > 2.345.$$

Since $2.345 < 2.354$, the detection boundary for higher criticism is smaller than that of max test despite F being exponential.

Proof of Propositions 2 and 3(b) for the modified higher criticism and Berk-Jones tests

Since Proposition 2 is a directly corollary of Proposition 3, we will only provide the proof of Part (b) of Proposition 3 for the modified higher criticism and Berk-Jones tests.

Proof. Under the condition of Part (b), there exists $\delta_0 \in (0,1)$ and constant $c_0 > 0$ such that $\lim_{n\to\infty} \lambda_n(\delta_0) - \frac{1-\delta_0}{2} = 2c_0 > 0$ for large enough n. Let $t = \bar{\Phi}(2\delta_0 \log_n) < n^{-\delta_0}$. Recall that F_n is the empirical distribution of p-values. Therefore $nF_n(t) = N(\delta_0)$ follows a binomial distribution with

$$\mathbb{E}_{H_1} N(\delta_0) = nt(1-n^{-\beta}) + n^{\lambda_n(\delta_0)} \geq nt(1-n^{-\beta}) + n^{\frac{1-\delta_0}{2}+c_0} \geq nt + \frac{1}{2} n^{\frac{1-\delta_0}{2}+c_0}.$$

for large enough n, and

$$\text{Var}_{H_1} N(\delta_0) = \mathbb{E}_{H_1} N(\delta_0) \left(1 - \frac{\mathbb{E}_{H_1} N(\delta_0)}{n}\right)$$

Therefore, by Chebyshev's inequality,

$$\begin{aligned}
\mathbb{P}_{H_1}[N(\delta_0) < nt + n^{\frac{1-\delta+c_0}{2}}] &\leq \frac{\text{Var}_{H_1} N(\delta_0)}{\left(\mathbb{E}_{H_1} N(\delta_0) - nt - n^{\frac{1-\delta_0+c_0}{2}}\right)^2} \\
&\leq \frac{\mathbb{E}_{H_1} N(\delta_0)}{\left(\mathbb{E}_{H_1} N(\delta_0) - nt - n^{\frac{1-\delta_0+c_0}{2}}\right)^2} \\
&\leq \frac{1}{\mathbb{E}_{H_1} N(\delta_0) - nt - 2n^{\frac{1-\delta_0+c_0}{2}}} \\
&\leq n^{-\frac{1-\delta_0+c_0}{2}}
\end{aligned}$$

for large enough n. Therefore, for the modified higher criticism statistics, we have

$$\begin{aligned}
\mathbb{P}_{H_1}(mHC_n \geq 2\sqrt{\log\log n}) &\geq \mathbb{P}_{H_1}\left(\frac{\sqrt{n}(F_n(t)-t)}{\sqrt{t(1-t)}} \geq 2\sqrt{\log\log n}\right) \\
&\geq \mathbb{P}_{H_1}\left(\frac{N(\delta_0)-nt}{\sqrt{nt(1-t)}} \geq 2\sqrt{\log\log n}\right) \\
&\geq 1 - \mathbb{P}_{H_1}\left[N(\delta_0) < nt + n^{\frac{1-\delta+c_0}{2}}\right] \to 1
\end{aligned}$$

as the $n \to \infty$, where the last inequality holds for large enough n. Now we turn to the Berk-Jones statistics. First, it can be easily verified that $\log(x+1) > x/2$ for $x \in (-1/2, 1/2)$. Without

loss of generality, suppose that $2c_0 < (1-\delta_0)/2$, then $\mathbb{E}_{H_1} N(\delta_0)/nt \to 1$, and $F_n(t)/t \xrightarrow{p} 1$. If $1/2 < F_n(t)/t < 3/2$, then

$$\begin{aligned} & 2n\left[F_n(t)\log\frac{F_n(t)}{t} + (1-F_n(t))\log\frac{(1-F_n(t))}{(1-t)}\right] \\ \geq\ & nF_n(t)\left(\frac{F_n(t)}{t}-1\right) + n(1-F_n(t))\left(\frac{(1-F_n(t))}{(1-t)}-1\right) = \frac{n(F_n(t)-t)^2}{t(1-t)}. \end{aligned}$$

Therefore

$$\lim_{n\to\infty}\mathbb{P}_{H_1}(BJ_n \geq 2\sqrt{\log\log n}) \geq \lim_{n\to\infty}\mathbb{P}_{H_1}\left(\frac{n(F_n(t)-t)^2}{t(1-t)} \geq 4\log\log n\right) = 1,$$

which completes the proof. □

Additional simulation results

We provide additional simulation results where G is the Gaussian (Figure 5.4), logistic (Figure 5.5), chi-squared (Figure 5.6), t_5 (Figure 5.7), and t_3 (Figure 5.8) distribution, and $G_n = rG$. In each simulation, $n = 50,000$ and there are $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from G_n. We find that in all settings, the power of max test is similar to the power of the higher criticism test when $\beta > 1/2$.

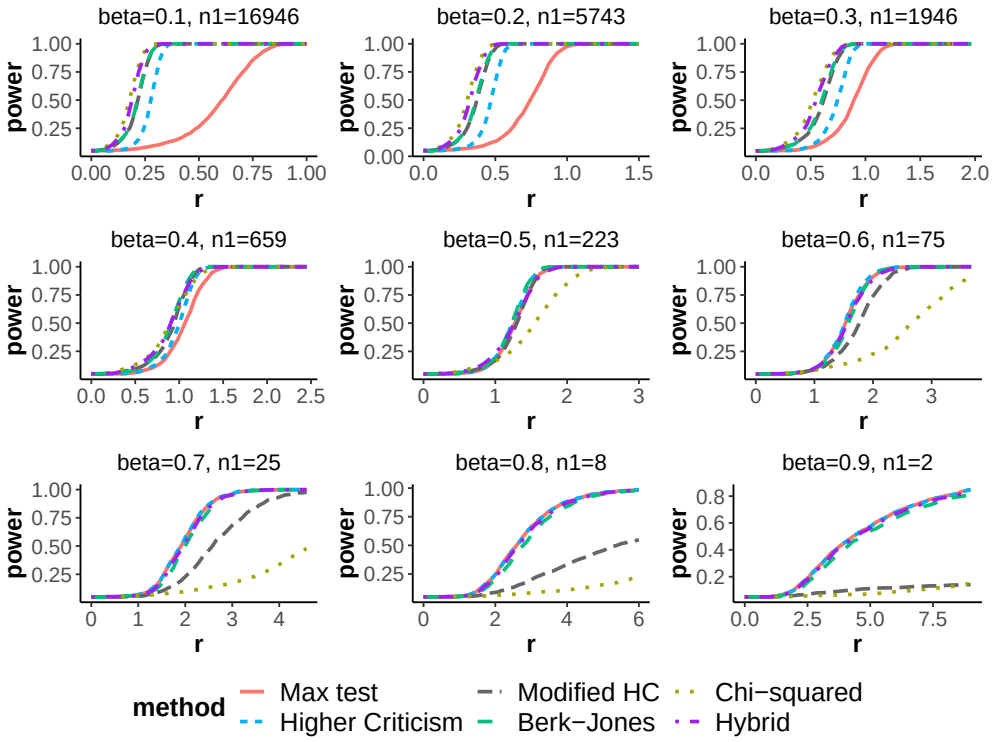


Figure 5.4: Comparison of power for different tests, where $n = 50,000$ with $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from $N(0, r^2)$.

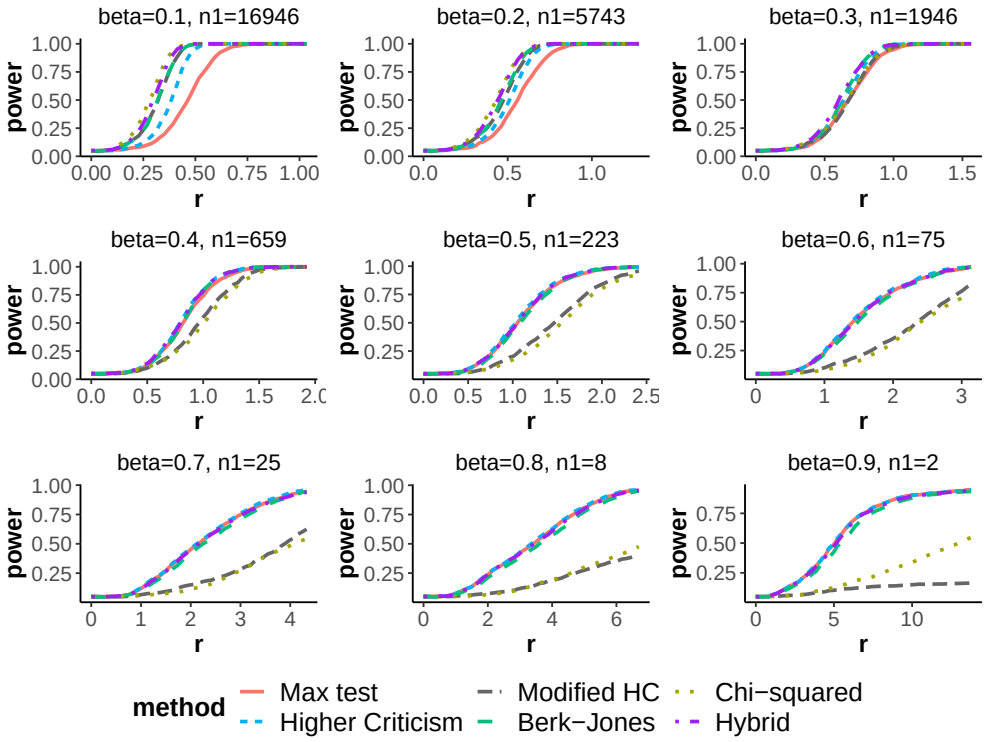


Figure 5.5: Comparison of power for different tests, where $n = 50,000$ with $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from $r * \text{Logistic}(0, 1)$.

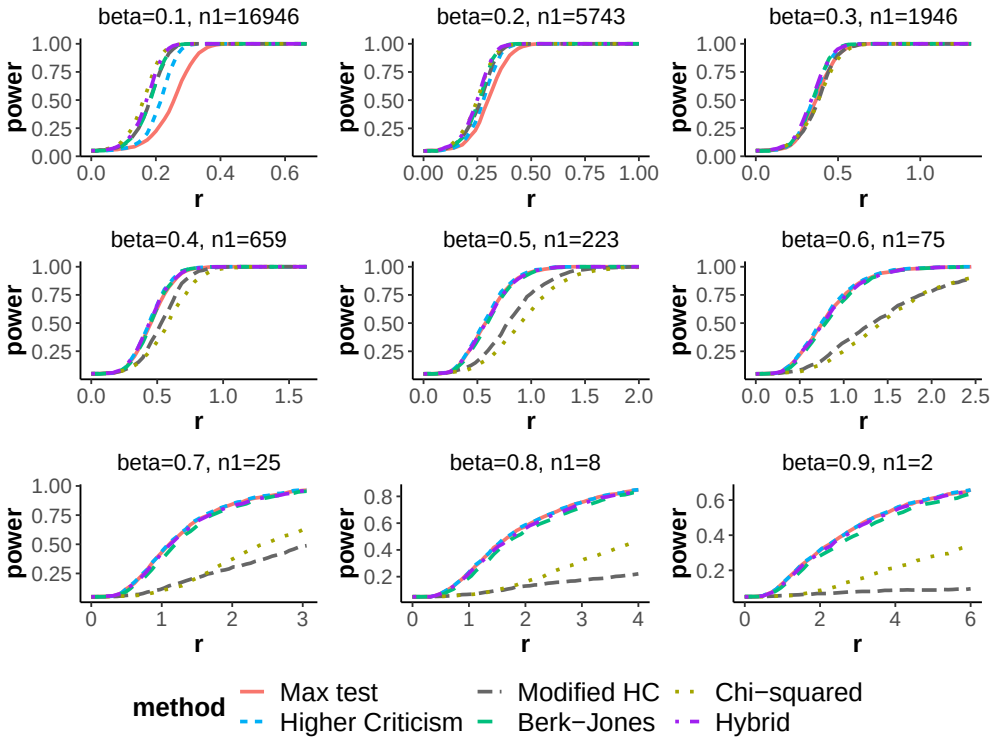


Figure 5.6: Comparison of power for different tests, where $n = 50,000$ with $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from $r * \chi^2(1)$.

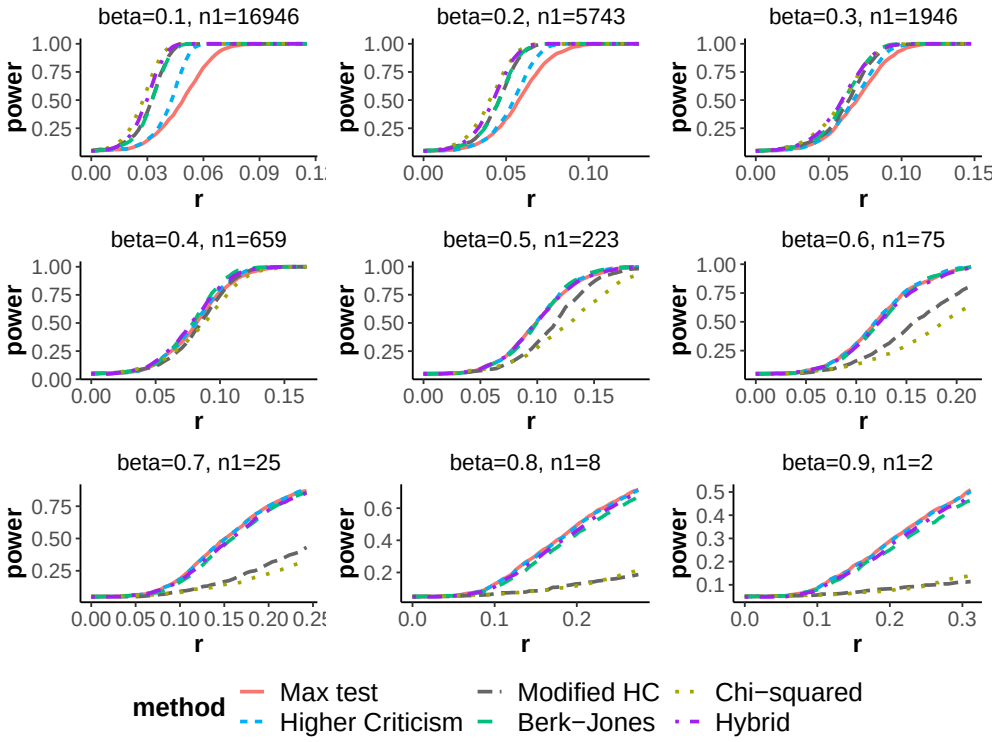


Figure 5.7: Comparison of power for different tests, where $n = 50,000$ with $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from $r * t_5$.

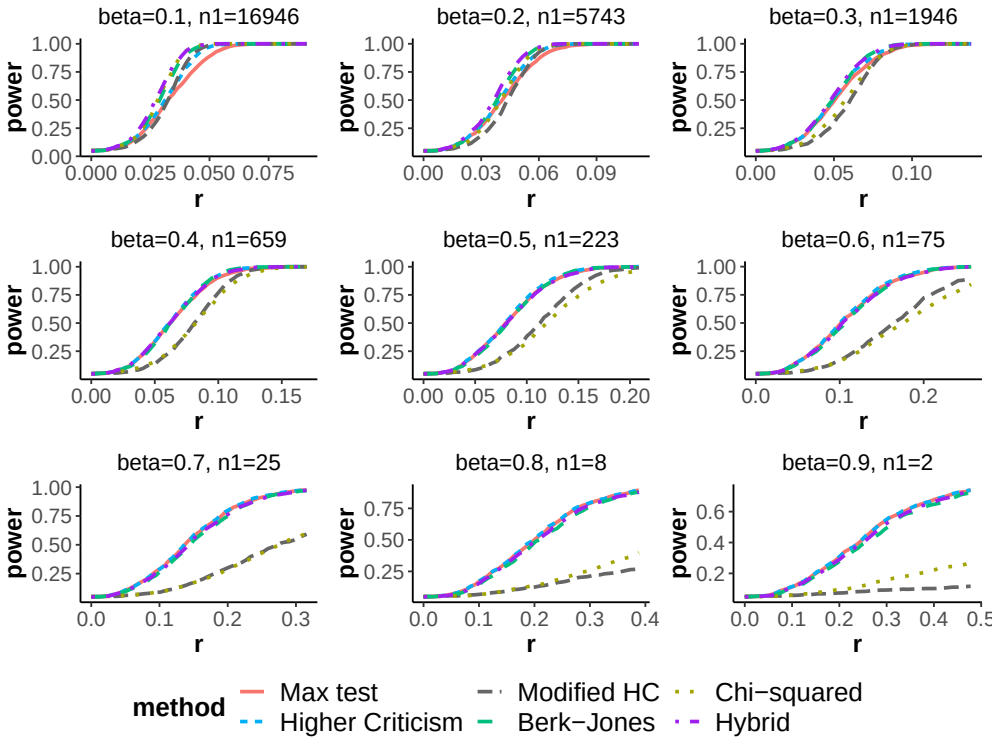


Figure 5.8: Comparison of power for different tests, where $n = 50,000$ with $n_1 = \lfloor n^{1-\beta} \rfloor$ non-null means drawn from $r * t_3$.

Chapter 6

Whiteout: when do fixed-X knockoffs fail?

6.1 Introduction

The knockoff filter in the Gaussian linear model

Knockoff methods are a flexible framework for multiple testing in supervised learning problems, that operate by introducing a "negative control" for each predictor variable in the model and testing an algorithm's ability to distinguish the true variables from the controls. In particular, the *fixed-X knockoff filter* introduces extra predictor variables in a Gaussian linear model whose joint correlation structure with each other and with the true predictors makes them appear indistinguishable from true predictors whose regression coefficients are zero. In this model we observe a fixed design matrix $X \in \mathcal{R}^{n \times d}$ and response

$$y = X\beta + \varepsilon, \quad \text{where } \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n,$$

where $\sigma^2 > 0$ and $\beta \in \mathcal{R}^d$ are unknown parameters, and the goal is to test each null hypothesis $H_j : \beta_j = 0$ against the two-sided alternative, for $j = 1, \ldots, d$. Following Barber, Candès, et al. [9], we assume throughout that X has full column rank with $2d \leq n$.

The uniformly most powerful unbiased (UMPU) test of H_j is the two-sided t-test that rejects for extreme values of the t-statistic $T_j = \hat{\beta}_j / \sqrt{\hat{\sigma}^2 \Sigma_{jj}}$, where $\Sigma = (X^\mathsf{T} X)^{-1}$ and $\hat{\beta}$ and $\hat{\sigma}$ are respectively the OLS estimator

$$\hat{\beta} \;=\; \left(X^\mathsf{T} X\right)^{-1} X^\mathsf{T} y \;\sim\; \mathcal{N}_d\left(\beta,\, \sigma^2 \Sigma\right)$$

and the residual variance $\hat{\sigma}^2 = \|y - X\hat{\beta}\|^2/(n-d)$. Taken together, these two estimators are a complete sufficient statistic for the model. Let p_j denote the p-value for the two-sided t-test on H_j.

The standard approach for multiple testing would operate on the t-test p-values for these tests, correcting for multiplicity. If R is the number of rejected hypotheses from a multiple testing procedure and V is the number of rejected true null hypotheses (false discoveries), [16] define the *false discovery proportion* as FDP $= V/\max\{R, 1\}$, and the *false discovery rate* (FDR) as

its expectation FDR $= \mathbb{E}\,\text{FDP}$. While the Benjamini–Hochberg (BH) procedure of [16] is not known to control the FDR in this problem unless Σ is diagonal (i.e., unless the columns of X are orthogonal), recent methods can directly adjust BH for the multivariate Gaussian dependence [43]. The FDR criterion relaxes the more conservative *family-wise error rate* FWER $= \mathbb{P}(V \geq 1)$, the probability of making any false rejections, which we can control using the conservative Bonferroni correction that rejects H_j when $p_j \leq \alpha/m$.[1]

The knockoff filter of Barber, Candès, et al. [9] takes a radically different approach, bypassing the t-test p-values entirely. The method begins by augmenting the design matrix X with a second matrix $\widetilde{X} \in \mathcal{R}^{n\times d}$ of negative control or *knockoff* variables, constructed to satisfy

$$\widetilde{X}^\mathsf{T}\widetilde{X} = X^\mathsf{T}X, \quad \text{and} \quad \widetilde{X}^\mathsf{T}X = X^\mathsf{T}X - D,$$

for some diagonal matrix $D \preceq 2X^\mathsf{T}X$, where $\preceq$ denotes the positive semidefinite ordering. As we will see, a larger entry of D_{jj} preserves more signal for the inference on H_j, but in general it is not possible to maximize all D_{jj} simultaneously.

Knockoffs then calculates so-called W-statistics $W_1, \ldots, W_d$ satisfying two properties:

1. Sufficiency W_j depends on $[X\ \widetilde{X}]$ and y only through the Gram matrix $[X\ \widetilde{X}]^\mathsf{T}[X\ \widetilde{X}]$ and $[X\ \widetilde{X}]^\mathsf{T}y$, and

2. Antisymmetry Swapping any variable X_j with its knockoff $\widetilde{X}_j$ would flip the sign of W_j and leave every other W_k fixed. That is, if $\text{Swap}_j([X\ \widetilde{X}])$ is the augmented design matrix with these variables swapped, then

$$W_k(\text{Swap}_j([X\ \widetilde{X}]), y) \;=\; \begin{cases} -W_k([X\ \widetilde{X}], y) & k = j \\ W_k([X\ \widetilde{X}], y) & k \neq j \end{cases}.$$

The absolute values $|W| = (|W_1|, \ldots, |W_d|)$ determine a data-adaptive hypothesis ordering, with larger values assigned higher priority. As we will see, the sufficiency and antisymmetry properties along with the model assumptions ensure that, conditional on $|W|$, $\text{sgn}(W_1), \ldots, \text{sgn}(W_d)$ are mutually independent, and under H_j, $\text{sgn}(W_j)$ is a Rademacher random variable (without loss of generality we can assume all $W_j \neq 0$ by construction). Once the W-statistics are calculated, the knockoff filter applies an ordered multiple testing method called *selective SeqStep* [9] treating each $\text{sgn}(W_j)$ as a "binary p-value" for H_j. To control the false discovery rate (FDR) at level α, the knockoff+ method rejects all hypotheses H_j for which W_j exceeds the adaptive threshold $\hat{t}$:

$$\hat{t} = \min\left\{t : \widehat{\text{FDP}}_t^{\text{kn}} \leq \alpha\right\}, \quad \text{where} \quad \widehat{\text{FDP}}_t^{\text{kn}} = \frac{1 + \#\{j :\ W_j \leq -t\}}{\#\{j :\ W_j \geq t\}}.$$

The fixed-X knockoff filter is a highly versatile method that offers the user flexibility at two stages of the procedure: first, in choosing the matrix D from among the many matrices satisfying

[1] A more accurate FWER correction would apply the closure of the max-t test: if S is the maximal set for which $\|T_S\|_\infty$ is below its $1-\alpha$ quantile under $H_S :\ \beta_S = 0$, then we can reject H_j for $j \notin S$; see [55] for more details.

$D \preceq 2X^\mathsf{T}X$, and second, in choosing how to define the W-statistics. Spector and Janson [115] show the importance of choosing D well and discuss ramifications on the procedure's power, and various other works detail myriad ways to tailor the W-statistics using machine learning methods that exploit structural assumptions or other prior beliefs about the coefficients. In particular, at either stage the analyst can choose to favor some hypotheses over others by increasing their values of D_{jj} or $|W_j|$. This flexibility poses a major challenge if we wish to upper-bound the method's power universally over the analyst's entire choice set, since it gives well-informed analysts ample opportunities to stack the deck in their own favor.

In light of this tremendous flexibility, it may come as a surprise to discover regimes where *no* knockoffs method — even one designed with full knowledge of the true regression coefficients — can achieve nontrivial power, even while Bonferroni-corrected inference achieves near-perfect power. To explain where knockoffs can go wrong, the next Section 6.1 formally recasts the knockoff filter as a conditional post-selection inference method with an unrestricted exploratory stage followed by a prescribed confirmatory stage. The key quantity is a randomized estimator $\widetilde{\beta} = \hat{\beta} + \omega$, where ω is user-generated Gaussian noise in the style [123], with $\mathrm{Var}(\omega)$ crafted to make $\mathrm{Var}(\widetilde{\beta})$ diagonal. Section 6.1 shows the two formulations are equivalent, building off a conditioning argument introduced in the Supplement of Barber, Candès, et al. [8].

Knockoffs as conditional inference on a whitened estimator

We now give an alternative but equivalent account of the knockoff filter without W-statistics, without sufficiency and antisymmetry properties, and indeed without any negative control variables at all. Instead, we view the method as conditional post-selection inference, where the key step is to construct a *whitened estimator* $\widetilde{\beta}$ with diagonal covariance. If $\Delta \in \mathcal{R}^{d\times d}$ is any diagonal matrix with $\Delta \succeq \Sigma = (X^\mathsf{T}X)^{-1}$, let

$$\widetilde{\beta} \;=\; \hat{\beta} + \omega \;\sim\; \mathcal{N}_d\left(\beta,\, \sigma^2\Delta\right), \quad \text{where} \;\; \omega \sim \mathcal{N}_d\left(0,\, \sigma^2(\Delta - \Sigma)\right) \tag{6.1}$$

is noise generated by the user, independently of $\hat{\beta}$. We will see in Section 6.1 that even when σ^2 is unknown, ω can be carved out of $\hat{\sigma}^2$ as long as $n \geq d + r$, where $r = \mathrm{rank}(\Delta - \Sigma) \leq d$.

By whitening the estimator, we buy independence of the coordinates at the price of increasing their variance, since $\Delta_{jj} \geq \Sigma_{jj}$. However, this price is recouped by an exploratory analysis using the statistic

$$\xi \;=\; X^\mathsf{T}X\hat{\beta} - \Delta^{-1}\widetilde{\beta} \;\sim\; \mathcal{N}_d\left(A\beta,\, \sigma^2 A\right), \quad \text{where} \;\; A \;=\; X^\mathsf{T}X - \Delta^{-1} \succeq 0. \tag{6.2}$$

ξ carries the information lost by whitening since $\hat{\beta} = (X^\mathsf{T}X)^{-1}(\xi + \Delta^{-1}\widetilde{\beta})$, and is independent of $\widetilde{\beta}$ since

$$\mathrm{Cov}(\xi, \widetilde{\beta}) \;=\; (X^\mathsf{T}X)\,\mathrm{Cov}(\hat{\beta}, \hat{\beta} + \omega) - \Delta^{-1}\,\mathrm{Cov}(\widetilde{\beta}, \widetilde{\beta}) \;=\; \sigma^2 - \sigma^2 \;=\; 0.$$

In terms of $\omega, \widetilde{\beta}$ and ξ, the fixed-X knockoff filter can be equivalently defined as follows:

Stage 1 (whitening). For any $\Delta \succeq \Sigma$, generate noise $\omega \sim \mathcal{N}_d(0, \sigma^2(\Delta - \Sigma))$ independently of $\hat{\beta}$.

Stage 2 (exploratory analysis). Observe ξ and $|\widetilde{\beta}|$ and use them to order the d hypotheses for Selective SeqStep as $H_{[1]}, \ldots, H_{[d]}$, where $[1]$ indexes the first hypothesis in order and $[d]$ the last.

In addition, select a one-sided alternative for each H_j. Let $\psi_j = +1$ if the right-tailed alternative is selected, and $\psi_j = -1$ for the left-tailed alternative.

Stage 3 (confirmatory analysis). Using Selective SeqStep, test the hypotheses in order using $\text{sgn}(\widetilde{\beta}_j)$ as the test statistic for β_j. The signs are conditionally independent given ξ and $|\widetilde{\beta}|$, with

$$\text{logit}\ \mathbb{P}\left(\text{sgn}(\widetilde{\beta}_j) = +1 \mid \xi,\, |\widetilde{\beta}|\right) \;=\; \frac{2\,|\widetilde{\beta}_j|}{\sigma^2 \Delta_{jj}} \cdot \beta_j, \quad \text{where } \text{logit}\ p \,=\, \log \frac{p}{1-p}. \tag{6.3}$$

The equivalence between this formulation and the standard formulation of knockoffs given in Section 6.1 is formally stated and proven in Section 6.1. Because $\text{sgn}(\widetilde{\beta}_j)$ is conditionally a Rademacher random variable if $\beta_j = 0$, and is stochastically increasing in β_j, the conditional p-value $\widetilde{p}_j$ is $1/2$ when $\text{sgn}(\widetilde{\beta}_j) = \psi_j$, and 1 when $\text{sgn}(\widetilde{\beta}_j) = -\psi_j$. To be fully explicit, Stage 3 rejects $H_{[j]}$ if $j \le \hat{k}$ and $\widetilde{p}_{[j]} = 1/2$, where

$$\hat{k} = \max\left\{k :\ \widehat{\text{FDP}}_k^{\text{wh}} \le \alpha\right\}, \quad \text{with} \quad \widehat{\text{FDP}}_k^{\text{wh}} = \frac{1 + \sum_{j=1}^{k} 1\{\widetilde{p}_{[j]} = 1\}}{\sum_{j=1}^{k} 1\{\widetilde{p}_{[j]} = 1/2\}}.$$

Given ξ and $|\widetilde{\beta}|$, the p-values are valid and independent, satisfying the requirements for Selective SeqStep, so the method described above controls the directional FDR, both conditionally and marginally.

The exploratory analysis is defined vaguely because the analyst can use ξ and $|\widetilde{\beta}|$ however they like, provided they have not yet observed anything else about y.[2] It is in this unrestricted stage that knockoffs can exploit prior information and structural assumptions. When all goes well, the ordering is highly informative, effectively reducing the multiplicity in Stage 3 by focusing inferential power on the first few hypotheses. In many problems, a good exploratory analysis can more than compensate for the additional noise and binarization of the confirmatory test statistics, helping knockoffs to outperform less flexible methods like BH.

However, this formulation also exposes an important vulnerability of the method. When the eigenstructure of Σ is unfavorable, the price of whitening can be devastating, dooming the confirmatory analysis before the exploratory analysis even begins. Roughly, when Σ has large leading eigenvalues and dense leading eigenvectors, we will necessarily have $\Delta_{jj} \gg \Sigma_{jj}$ for most of the variables, rendering the conditional distribution of $\text{sgn}(\widetilde{\beta}_j)$ nearly uninformative about β_j even in regimes where Bonferroni inference enjoys full asymptotic power. In short, we cannot whiten the estimator without totally obscuring the signal.

[2] While the sufficiency property of Barber, Candès, et al. [9] also regulates the analyst's use of the fixed design matrix X, we will see in Section 6.1 that this restriction can be relaxed.

Equivalence of the two formulations

Next we will show that the two formulations of knockoffs in Sections 6.1 and 6.1, which we will respectively call the *standard method* and the *whitening method*, are essentially equivalent. In the standard method, an implementation of the knockoff filter is fully defined by a valid knockoff matrix $\widetilde{X}$ and a recipe for computing W-statistics satisfying the sufficiency and antisymmetry properties. In the whitening method, an implementation is fully defined by a diagonal matrix $\Delta \succeq (X^\mathsf{T}X)^{-1}$ and a recipe for computing a hypothesis ordering and ψ_j values from $|\widetilde{\beta}|$ and ξ. We will show that the implementations of each method are essentially in one-to-one correspondence with each other, using a coupling between $[X\ \widetilde{X}]^\mathsf{T}y$ and ω.

To begin our analysis, assume that $\widetilde{X}$ is a valid knockoff matrix with $X^\mathsf{T}\widetilde{X} = X^\mathsf{T}X - D$, which can be constructed for any $D \preceq 2X^\mathsf{T}X$ [9], and assume all $D_{jj} > 0$. Set $\Delta = 2D^{-1}$, so $A = X^\mathsf{T}X - \Delta^{-1} = X^\mathsf{T}X - \frac{1}{2}D$.

Following a conditioning argument in Barber, Candès, et al. [8], if we add and subtract $\widetilde{X}^\mathsf{T}y$ from $X^\mathsf{T}y$ we obtain a useful $2d$-variate Gaussian statistic whose mean and covariance matrix can be calculated from $y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$:

$$\begin{pmatrix} (X+\widetilde{X})^\mathsf{T}y \\ (X-\widetilde{X})^\mathsf{T}y \end{pmatrix} \;=\; \mathcal{N}_{2d}\left(\begin{pmatrix} 2A\beta \\ D\beta \end{pmatrix},\; \begin{pmatrix} 4\sigma^2 A & 0 \\ 0 & 2\sigma^2 D \end{pmatrix}\right). \tag{6.4}$$

It is suggestive that by rescaling the second component in (6.4) we obtain $D^{-1}(X-\widetilde{X})^\mathsf{T}y \sim \mathcal{N}_d(\beta, \sigma^2\Delta)$, which is the desired distribution for $\widetilde{\beta}$. Pursuing this ansatz, set

$$\omega \;=\; D^{-1}(X-\widetilde{X})^\mathsf{T}y - \hat{\beta} \;\sim\; \mathcal{N}_d\left(0,\, \sigma^2(\Delta - \Sigma)\right). \tag{6.5}$$

In the model where σ^2 is known, ω is ancillary and $\hat{\beta}$ complete sufficient, so ω is independent of $\hat{\beta}$ by Basu's Theorem. As a result we have $\widetilde{\beta} = D^{-1}(X-\widetilde{X})^\mathsf{T}y$ and

$$\xi \;=\; X^\mathsf{T}X\hat{\beta} - \Delta^{-1}\widetilde{\beta} \;=\; X^\mathsf{T}y - \frac{1}{2}(X-\widetilde{X})^\mathsf{T}y \;=\; \frac{1}{2}(X+\widetilde{X})^\mathsf{T}y.$$

Next we relate the whitening method's exploratory stage to a slightly weakened substitute for the sufficiency and antisymmetry properties.

Proposition 4. *For $j = 1, \ldots, d$ define*

$$W_j^*([X\ \widetilde{X}], y) = \mathit{sgn}\left((X_j - \widetilde{X}_j)^\mathsf{T}y\right) \cdot W_j([X\ \widetilde{X}], y). \tag{6.6}$$

If W satisfies the sufficiency and antisymmetry properties, then W^* depends on y only through the unordered pairs $\{X_1^\mathsf{T}y,\, \widetilde{X}_1^\mathsf{T}y\},\ \ldots,\ \{X_d^\mathsf{T}y,\, \widetilde{X}_d^\mathsf{T}y\}$. In terms of the coupling defined by (6.5)*, if W satisfies the sufficiency and antisymmetry properties then W^* depends on y only through ξ and $|\widetilde{\beta}|$.*

Proposition 4 is proven in Barber, Candès, et al. [8] but we provide a proof here for completeness.

Proof. By the sufficiency property, W only depends on y through $(X + \widetilde{X})^\mathsf{T} y$ and $(X - \widetilde{X})^\mathsf{T} y$, so the same is true of W^*. By the antisymmetry property we have

$$\begin{aligned} W_j^*(\text{Swap}_j([X\ \widetilde{X}]), y) &= \text{sgn}\left((\widetilde{X}_j - X_j)^\mathsf{T} y\right) \cdot W_j(\text{Swap}_j([X\ \widetilde{X}]), y) \\ &= \text{sgn}\left(-(X_j - \widetilde{X}_j)^\mathsf{T} y\right) \cdot \left(-W_j([X\ \widetilde{X}], y)\right) \\ &= W_j^*([X\ \widetilde{X}], y), \end{aligned}$$

so $W_k^*(\text{Swap}_j([X\ \widetilde{X}]), y) = W_k^*([X\ \widetilde{X}], y)$ for *all* k including $k = j$.

Since W^* is invariant to flipping the sign of any $(X_j - \widetilde{X}_j)'y$, it actually depends on y only through $(X + \widetilde{X})^\mathsf{T} y$ and $|(X - \widetilde{X})^\mathsf{T} y|$, i.e., only through the unordered pairs $\{X_j^\mathsf{T} y,\ \widetilde{X}_j^\mathsf{T} y\}$, for $j = 1, \ldots, d$.

For the coupling in (6.5), we have $2\xi = (X + \widetilde{X})^\mathsf{T} y$ and $D\widetilde{\beta} = (X - \widetilde{X})^\mathsf{T} y$. Hence, observing ξ and $|\widetilde{\beta}|$ is the same as observing $(X + \widetilde{X})^\mathsf{T} y$ and $|(X - \widetilde{X})^\mathsf{T} y|$, which is the same as observing the unordered pairs. □

We will say that W satisfies the *unordered pair property* if W^* depends on y only through $\{X_j^\mathsf{T} y, \widetilde{X}_j^\mathsf{T} y\}$ for $j = 1, \ldots, d$. This amounts to a slight relaxation of the sufficiency and antisymmetry properties because it allows for W^* to have unrestricted dependence on the fixed matrix $[X\ \widetilde{X}]$. With these relationships established, we are now prepared to prove formal equivalence.

Theorem 4. *Assume that $n \geq 2d$ and let $\widetilde{X}$ be a knockoff matrix with $X^\mathsf{T}\widetilde{X} = X^\mathsf{T}X - D$. Let $\Delta = 2D^{-1}$ and define ω as in* (6.5)*. Then*

(a) For any implementation of the whitening method, we can construct W-statistics satisfying the unordered pair property so that the two methods give identical rejection sets.

(b) For any W-statistics satisfying the unordered pair property with $|W_1|, \ldots, |W_d|$ almost surely positive with no ties, we can construct an implementation of the whitening method so that the two methods give identical rejection sets.

Proof. For (a), take $W_{[j]} = (d+1-j) \cdot \psi_{[j]} \cdot \text{sgn}(\widetilde{\beta}_{[j]})$, for $j = 1, \ldots, d$. W satisfies the unordered pair property because $W_{[j]}^* = (d + 1 - j) \cdot \psi_{[j]}$ depends on y only through ξ and $|\widetilde{\beta}|$, and the same is true for the ordering indices $[j]$.

For (b), because the method depends on $|W|$ only through the ordering of its coordinates, we can assume without loss of generality that $|W|$ is a permutation of $\{1, \ldots, d\}$. Take $[j]$ to be the index of the jth largest $|W_j|$ value, so that $|W_{[j]}| = d + 1 - j$, and set $\psi_j = \text{sgn}(W_j^*)$, which is a function of ξ and $|\widetilde{\beta}|$. As a result, we again have $W_{[j]} = (d + 1 - j) \cdot \psi_{[j]} \cdot \text{sgn}(\widetilde{\beta}_{[j]})$.

To see why these two methods return the same rejection sets when $W_{[j]} = (d+1-j) \cdot \psi_{[j]} \cdot \text{sgn}(\widetilde{\beta}_{[j]})$, note first that $\widetilde{p}_j = 1/2 \iff \text{sgn}(\widetilde{\beta}_j) = \psi_j \iff W_j > 0$. As a result

$$\begin{aligned}\widehat{\text{FDP}}_k^{\text{wh}} &= \frac{1 + \sum_{j=1}^d 1\{\widetilde{p}_{[j]} = 1,\, [j] \le k\}}{\sum_{j=1}^d 1\{\widetilde{p}_{[j]} = 1/2,\, [j] \le k\}} \\ &= \frac{1 + \sum_{j=1}^d 1\{W_{[j]} \le -(d+1-k)\}}{\sum_{j=1}^d 1\{W_{[j]} \ge d+1-k\}} \\ &= \widehat{\text{FDP}}_{d+1-k}^{\text{kn}}.\end{aligned}$$

For other values of $t \in (0, d+1]$, $\widehat{\text{FDP}}_t^{\text{kn}} = \widehat{\text{FDP}}_{\lceil t \rceil}^{\text{kn}}$, where $\lceil t \rceil$ is the integer ceiling of t. Therefore, $\hat{t} = d + 1 - \hat{k}$ and the rejection sets are the same.

□

Because the whitening method controls FDR, Theorem 4 implies that the unordered pair property is a sufficient condition for the standard knockoff filter to control FDR as well, relaxing the sufficiency and antisymmetry properties. The requirement in (b) that all $|W_j|$ be positive and distinct is not really necessary: we could break ties or generate "signs" at random, or generalize the whitening method so that $\widehat{\text{FDP}}_k^{\text{wh}}$ is only calculated for a subset of $k = \{1, \ldots, d\}$ corresponding to the indices where $|W_{[k]}|$ is positive and decreases. We ignore these generalizations for the sake of brevity.

Knockoffs without knockoffs: implementing the whitening method

While the coupling in Section 6.1 gives a recipe for generating ω using a knockoff matrix $\widetilde{X}$, we can alternatively skip creating $\widetilde{X}$ and instead generate the noise directly using the sample variance $\hat{\sigma}^2$. We can relax the usual dimension requirement, that $n \ge 2d$, and require only that $n \ge d + r$ where $r = \text{rank}(\Delta - \Sigma) \le d$. In that case let $M \in \mathcal{R}^{d \times r}$ be any fixed matrix with $MM^\mathsf{T} = \Delta - \Sigma$, and set

$$\omega = \sqrt{\frac{(n-d)v\hat{\sigma}^2}{\|z\|^2}} \cdot Mz, \quad \text{for} \quad z \sim \mathcal{N}_r(0, I_r) \quad \text{and} \quad v \sim \text{Beta}\left(\frac{r}{2}, \frac{n-d-r}{2}\right), \tag{6.7}$$

or $v = 1$ if $n = d + r$. z and v are auxiliary random variables generated independently of the data and each other. Lemma 4 shows ω has the desired distribution; furthermore it is independent of $\hat{\beta}$ because $\hat{\sigma}^2$ is.

Lemma 4. *Define z, v, and ω as in* (6.7)*. Then $(n-d)v\hat{\sigma}^2 \sim \sigma^2\chi_r^2$, and $\omega \sim \mathcal{N}_r(0, \sigma^2(\Delta - \Sigma))$.*

Proof. If $n = d+r$ and $v = 1$ then $(n-d)v\hat{\sigma}^2 \sim \sigma^2\chi_{n-d}^2$ is immediate. If $n > d+r$, independently generate $a_1 \sim \chi_r^2 = \text{Gamma}\left(\frac{r}{2}, 2\right)$ and $a_2 \sim \chi_{n-d-r}^2 = \text{Gamma}\left(\frac{n-d-r}{2}, 2\right)$. Then it is a standard

fact that the ratio $v = \frac{a_1}{a_1+a_2} \sim \text{Beta}\left(\frac{r}{2}, \frac{n-d-r}{2}\right)$ is independent of $a_1 + a_2 \sim \chi^2_{n-d}$. As a result,

$$v \cdot (n-d)\hat{\sigma}^2 \overset{\mathcal{D}}{=} v \cdot (a_1 + a_2)\sigma^2 = \sigma^2 a_1 \sim \sigma^2 \chi^2_r.$$

Similarly, $z/\|z\| \sim \text{Unif}(\mathbb{S}^{r-1})$ is independent of $\|z\|^2 \sim \chi^2_r$. As a result,

$$\sqrt{(n-d)v\hat{\sigma}^2} \cdot \frac{z}{\|z\|} \overset{\mathcal{D}}{=} \sqrt{\sigma^2\|z\|^2} \cdot \frac{z}{\|z\|} \sim \mathcal{N}_r(0, \sigma^2 I_r).$$

□

Because any implementation of knockoffs in its standard formulation can be viewed as an implementation of the whitening method by taking ω as in (6.5) and using the construction in Theorem 4(b), state-of-the-art knockoff implementations such as those in the `knockoff` package [99] can also be deployed as implementations of the whitening method.

In addition the whitening method may be a fruitful starting point for generalizing knockoffs, for example by altering what information is available in Stage 2, or by replacing Selective SeqStep with a different multiple testing method in Stage 3. Fithian, Sun, and Taylor [44]

One immediate generalization when $n > d + r$ is involves the "left-over" variance estimator

$$\widetilde{\sigma}^2 = \frac{(n-d)(1-v)}{n-d-r}\hat{\sigma}^2 \sim \frac{\sigma^2}{n-d-r}\chi^2_{n-d-r}.$$

It is easily shown that $\hat{\beta}, \omega$, and $\widetilde{\sigma}^2$ are mutually independent, and ξ and $\widetilde{\beta}$ depend only on $\hat{\beta}$ and ω, so allowing the analyst to use $\widetilde{\sigma}^2$ in Stage 2 has no effect on the conditional inference in Stage 3.

The analogous quantity in the standard implementation is $\widetilde{\sigma}^2 = \frac{1}{n-d-r}\|y - \hat{y}\|^2$, where $\hat{y}$ is the projection of y on the columns of $[X\ \widetilde{X}]$. Because $\widetilde{\sigma}^2$ is independent of $[X\ \widetilde{X}]^\mathsf{T} y$, the unordered pair property could be immediately relaxed to a requirement that W^* be a function of the unordered pairs and $\widetilde{\sigma}^2$ without affecting the FDR control proof. One use for $\widetilde{\sigma}^2$ could be as an input to generalized cross-validation [46].

Instead of pursuing methodological generalizations, we focus our attention for the remainder of this work on using the whitening method as a lens through which to gain a better theoretical understanding of knockoffs.

Related work

The whitening method presented in Section 6.1 is presaged in several prior works. Most notably, an argument in the Supplement of Barber, Candès, et al. [8] conditions on $(X + \widetilde{X})^\mathsf{T} y$ and $|(X - \widetilde{X})^\mathsf{T} y|$ to prove knockoffs control the *directional FDR*, a more stringent version of FDR where the analyst must draw a conclusion about $\text{sgn}(\beta_j)$ when rejecting H_j. An analogous argument to theirs shows that the whitening method controls the directional FDR as well: because $\widetilde{p}_j$ is also a valid conditional p-value for the data-dependent one-sided null $H_j^{\psi_j} : \psi_j \beta_j \leq 0$, the rejections in Stage

3 can be viewed as concluding that $\text{sgn}(\beta_j) = \psi_j$, yielding directional FDR control conditional on ξ and $|\widetilde{\beta}_j|$, and therefore also marginally.

Likewise, assuming $A = 2(X^\mathsf{T}X) - D$ is nonsingular Sarkar and Tang [108] consider $D^{-1}(X - \widetilde{X})^\mathsf{T}y$ and $A^{-1}(X + \widetilde{X})^\mathsf{T}y$ (equivalent to $\widetilde{\beta}$ and $A^{-1}\xi$ in our notation) as two independent unbiased estimators of β, taking this perspective as a starting point from which to derive hybrid multiple testing procedures blending the knockoff filter with the BH procedure.

There has been limited work on the power of knockoff filters. When studying the TPP-FDP tradeoff on the Lasso path, it is frequently assumed that there exists a constant fraction of non-zero coefficients and the coefficients are sampled from a fixed distribution [14, 121]. In addition, it is often assumed that $n/p \to \delta$ for some positive constant δ. Under this low-dimensional linear sparsity regime, [129] used results from [14] to derive the asymptotic TPP-FDP tradeoff of knockoff filters under i.i.d Gaussian design. [76] studied the power of knockoff under correlated design in the low dimensional setting, and showed that the knockoff filter has full asymptotic power when the precision matrix has vanishing diagonal entries. Going beyond the aforementioned linear sparsity assumptions, [42] studied the power of the *oracle* knockoff filter (assuming that the oracle covariance structure of the variables is known). They assumed that the coefficients are fixed and relatively large, and showed under certain regularity conditions that the oracle knockoff filter is consistent.

Recently, [115] showed that the SDP Knockoff can be asymptotically powerless for equicorrelated Gaussian design with correlation $\rho \geq 0.5$, due to the strong joint dependencies in the distribution of $[X, \tilde{X}]$. They proposed the minimum variance-based reconstructability (MVR) knockoffs, and showed that the TPR of the MVR knockoff converges to 1 under regularity conditions. However, it can be proved that the TPR of the Bonferroni correction also converges to 1 under the same set of conditions. [65] analyzed the phase diagram of the SDP knockoff, but their results are restricted to block-equicorrelated correlation structure with block size 2.

To the best of our knowledge, there has been no formal theoretical results comparing the asymptotic power of knockoff to that of other baseline FDR controlling techniques. When $p < n$, a crude baseline is to perform Bonferroni correction or the Benjamini-Yekutieli procedure on the p values from OLS. The main contribution of this paper is to identify failure mode of the knockoff filter relative to this baseline.

6.2 Finite sample upper bounds on the power of knockoff filter

Oracular ordering and the knockoff* procedure

We now turn our attention to the analyst's situation after Stage 1 of the fixed-X knockoff filter, to determine the best achievable knockoff method. Conditional on ξ and $|\widetilde{\beta}|$, the best possible ordering of variables is according to their conditional likelihood of resulting in a small p-value.

Assuming that $\psi_j = \text{sgn}(\beta_j)$ for all $j \in \mathcal{H}_0^c$, the log-odds of observing a small p-value is

$$\eta_j \;=\; \text{logit}\,\mathbb{P}(\widetilde{p}_j = 1/2 \mid C) \;=\; \frac{2|\widetilde{\beta}_j|}{\sigma^2 \Delta_{jj}} \cdot |\beta_j|. \tag{6.8}$$

As such, the best possible ordering of predictor variables — achievable only by an analyst with knowledge of β_j — is to order them in decreasing order of η_j, with ties broken arbitrarily. Define the knockoff* procedure as the procedure that correctly predicts all β_j, and orders the variables in decreasing order of η_j.

Next, for $0 < p < q < 1$, define the random walk

$$S_k^{p,q} = \sum_{j=1}^{k} p - Z_k, \quad \text{where } Z_1, Z_2, \ldots \overset{i.i.d.}{\sim} \text{Bern}\,(q)\,.$$

Let $F(\,\cdot\,; p, q)$ represent the distribution function of $\max\{k : \; S_k^{p,q} \geq 1 - p\}$, the last time the random walk exceeds $1 - q$. It can be easily shown by Chernoff's inequality that the expectation of $\max\{k : \; S_k^{p,q} \geq 1 - p\}$ is bounded.

Assume the η values are ordered as

$$\eta_{(1)} \geq \eta_{(2)} \geq \cdots \geq \eta_{(d)}.$$

Proposition 5. *If $\eta_{(1)} < -\log \alpha$, then the number of rejections for any knockoff procedure at FDR significance level α is stochastically smaller than $F\left(\,\cdot\,; \frac{\alpha}{1+\alpha}, \frac{e^{-\eta_{(1)}}}{1+e^{-\eta_{(1)}}}\right)$.*

In any asymptotic regime where $\eta_{(1)} \xrightarrow{p} 0$, the number of rejections for any knockoff procedure at any significance level is $O_p(1)$ with limiting distribution function stochastically smaller than $F\left(\,\cdot,\; \frac{\alpha}{1+\alpha}, \frac{1}{2}\right)$.

Proof. Let $p = \frac{\alpha}{1+\alpha}, q = \frac{e^{-\eta_{(1)}}}{1+e^{-\eta_{(1)}}}$, $\zeta_j = e^{-\eta_{(j)}}$ and $Z'_j = 1\{p_j = 1\} \overset{ind}{\sim} \text{Bern}\left(\frac{\zeta_j}{1+\zeta_j}\right)$. Then

$$A_k = \sum_{j=1}^{k} Z'_j, \qquad \text{and } R_k = k - A_k = \sum_{j=1}^{k} 1 - Z'_j.$$

The number of knockoff rejections is $R_{\hat{k}} \leq \hat{k}$, where $\hat{k}$ is the largest index k with

$$\widehat{\text{FDP}}_k = \frac{1 + A_k}{R_k} \leq \alpha \iff 1 \leq \alpha R_k - A_k = \sum_{j=1}^{k} \alpha - (1+\alpha) Z'_j.$$

If we define $Z_j \overset{i.i.d.}{\sim} \text{Bern}\left(\frac{\zeta_1}{1+\zeta_1}\right)$ and construct Z'_j so that $Z'_j \leq Z_j$ almost surely, then

$$\alpha R_k - A_k \leq (1+\alpha) S_k^{p,q} = \sum_j \left(\alpha - (1+\alpha) Z_j\right),$$

and $\hat{k}$ is almost surely no larger than the last time S_k exceeds $\frac{1}{1+\alpha} = 1 - p$. □

Proposition 5 gives us a tool for proving negative results that hold uniformly over all possible methods for calculating knockoff W-statistics, including methods that exploit informal prior information. It characterizes the regime where even an omniscient analyst cannot possibly order the hypotheses well enough to achieve nontrivial power, because there is simply not enough information available after conditioning on C. Naturally, less-than-omniscient analysts may still struggle due to the difficulty of ordering the predictor variables even when the knockoff* procedure has high power.

Proposition 5 assumes that all the log-odds are bounded away from infinity. As such, even for the variables ordered at the front of the list, the probability of observing a large p-value becomes nontrivial. The next proposition generalize Proposition 5 under the regime where there are no more than k^* "large" log-odds. We show that we can expect no more than $O(k^*)$ rejections from any knockoff method.

Proposition 6. *Fix a significance level $\alpha > 0$. If*

$$\eta_{(k^*)} < -\log(\alpha + \delta), \delta > 0$$

then the expected number of rejections for any knockoff procedure at FDR significance level α is upper bounded by $C_1(\delta)k^* + C_2(\delta)$, where

$$C_1(\delta) = \max\left\{1, \frac{4\alpha(1+\alpha+\delta)}{\delta}\right\} + 1, \quad \textit{and} \quad C_2(\delta) = O\left(\frac{1}{\delta^2}\right).$$

In particular, if

$$\eta_{(\lfloor cd \rfloor)} \xrightarrow{p} 0, \textit{ for all } c > 0,$$

then the expected number of rejections for any knockoff procedure at any FDR significance level is $o(d)$.

We defer the proof of this proposition to Section 6.4, where we also give a more precise bound on the constant $C_2(\delta)$. For example, when $\alpha = 0.1$ and $\delta \leq \alpha$, it can be shown that $C_2(\delta) \leq 5\delta^{-2}$. We note that the constant 5 is not optimal and better bounds could be obtained by simulation.

Whereas Proposition 6 gives us a way to bound the power of a knockoff procedure in terms of the information left over after the whitening step, the next section analyzes when it is impossible to carry out the whitening step without dramatic information loss. We will show that this loss is determined by the eigen-decomposition of $\Sigma = (X^\mathsf{T}X)^{-1}$, the covariance matrix of the OLS estimate, and characterize how large the signal must be to overcome this information loss and achieve nontrivial TPR, and illustrate our analysis with examples.

Upper bounding the power

The power of any knockoff procedure can be upper bounded by the knockoff* procedure introduced in the previous section. Therefore, Proposition 4 gives us a tool to prove negative results of any knockoff procedure, by calculating the odds ratio for hypothesis after the initial whitening

step. Specifically, we want to answer the following question: when will the initial whitening step destroy almost all the signals in z, such that the TPR of any knockoff procedure must converge to 0, even for effect sizes large enough that the TPR of the Bonferroni procedure converges to 1? Throughout this section, we assume that the covariance matrix (of $\hat{\beta}$) $\Sigma = (X^T X)^{-1}$ has unit diagonal entries.

Roughly speaking, we find that knockoffs fail when the leading eigenvalue λ_1 of Σ is much larger than $(\log d)^2$, and the leading eigenvector u_1 is dense, in a sense we will define later. Under these conditions, the diagonal matrix Δ can only dominate the covariance matrix Σ when the diagonal entries of Δ are much larger than $(\log d)^2$, so that the variance of the added artificial noise is more than $(\log d)^2$ times larger than the variance of the observed Gaussian vector. Recall that when the non-null means are equal to $\sigma\sqrt{2(1+r)\log d}$, with $r > 0$, the TPR of the Bonferroni procedure will converge to 1. However, under the same conditions, we will show that both the mean and standard deviation of the log odds η_j in equation (6.8) will converge to 0 at a faster rate than $1/\log d$, leaving essentially no information for the inference stage.

Most notably, when Σ has a factor model structure, the leading eigenvalue of Σ is typically on the order of d, and the leading eigenvector is spread more or less evenly. Our main result shows that the any knockoff procedure is powerless in this regime unless the signal is extremely strong (roughly on the order of $\sqrt{d}$).

We pause to note that this should not be confused with the case where the design matrix X in linear regression has a factor model structure. Since the test statistics (the OLS estimate) $\hat{\beta}$ has covariance matrix $(X^\mathsf{T} X)^{-1}$ rather than $X^\mathsf{T} X$, our results would not generally apply. To illustrate the difference between these two cases, consider the equicorrelated Gaussian design with correlation ρ, which we will use as a running example throughout this section. When the covariates are positively correlated ($\rho > 0$), $\frac{1}{n} X^\mathsf{T} X \approx \text{diag}(1-\rho) + \rho \mathbf{1}_d \mathbf{1}_d^\mathsf{T}$, so that $\lambda_1 \approx \rho d$ and $u_1 \approx \mathbf{1}_d/\sqrt{d}$, but the first eigenvalue of $\left(\frac{1}{n} X^\mathsf{T} X\right)^{-1}$ is not too large. This regime has been studied is several previous works, especially Spector and Janson [115], but is not the focus of our results. By contrast, if the covariates are negatively correlated ($\rho < 0$), then $(X^\mathsf{T} X)^{-1}$ may have a very large eigenvalue, possibly making knockoff inference powerless. A concrete example is the problem of testing for an effect of multiple treatments measured against a common control, which we describe next:

Example 1 (Multiple comparisons to control)**.** *Assume that we observe a continuous response on m units under each of several treatments, where $y_{j,i}$ represents the ith response under treatment $j = 0, 1, \ldots, d$, where $j = 0$ corresponds to the control condition. We can represent this as a linear model by writing $y_{0,i} = \beta_0 + \epsilon_{0,i}$ for the control group, and*

$$y_{j,i} = \beta_0 + \beta_j + \epsilon_{j,i}$$

for the jth treatment group, so that β_j corresponds to the differential effect of treatment j relative to control. If we compose a vector $y = (y_0^\mathsf{T}, y_1^\mathsf{T}, \cdots, y_d^\mathsf{T})^\mathsf{T}$, we obtain a standard linear regression problem with intercept β_0 and can perform multiple testing on the coefficients $\beta_1, \ldots, \beta_d$.

Under this model, the OLS estimator for β_j with $j > 0$ is of the form

$$\hat{\beta}_j = \bar{y}_j - \bar{y}_0, \quad \text{where } \bar{y}_k = \sum_{i=1}^{m} y_{k,i},$$

so that $\text{corr}(\hat{\beta}_j, \hat{\beta}_k) = 0.5$ *for distinct* $j, k > 0$.

With these examples in mind, let us go back to investigate how the leading eigenvalue and eigenvector of Σ affect how much information is destroyed in the whitening step. Recall that Δ must satisfy $\Delta \succeq \Sigma$, and let λ_1 be the leading eigenvalue of Σ, and u_1 be the corresponding eigenvector. Without loss of generality, suppose that $\Delta_{11} \leq \cdots \leq \Delta_{dd}$, and let $u^2_{1,(1)} \leq \cdots \leq u^2_{1,(d)}$ be the order statistics of $(u^2_{1,1}, \ldots, u^2_{1,d})$. For any subset $S \subset \{1, \ldots, d\}$, we have

$$\sum_{j \in S} \Delta_{jj} u^2_{1,j} \;=\; u^\mathsf{T}_{1,S} \Delta_{S,S} u_{1,S} \;\geq\; \sum_{\ell'=1}^{L} \lambda_{\ell'} (u^\mathsf{T}_{\ell',S} u_{1,S})^2 \;\geq\; \lambda_1 \|u_{1,S}\|_2^4.$$

Therefore

$$\|u_{1,S}\|_2^2 \max_{j \in S} \Delta_{jj} \;\geq\; \sum_{j \in S} \Delta_{jj} u^2_{1,j} \;\geq\; \lambda_\ell \|u_{1,S}\|_2^4,$$

and we have

$$\max_{j \in S} \Delta_{jj} \;\geq\; \lambda_1 \|u_{S,1}\|_2^2. \tag{6.9}$$

The above equation relates the amount of noise added in the whitening step with the eigen-decomposition of Σ. The following theorem combines the above argument with Proposition 4. It shows that we can determine an upper bound on the power of any knockoff procedure by inspecting the eigen-decomposition of Σ.

Theorem 5. *Let* $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$ *be the eigenvalues of* $\Sigma = (X^\mathsf{T} X)^{-1}$*, and* $u_1, \ldots, u_d$ *be the corresponding eigenvectors. Let* $\beta^2_{(1)} \geq \cdots \geq \beta^2_{(d)}$ *be the order statistics of* $(\beta^2_1, \ldots, \beta^2_d)$*. For* $1 \leq \ell \leq d$*, let* $u^2_{\ell,(1)} \leq \cdots \leq u^2_{\ell,(d)}$ *be the order statistics of* $(u^2_{\ell,1}, \ldots, u^2_{\ell,d})$*. For any targeted FDR level* $\alpha > 0$*, let* $k_{d,\alpha}$ *be the smallest integer* k *such that*

$$\max_{1 \leq \ell \leq d} \lambda_\ell \sum_{j=1}^{k} u^2_{\ell,(j)} > \frac{32 \log d}{(\log \alpha)^2} \frac{\beta^2_{(k)}}{\sigma^2} \tag{6.10}$$

Then there exists constant C_α *such that the expected number of rejections for any knockoff procedure at FDR level* α *is upper bounded by* $C_\alpha k_{d,\alpha}$*.*

Theorem 5 provides a finite-sample, deterministic upper bound on the number of rejections to be expected from any knockoff procedure, without regard to choices made by the analyst. Equation 6.10 provides a condition that practitioners can check to ascertain the smallest detectable SNR β_j/σ. Although the left hand side of Equation 6.10 may seem complicated at first sight, we point out that it is simply a generalization of the term in Equation 6.9. Indeed, we can easily generalize Equation 6.9 to show that

$$\max_{j \in S} \Delta_{jj} \;\geq\; \lambda_\ell \|u_{S,\ell}\|_2^2. \tag{6.11}$$

for any $1 \leq \ell \leq d$. Let $\ell^* = \arg\max_{1 \leq \ell \leq d} \lambda_\ell \sum_{j=1}^{k} u^2_{\ell,(j)}$. If we take $\ell = \ell^*$ and $S = \lfloor k \rfloor$, then the right hand side in the above equation will coincide with the left hand side of Equation

6.10. Therefore, Equation 6.10 and 6.9 together provide a lower bound on the amount of extra noise added in the whitening step. Recall that the eigenvectors $u_i, 1 \leq i \leq d$ have unit length. Therefore, when the terms in u_1 are roughly evenly spread, $\lambda_1 \sum_{j=1}^{k} u_{1,(j)}^2$ approximately scales as $k\lambda_1/d$. In that case, the above theorem implies that the expected number of rejections is on the order of $\frac{d \log d \max_j \beta_j^2}{\lambda_1}$.

As a running example, let us consider the case where Σ is an equi-correlated matrix with off-diagonal entry $\rho > 0$. Recall that in the Gaussian linear regression setup, this is the case where $X^\mathsf{T}X$ is an equi-correlated matrix with off-diagonal entry $\rho' < 0$. For the equi-correlated matrix Σ with $\Sigma_{jj} = 1$ and $\Sigma_{ij} = \rho, i \neq j$, its largest eigenvalue is $\lambda_1 = (d-1)\rho - 1$ and the corresponding eigenvector is $u_1 = \frac{1}{\sqrt{d}} 1^\mathsf{T}$. Thus, the term $\lambda_1 \sum_{j=1}^{k} u_{1,(j)}^2$ scales as $O(k)$. Therefore, the expected number of rejections is on the order of $\log d \max_j \mu_j^2$, regardless of the actual number of non-nulls.

The above theorem establishes the limit on the number of rejections any knockoff filter can have. However, it leaves open a couple questions which will be answered in the corollary below. First, the theorem does not provide a direct upper bound on the TPR of knockoff, and its comparison to the TPR of Bonferroni correction. To bound the TPR of knockoff, we need not only an upper bound on the number of rejections, but also a model on the distribution of the means μ_j. In this line of work, it is frequently assumed that $d\pi_d$ non-null positions are chosen uniformly at random from $\{1, \ldots, d\}$, with the non-null proportion $\pi_d > 0$. We adopt a similar condition in the following corollary. Second, this theorem shows that, in addition to a large leading eigenvalue, the power of knockoff will be negatively affected when u_1 does not have too many "small" entries. However, this property of u_1 is not made precise in the above theorem. When studying the power of FDR controlling methods, it is frequently assumed that the covariance matrix Σ has a limiting distribution [76, 129]. Here, to characterize the property of u_1, we will use a similar but less stringent condition. Let

$$\nu_d(|u_1|, c) = \frac{1}{d} \sum_{j=1}^{d} 1(\sqrt{d}|u_{1,j}| < c), \quad c > 0$$

be the empirical distribution function of the terms in the (scaled) leading eigenvector $\sqrt{d}u_1$. We will assume that $\nu_d(|u_1|, \cdot)$ has a limiting distribution $F(\cdot)$.

The following corollary shows that, under relatively weak conditions on μ and $F(\cdot)$, any knockoff procedure will have zero asymptotic TPR when the leading eigenvector λ_1 is much larger than $\log(d)^2$.

Corollary 2. *Suppose that*

1. *$\nu_d(|u_1|, \cdot)$ has a distribution limit $F(\cdot)$ which does not have a point mass at 0, i.e. $F(0) = 0$.*
2. *All but a uniformly drawn subset of $d\pi_d$ entries from the coefficient vector β are zero, where $\pi_d > 0$ and $d\pi_d \to \infty$. The non-zero means all equal to $\sqrt{2r \log d}$ with some $r > 1$.*

Then

1. *the TPR of Bonferroni correction always converges to 1;*

2. *the TPR of any knockoff procedure must converge to 0 if* $\lambda_1/(\log d)^2 \to \infty$*,*

where the expectation is taken over both the coefficient β *and the Gaussian error* ϵ*.*

To the best of our knowledge, this is the first result that shows knockoff can have zero asymptotic TPR when the TPR of knockoff converges to 1. Going back to the equi-correlated design example, we note that the empirical distribution of the entries in $\sqrt{d}u_1$ has a point mass at 1. As such, the first condition of the corollary is satisfied with $F(\cdot) = \delta_1$. In addition, recall that the leading eigenvalue $\lambda_1 = (d-1)\rho - 1$, where ρ is the off-diagonal entry. Therefore, we have $\lambda_1/(\log d)^2 \to \infty$ and it follows that the TPR of any knockoff procedure must converge to 0.

Compared to prior works on the negative results of knockoff, which are limited in both the covariance structure and the knockoff test statistics considered, our result here makes a number of unique and important contributions. First, prior works typically study a particular realization of the knockoff framework. For example, [65] and [76] study only lasso-type type statistics, and [115] only shows negative results for SDP-knockoff. By contrast, the result here applied to *any* test statistics and any strategy to create the knockoff matrix. Second, the covariance structure studied here is much more general than the equi-correlated block-diagonal matrices studied in [65] and [115]. Third, We note that although most prior works assume that there are a polynomial number of non-nulls (i.e. $\pi_d \sim d^{-\beta}$ for some $0 \leq \beta < 1$), our result does not require any assumptions on the sparsity parameter π_d, except that the expected number of true non-nulls $d\pi_d$ converges to infinity. Finally, in the OLS setup, results in [76] and [42] rely on conditions of the covariance matrix of the augmented matrix $(X, \tilde{X})$. In contrast, our conditions is stated in terms of the original covariance matrix, and can be checked without creating the knockoff matrix.

6.3 Numerical results

Fixed knockoff for Gaussian linear model

In this section, we use simulations studies under the Gaussian linear model setup to illustrate scenarios where the TPR of any knockoff method is close to zero. For the design matrix $\mathbf{X} \in \mathcal{R}^{n\times d}$, we generate random matrices whose rows are generated i.i.d from $N(0, K), K \in \mathcal{R}^{d\times d}$. We consider the following two regimes:

(a) **Positively equi-correlated OLS estimator** K^{-1} is an equicorrelation matrix with correlation $\rho = 0.2$, i.e. $K^{-1}_{ii} = 1, 1 \leq i \leq d$, and $K^{-1}_{ij} = 0.2, 1 \leq i < j \leq d$;

(b) **Positively equi-correlated covariates** K is an equicorrelation matrix with correlation $\rho = 0.2$.

In regime (a), the columns of $\mathbf{X}$ are negatively correlated and the OLS test statistics are positive correlated, and we expect that any knockoff methods will have trivial power. In regime (b), the columns of $\mathbf{X}$ are positively correlated and the OLS test statistics are negatively correlated.

We choose $n = 3000$ and $d = 1000$. For each K, we fix one realization of the random matrix, and then normalize its columns to obtain the design matrix $\mathbf{X}$. Next, we generate the response $Y \in \mathcal{R}^d$ as follows. First, to define β, we choose $s = 30$ coefficients uniformly at random and let $\beta_j = 5$ for each of the selected coefficients. We then generate $Y = \mathbf{X}\beta + \epsilon$, where ε_i are i.i.d standard normal errors. The above data generating procedure is repeated 600 times for each K.

For each design matrix $\mathbf{X}$, we generate the knockoff matrix $\widetilde{\mathbf{X}}$ using the Equicorrelated-knockoff and SDP-knockoff algorithms [9]. For each instance of $(\mathbf{X}, \widetilde{\mathbf{X}}, Y)$, we first consider the knockoff* procedure introduced in Section 6.2. The knockoff* procedure is the best achiveable knockoff method in terms of TPR, and can only be carried out with oracular knowledge about the true coefficients β. Next, we consider a practically feasible knockoff method which uses the maximum lasso penalty level as test statistic. Finally, we consider the BH procedure and the Bonferroni test on OLS p-values for baseline comparison. Note that the BH procedure is not guaranteed to control the FDR at the desired level when K is an equicorrelation matrix with $\rho = 0.2$.

Overall, the performance of knockoff shows stark contrast under these two regimes, while the performance of BH and Bonferroni appear to be much more stable. Figure 6.1 shows the power of the BH, the Bonferroni and the oracular knockoff tests. We see that when the OLS test statistics are positively correlated, the TPRs of both oracular knockoff tests (i.e. knockoff*) are close to zero. Table 6.1 shows the FDR and TPR of all methods for targeted FDR level $\alpha = 0.1$ and 0.2. We find that the TPR of the best knockoff method is smaller than 0.02 when controlling the FDR at 0.2.

Recall that when the covariance matrix of the test statistics has factor model structure, the whitening step of knockoff destroys virtually all the information, and the log-odds of observing small p-values in the inference stage becomes small. This is again confirmed by Figure 6.2, which shows the average of the $s = 30$ largest log-odds across different trials. The rest of the log-odds are zero since the corresponding coefficient β_j is zero. In particular, we find that most of the log-odds are smaller than $-\log \alpha$ with $\alpha = 0.1$. Thus by Proposition 6, the number of rejections of any knockoff methods must be small.

Knockoff for multivariate Gaussian statistics

In Section 6.1, we reinterpreted the knockoff method and generalized the fixed-X knockoff procedure to multivariate normal test statistics. Here we use simulations to investigate the performance of different methods for testing the means of multivariate normal, and hint at the possible use cases and limitations of knockoffs for such problems.

We generate $d = 1000$ dimensional multivariate Gaussian vectors $\hat{\mu} \sim N(\mu, K)$. The mean vector μ is generated in the same way as the linear coefficient β in the previous simulation, except that the non-zero means μ_j are set to 3.5. We consider two types of covariance matrices:

(a) K **has factor model structure**. In particular, we let

$$K = I_d + \lambda \sum_{\ell=1}^{k} u_\ell u_\ell^T, \tag{6.12}$$

where u_ℓ are drawn independently from the uniform sphere.

(b) K^{-1} **has factor model structure**, i.e.

$$K^{-1} = I_d + \lambda \sum_{\ell=1}^{k} u_\ell u_\ell^T, \tag{6.13}$$

In particular, we choose $k = 2$ and $\lambda \in \{20, 100\}$. We normalize the covariance matrix K such that K has unit diagonal entries. After normalization, the largest eigenvalue of K is approximately 19 and 75 for $\lambda = 20$ and 100, respectively (the sum of all eigenvalue is $d = 1000$). For each setup, we consider one fixed realization of the random matrix K. In each trial of our simulation, we first generate the mean vector μ and then the multivariate normal vector $\hat{\mu}$.

For each matrix K, we repeat the above data generating procedure $N = 600$ time. Again, we consider both the SDP knockoff and the equi-correlated knockoff to create the diagonal matrix D that satisfies $D \succeq K$. For each diagonal matrix D, we first implement the oracular ordering and the associated knockoff* procedure defined in Section 6.2. We consider For each observed $\hat{\mu} \sim N(\mu, K)$, we can generate an artificial design matrix $X \in \mathcal{R}^{2d \times d}$, where the first d rows of X equal to $K^{-1/2}$, and the last d rows are filled with zero. We can then generate an response vector $y \in \mathcal{R}^{2d}$ as $y = ((K^{-1/2}\hat{\mu})^{\mathsf{T}}, \epsilon_{d+1}, \ldots, \epsilon_{2d})^{\mathsf{T}}$, where $\epsilon_{d+1}, \ldots, \epsilon_{2d}$ are i.i.d standard Gaussian noise. As such, $y - X^{\mathsf{T}}\mu \sim N(0, I_{2d \times 2d})$, and we can use the fixed-X knockoff on the pair (X, y).

In addition, we also apply the Bonferroni-BH procedure proposed by Sarkar and Tang [108], with the diagonal matrix D created via SDP-knockoff.

Table 6.2 shows the FDR and TPR of different methods for $\alpha = 0.2$. As expected, we found that the power of even the best achieveable knockoff method is less than that of Bonferroni when the covariance matrix K has a factor model structure with reasonably large leading eigenvalue. This suggests that knockoff-type approaches suffer from severe power loss when applied to general test statistics with factor model structure. However, when the precision matrix K^{-1} has a factor model structure, it maybe possible to use the knockoff framework procedure to design a test with superior TPR than baseline methods such as the BH. We leave this for future research.

6.4 Proofs

Proof of Proposition 6

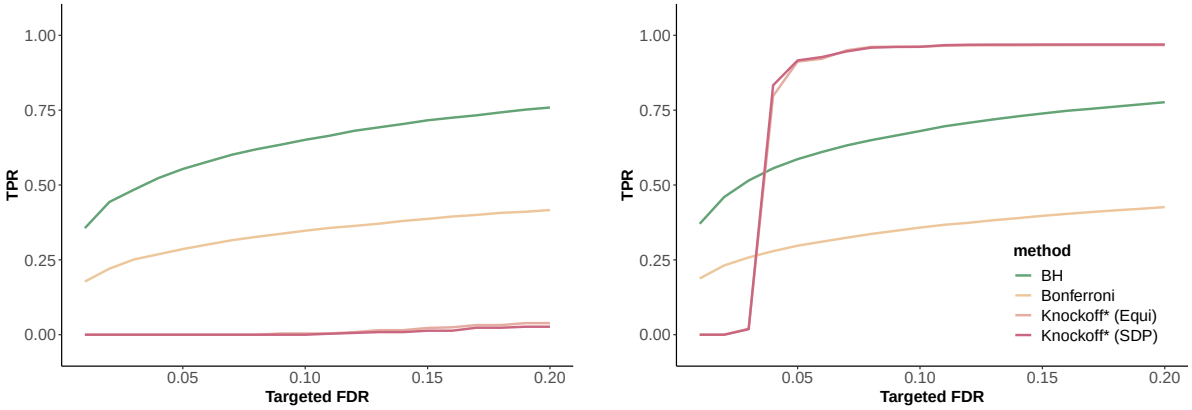
For any α and δ, we define

$$p = \frac{\alpha}{1+\alpha}, \quad q_\delta = \frac{\alpha+\delta}{1+\alpha+\delta}.$$

We will now prove the following proposition, which is stronger and more precise than Proposition 6.

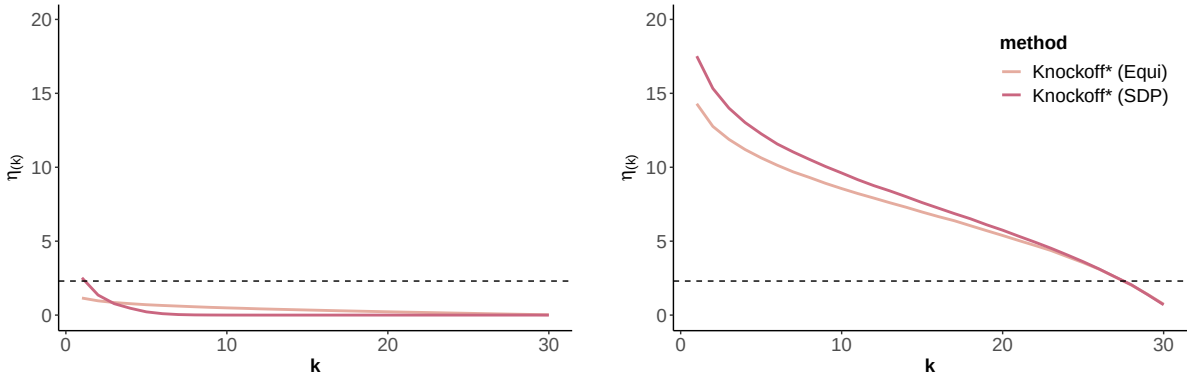
Proposition 7. *Suppose that*

$$\eta_{(k^*)} < -\log(\alpha+\delta), \delta > 0,$$



(a) Positively equi-correlated OLS estimator (b) Positively equi-correlated covariates

Figure 6.1: TPR of different tests under different target FDR levels. Given the knockoff matrix, he knockoff* test is the best achievable knockoff method. "Equi" stands for equi-correlation knockoffs, while "SDP" stands for the SDP knockoffs. In Figure (a), the covariates in the design matrix are positively correlated with correlation approximately 0.2. In Figure (b), the covariates in the design matrix are negatively correlated, and the entries of the OLS estimate $\hat{\beta}$ are positively correlated with correlation approximately 0.2.



(a) Positively equi-correlated OLS estimator (b) Positively equi-correlated covariates

Figure 6.2: The log-odds of observing a small p-value in the inference stage. The log-odds are computed as Equation 6.8. Only the $s = 30$ non-zero log odds are shown in both figures. $\eta_{(k)}$ denotes the kth largest log-odd and $\eta_{(k)} = -\log 0.1$ is shown in the dashed black line.

		SDP knockoff		Equicorrelated knockoff		Other methods	
		Knockoff*	Maximum penalty level	Knockoff*	Maximum penalty level	BH	Bonferroni
$\alpha = 0.1$	FDR	0.00	0.00	0.00	0.00	0.09	0.01
	TPR	0.00	0.00	0.00	0.00	0.65	0.35
$\alpha = 0.2$	FDR	0.00	0.01	0.00	0.00	0.18	0.02
	TPR	0.03	0.00	0.04	0.01	0.76	0.42

(a) Positively equi-correlated OLS estimator

		SDP knockoff		Equicorrelated knockoff		Other methods	
		Knockoff*	Maximum penalty level	Knockoff*	Maximum penalty level	BH	Bonferroni
$\alpha = 0.1$	FDR	0.06	0.07	0.05	0.07	0.10	0.01
	TPR	0.96	0.41	0.96	0.41	0.68	0.36
$\alpha = 0.2$	FDR	0.16	0.17	0.15	0.17	0.19	0.01
	TPR	0.97	0.73	0.97	0.74	0.78	0.43

(b) Positively equi-correlated covariates

Table 6.1: FDR and TPR of different methods under different target FDR levels.

		SDP knockoff		Equicorrelated knockoff		Other methods		
		Knockoff*	Maximum penalty level	Knockoff*	Maximum penalty level	Sarkar-Tang	BH	Bonferroni
$\lambda = 20$	FDR	0.00	0.04	0.02	0.03	0.19	0.19	0.01
	TPR	0.52	0.17	0.58	0.21	0.03	0.77	0.42
$\lambda = 100$	FDR	0.00	0.02	0.00	0.00	0.00	0.18	0.01
	TPR	0.21	0.05	0.13	0.04	0.01	0.77	0.41

(a) When K has factor model structure

		SDP knockoff		Equicorrelated knockoff		Other methods		
		Knockoff*	Maximum penalty level	Knockoff*	Maximum penalty level	Sarkar-Tang	BH	Bonferroni
$\lambda = 20$	FDR	0.18	0.17	0.18	0.19	0.20	0.19	0.02
	TPR	1.00	0.67	0.99	0.68	0.40	0.76	0.41
$\lambda = 100$	FDR	0.18	0.12	0.17	0.17	0.19	0.19	0.02
	TPR	0.99	0.39	0.99	0.47	0.49	0.77	0.41

(b) When K^{-1} has factor model structure

Table 6.2: FDR and TPR of different methods for testing means of multivariate Gaussian with correlation matrix K defined in Equations 6.12 and 6.13. Target FDR level $\alpha = 0.2$. "Maximum Penalty Level" refers to the method where we first generate artificial design matrix and response (X, y), and then apply the fixed-X knockoff.

where $k^* \geq 1/\alpha$. Define

$$C_1(\alpha, \delta) = \max\left\{1, \frac{4\alpha(1+\alpha+\delta)}{\delta}\right\} + 1,$$

and

$$C_2(\alpha, \delta) = 5.142 C(p, q_\delta)^{-3/2} \frac{(1+\alpha)^{9/2}(\alpha+\delta)}{\alpha^{3/2}} \left(\frac{1+\alpha+\delta}{\delta}\right)^2,$$

where

$$C(p, q_\delta) = 1 + \max\left\{0, \min\left\{\frac{5}{3p} - \frac{2q_\delta}{3p^2}, \frac{1}{q_\delta}\right\}\right\} \geq 1.$$

Then the expected number of rejections for any knockoff procedure at FDR significance level α is upper bounded $C_1(\alpha, \delta)k^* + C_2(\alpha, \delta)$.

Proof. Consider the random walk

$$S_k = \sum_{j=1}^{k}(p - Z_j), \quad \text{where } Z_j \overset{ind}{\sim} \text{Bern}\left(q_j\right), \quad p = \frac{\alpha}{1+\alpha}, \quad q_j = \frac{e^{-\eta_{(j)}}}{1+e^{-\eta_{(j)}}}.$$

Then for any knockoff procedure, the number of rejections R is upper bounded by

$$\max\left\{k : S_k \geq \frac{1}{1+\alpha}\right\} \leq \max\left\{k : S_k \geq 0\right\}.$$

Consider another random walk

$$\tilde{S}_k = \sum_{j=1}^{k}(p - \tilde{Z}_j), \quad \text{where } \tilde{Z}_j \overset{i.i.d.}{\sim} \text{Bern}\left(q_\delta\right), \quad q_\delta = \frac{\alpha+\delta}{1+\alpha+\delta} \leq q_{(k^*)}$$

Note that

$$S_k \;=\; \sum_{j=1}^{k^*}(p - Z_j) + \sum_{j=k^*}^{d}(p - Z_j) \;\leq\; pk^* + \sum_{j=k^*}^{d}(p - Z_j).$$

Since $q_1 \leq q_2 \leq ...$, we know that $\sum_{j=k^*}^{d}(p - Z_j)$ is stochastically smaller than $\sum_{j=k^*}^{d}(p - \tilde{Z}_j)$, where $\tilde{Z}_j \overset{i.i.d.}{\sim} \text{Bern}\left(q_{(k^*)}\right)$. Therefore S_k is stochastically smaller than $pk^* + \tilde{S}_{k-k^*}$. Therefore

$$\begin{aligned}
\mathbb{E}[R] \;&\leq\; \mathbb{E}[\max\left\{k : S_k \geq 0\right\}] \\
&\leq\; \mathbb{E}\left[\max\left\{k : \tilde{S}_{k-k^*} \geq -pk^*\right\}\right] \\
&=\; k^* + \mathbb{E}\left[\max\left\{k : \tilde{S}_k \;\geq\; -pk^*\right\}\right].
\end{aligned} \tag{6.14}$$

Define

$$p(r) \triangleq \mathbb{P}\left(\max\left\{k : \tilde{S}_k \geq -pk^*\right\} \;=\; r\right).$$

Let $C_m = \max\{4\left(\frac{q_\delta}{p} - 1\right)^{-1}, 1\}$, and then

$$\begin{aligned}
\mathbb{E}\left[\max\left\{k : \tilde{S}_k \geq -pk^*\right\}\right] &= \sum_{r=1}^{\infty} r\mathbb{P}\left(\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right) \\
&\leq C_m k^* + \sum_{r=C_m k^*+1}^{\infty} r\mathbb{P}\left(\max\left\{k : \tilde{S}_k - pk^*\right\} = r\right) \qquad (6.15) \\
&= C_m k^* + \sum_{r=C_m k^*+1}^{\infty} rp(r).
\end{aligned}$$

Therefore, combining Equations 6.15 and 6.14, we have

$$\begin{aligned}
\mathbb{E}[R] &\leq k^* + \mathbb{E}\left[\max\left\{k : \tilde{S}_k \geq -pk^*\right\}\right] \\
&\leq k^* + C_m k^* + \sum_{r=C_m k^*+1}^{\infty} rp(r) \\
&= (C_m + 1)k^* + \sum_{r=C_m k^*+1}^{\infty} rp(r).
\end{aligned}$$

Recalling the definitions of C_m and $q_{(k^*)}$, we have

$$4\left(\frac{q_\delta}{p} - 1\right)^{-1} = 4\left(\frac{(\alpha+\delta)(1+\alpha)}{1+\alpha+\delta} - 1\right)^{-1} = \frac{4\alpha(1+\alpha+\delta)}{\delta}.$$

Therefore we have $C_m + 1 \leq C_1(\alpha, \delta)$. Turning to the second term, by Lemma 7 we have

$$\sum_{r=C_m k^*+1}^{\infty} rp(r) \leq \frac{eq_\delta}{p(1-q_\delta)\pi\sqrt{p(1-p)}}(p - q_\delta) \sum_{r=C_m k^*+1}^{\infty} \sqrt{r}e^{-c_h r}. \qquad (6.16)$$

Note that we can bound the summation $\sum_{r=C_m k^*+1}^{\infty} \sqrt{r}e^{-c_h r}$ by (note that $C_m k^* + 1 \geq 1/\alpha + 1$)

$$\begin{aligned}
\sum_{r=1/\alpha+1}^{\infty} \sqrt{r}e^{-c_h r} &\leq \sqrt{1+\alpha} \sum_{r=1/\alpha+1}^{\infty} \sqrt{r-1}e^{-c_h r} \\
&\leq \sqrt{1+\alpha} \int_{1/\alpha}^{\infty} \sqrt{r}e^{-c_h r} dr \qquad (6.17) \\
&\leq \sqrt{1+\alpha} \int_{0}^{\infty} \sqrt{r}e^{-c_h r} dr,
\end{aligned}$$

where

$$\begin{aligned}\int_0^\infty \sqrt{r}e^{-c_h r}dr &= \int_0^\infty 2(2c_h)^{-3/2}y^2 e^{-\frac{y^2}{2}}dy \\ &= (2c_h)^{-3/2}\int_{-\infty}^\infty y^2 e^{-\frac{y^2}{2}}dy \\ &= \sqrt{2\pi}(2c_h)^{-3/2}\end{aligned}$$

The first equality above is obtained by change of variable $r = y^2/2c_h$. Therefore,

$$\sum_{r=C_m k^*+1}^{\infty} \sqrt{r}e^{-c_h r} \leq \sqrt{1+\alpha}\sqrt{2\pi}(2c_h)^{-3/2}$$

Note that

$$c_h = \frac{1}{2}t_*^2 C(p, q_\delta) = \frac{9(p-q_\delta)^2}{32}C(p, q_\delta).$$

Thus,

$$\begin{aligned}\sum_{r=C_m k^*+1}^{\infty} rp(r) &\leq \frac{e}{2\sqrt{\pi}}\left(\frac{32}{9}\right)^{3/2}\frac{\sqrt{1+\alpha}}{(p-q_\delta)^2}\frac{q_\delta}{p(1-q_\delta)\sqrt{p(1-p)}} \\ &\leq \frac{5.142C(p,q_\delta)^{-3/2}}{(p-q_\delta)^2}\frac{q_\delta\sqrt{1+\alpha}}{p(1-q_\delta)\sqrt{p(1-p)}}.\end{aligned}$$

Turning to the term $(p-q_\delta)^2$ in the denominator, we have

$$(p-q_\delta)^2 = \left(\frac{\alpha}{1+\alpha} - \frac{\alpha+\delta}{1+\alpha+\delta}\right)^2 = \left(\frac{\delta}{(\alpha+1)(1+\alpha+\delta)}\right)^2.$$

Therefore

$$\frac{1}{(p-q_\delta)^2}\frac{q_\delta}{p(1-q_\delta)\sqrt{p(1-p)}} = \frac{(1+\alpha)^4(\alpha+\delta)}{\alpha^{3/2}}\left(\frac{1+\alpha+\delta}{\delta}\right)^2.$$

Combining the three equations above, we have

$$\sum_{r=C_m k^*+1}^{d} re^{-c_h r} \leq C_2(\alpha, \delta),$$

and the proposition is proved. □

Proof of Theorem 5 and Corollary 2

Theorem 5. *Let $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$ be the eigenvalues of $\Sigma = (X^\mathsf{T}X)^{-1}$, and $u_1, \ldots, u_d$ be the corresponding eigenvectors. Let $\beta^2_{(1)} \geq \cdots \geq \beta^2_{(d)}$ be the order statistics of $(\beta^2_1, \ldots, \beta^2_d)$. For $1 \leq \ell \leq d$, let $u^2_{\ell,(1)} \leq \cdots \leq u^2_{\ell,(d)}$ be the order statistics of $(u^2_{\ell,1}, \ldots, u^2_{\ell,d})$. For any targeted FDR level $\alpha > 0$, let $k_{d,\alpha}$ be the smallest integer k such that*

$$\max_{1\leq\ell\leq d} \lambda_\ell \sum_{j=1}^{k} u^2_{\ell,(j)} > \frac{32\log d}{(\log\alpha)^2}\frac{\beta^2_{(k)}}{\sigma^2} \tag{6.10}$$

Then there exists constant C_α such that the expected number of rejections for any knockoff procedure at FDR level α is upper bounded by $C_\alpha k_{d,\alpha}$.

Proof. By Proposition 6, it suffices to show that $\eta_{(k_{d,\alpha})} < -C\log\alpha$ for constant $C < 1$. Note that by Equation 6.8, this is equivalent to showing that

$$2\max_{j\geq k_{d,\alpha}} |\widetilde{\beta}_j||\beta_j|\Delta^{-1}_{jj} < -C\sigma^2\log\alpha.$$

Recall that Δ must satisfy $\Delta \succeq (X^\mathsf{T}X)^{-1}$. We have shown in Equation 6.9 that

$$\max_{j\in S} \Delta_{jj} \;\geq\; \lambda_\ell \|u_{S,\ell}\|_2^2.$$

Let $k_{d,\alpha}$ be the integer defined in the statement of the theorem. Suppose that

$$\ell^* = \arg\max_{1\leq\ell\leq d} \lambda_\ell \sum_{j=1}^{k} u^2_{\ell,(j)}.$$

Without loss of generality, suppose that $\Delta_{11} \leq \cdots \leq \Delta_{dd}$. If we take $S = [k_{d,\alpha}]$, then we have

$$\Delta_{k_{d,\alpha},k_{d,\alpha}} \;\geq\; \lambda_{\ell^*}\|u_{\ell^*,S}\|_2^2 \;\geq\; \lambda_{\ell^*}\sum_{j=1}^{k_{d,\alpha}} u^2_{\ell^*,(j)} \;\geq\; \frac{32\log d}{(\log\alpha)^2}\frac{\beta^2_{(k_{d,\alpha})}}{\sigma^2}.$$

Therefore, for any $j \geq k_{d,\alpha}$, if $\beta^2_j \leq \beta^2_{(k_{d,\alpha})}$, then

$$2\Delta^{-1}_{jj}\beta^2_j \leq \frac{(\log\alpha)^2\sigma^2}{32\log d}.$$

Since $\beta^2_{(k_{d,\alpha})}$ is the $k_{d,\alpha}$th largest element in $(\beta^2_1, \ldots, \beta^2_d)$, we know that

$$|\{j \geq k_{d,\alpha} : \beta^2_j > \beta^2_{(k_{d,\alpha})}\}| \leq k_{d,\alpha}$$

Therefore, if we denote $\mathcal{I}_\alpha$ as

$$\mathcal{I}_\alpha = \left\{ j : 2\Delta^{-1}_{jj}\mu^2_j \leq \frac{(\log\alpha)^2\sigma^2}{16\log d} \right\},$$

i.e. the collection of indices j such that $2\Delta_{jj}^{-1}\beta_j^2 < \frac{(\log\alpha)^2\sigma^2}{16\log d}$, then we must have

$$|\mathcal{I}_\alpha^c| \le k_{d,\alpha} + |\{j \ge k_{d,\alpha} : \beta_j^2 > \beta_{(k_{d,\alpha})}^2\}| \le 2k_{d,\alpha}.$$

Note that $2\widetilde{\beta}_j\beta_j\Delta_{jj}^{-1} \sim \mathcal{N}(2\Delta_{jj}^{-1}\beta_j^2, 4\Delta_{jj}^{-1}\beta_j^2\sigma^2)$, where $\max_{j\in\mathcal{I}_\alpha} 2\Delta_{jj}^{-1}\beta_j^2 \le \frac{(\log\alpha)^2\sigma^2}{16\log d}$, so that $2|\widetilde{\beta}_j\beta_j\Delta_{jj}^{-1}|$ is stochastically smaller than

$$\frac{(\log\alpha)^2\sigma^2}{16\log d} + \sqrt{\frac{(\log\alpha)^2\sigma^4}{8\log d}}|\mathcal{N}(0,1)|.$$

As a result,

$$\max_{j\in\mathcal{I}_\alpha} 2|\widetilde{\beta}_j||\beta_j|\Delta_{jj}^{-1} < \frac{(\log\alpha)^2\sigma^2}{16\log d} + \sqrt{\frac{(\log\alpha)^2\sigma^4}{8\log d}}\sqrt{4\log d} < \frac{-\sigma^2\log\alpha}{8} + \frac{-\sigma^2\log\alpha}{\sqrt{2}} \quad (6.18)$$

for $d > \alpha^{-1}$ with probability $1 - O(d^{-1})$. Note that we have proved that $|\mathcal{I}_\alpha^c| \le 2k_{d,\alpha}$. As such,

$$\eta_{(2k_{d,\alpha})} \le 2\max_{j\in\mathcal{I}_\alpha}\frac{|\widetilde{\beta}_j|\Delta_{jj}^{-1}}{\sigma^2}\cdot|\mu_j| < -\left(\frac{1}{\sqrt{2}}+\frac{1}{8}\right)\log\alpha < -0.9\log\alpha$$

with probability $1 - O(d^{-1})$. Let $\delta(\alpha) = \alpha^{0.9} - \alpha$. Then we have $\eta_{(2k_{d,\alpha})} < -\log(\alpha+\delta) = -0.9\log\alpha$. It follows from proposition 7 that the expected number of rejections is upper bounded by $2C_1(\alpha,\delta(\alpha))k_{d,\alpha} + C_2(\alpha,\delta(\alpha)) \le C_\alpha^* k_{d,\alpha}$, where $C_\alpha^* = 2C_1(\alpha,\delta(\alpha)) + C_2(\alpha,\delta(\alpha))$ is a constant that depends only on α. Therefore, the expected number of rejections

$$\begin{aligned}\mathbb{E}[R] &= \mathbb{E}\left[R|\eta_{(2k_{d,\alpha})} < -0.9\log\alpha\right]\mathbb{P}\left(\eta_{(2k_{d,\alpha})} < -0.9\log\alpha\right)\\ &\quad + \mathbb{E}\left[R|\eta_{(2k_{d,\alpha})} > -0.9\log\alpha\right]\mathbb{P}\left(\eta_{(2k_{d,\alpha})} > -0.9\log\alpha\right)\\ &\le 2C_\alpha^* k_{d,\alpha} + dO(d^{-1})\\ &\le C_\alpha k_{d,\alpha},\end{aligned}$$

where C_α is a constant that depends only on α.

□

Corollary 2. *Suppose that*

1. *$\nu_d(|u_1|,\cdot)$ has a distribution limit $F(\cdot)$ which does not have a point mass at 0, i.e. $F(0) = 0$.*
2. *All but a uniformly drawn subset of $d\pi_d$ entries from the coefficient vector β are zero, where $\pi_d > 0$ and $d\pi_d \to \infty$. The non-zero means all equal to $\sqrt{2r\log d}$ with some $r > 1$.*

Then

1. *the TPR of Bonferroni correction always converges to 1;*

2. the TPR of any knockoff procedure must converge to 0 if $\lambda_1/(\log d)^2 \to \infty$*,*

where the expectation is taken over both the coefficient β *and the Gaussian error* ϵ*.*

Proof. The first part of Corollary follows directly from the fact that the p-value for each hypothesis is smaller than $1/d$ when the SNR is larger than $\sqrt{2\log d}$. To prove the second part of the Corollary, it suffices to show that the expected number of rejections $\mathbb{E}[R] = o(d\pi_d)$, where the expectation is taken over both μ and z. By Proposition 6, it suffices to show that all but a vanishing proportion of the log-odds at the $d\pi_d$ non-null positions must converge in probability to zero, i.e.

$$\eta_{(2c\pi_d d)} \xrightarrow{p} 0 \tag{6.19}$$

for any $c > 0$. The rest of the proof proceeds in two steps. First, using Theorem 5, we will establish an intermediate result: all but cd diagonal entries of D^{-1} must be prohibitively large. As such, all but cd log-odds must converge to zero. Then, using the fact that the $d\pi_d$ non-null positions are chosen at random, we will show that among these cd positions where D^{-1} could be small, only $cd\pi_d$ are true non-nulls. As such, there could only be $O(cd\pi_d)$ rejections on average. Since c is arbitrary, the desired result will follow.

To prove the aforementioned intermediate result, we will use the first condition that the limiting distribution of the entries in u_1 does not have a point mass. By assumption, there exists constant $\epsilon > 0$ such that $F(\epsilon) < c$. for large enough d, we have

$$\nu_d(|u_1|, \epsilon) \le c/2.$$

Therefore for large enough d,

$$\sqrt{d}|u_{1,(cd/2)}| \ge \epsilon.$$

Therefore,

$$\sum_{j=1}^{cd} u_{1,(j)}^2 \;\ge\; \frac{cd}{2}(u_{1,(cd/2)})^2 \;\ge\; \frac{cd}{2}\frac{\epsilon}{d} \;\ge\; \frac{c\epsilon}{2}.$$

Hence, for any $\alpha < 1$, Equation 6.10 is satisfied with $k_{d,\alpha} = cd$ when d is large enough. Therefore, by Equation 6.18,

$$2\max_{j \ge cd} \frac{|\widetilde{\beta}_j|\Delta_{jj}^{-1}}{\sigma^2} \cdot |\beta_j| \overset{p}{<} -0.9\log\alpha.$$

for any $\alpha < 1$. Take $\alpha \to 1$, we know that

$$\max_{j \ge cd} \frac{|\widetilde{\beta}_j|\Delta_{jj}^{-1}}{\sigma^2} \cdot |\beta_j| \xrightarrow{p} 0.$$

This finishes the first step of the proof.

Turning to the second step, note that we always have

$$\frac{|\widetilde{\beta}_j|\Delta_{jj}^{-1}}{\sigma^2} \cdot |\beta_j| = 0$$

when $\beta_j = 0$. Therefore it follows that for the set $\mathcal{A}_j = \{j : j \geq cd\} \cup \{j : j \leq cd, \beta_j = 0\}$, we also have

$$\max_{j \in \mathcal{A}_j} \frac{|\widetilde{\beta}_j| \Delta_{jj}^{-1}}{\sigma^2} \cdot |\mathrm{m}\beta_j| \xrightarrow{p} 0.$$

That is, all the log-odds at the positions in $\mathcal{A}_j$ must converge to 0. Therefore, in order to show the desired inequality 6.19, it suffices to show that there are (with probability going to 1) no more than $2cd\pi_d$ indices outside the set $\mathcal{A}_j$. Since the non-null positions are chosen at random, we apply Hoeffding's bound (for sampling without replacement) again and obtain

$$\mathbb{P}\left(\frac{|\mathcal{A}_j^c|}{cd} > 2\pi_d\right) \;=\; \mathbb{P}\left(\frac{\#\{j \leq cd : \beta_j \neq 0\}}{cd} > 2\pi_d\right) \;\to\; 0.$$

This proves the Corollary.

□

Proof of technical lemmas

Lemma 5. *For any $q, \delta \in (0, 1)$, we have*

$$\frac{q \log(1-\delta)}{\log(1-q\delta)} - 1 \geq \frac{(1-q)}{2}\delta.$$

Proof. For any $q \in (0, 1)$, let

$$f(\delta) = \frac{(1-q)}{2}\delta \log(1-q\delta) + \log(1-q\delta) - q\log(1-\delta).$$

It suffices to show that $f(\delta) \geq 0$ for $\delta \in (0, 1)$. First, note that $f(0) = 0$. We will now show that the derivative of f w.r.t to δ is always positive when $\delta > 0$. In fact, we have

$$\begin{aligned} f'(\delta) &= \frac{(1-q)}{2}\left(\log(1-q\delta) - \frac{q\delta}{1-q\delta}\right) - \frac{q}{1-q\delta} + \frac{q}{1-\delta} \\ &= \frac{(1-q)}{2}\left(\log(1-q\delta) + \frac{q\delta(1+\delta)}{(1-\delta)(1-q\delta)}\right). \end{aligned}$$

Note that

$$\log(1-q\delta) = -\log\frac{1}{(1-q\delta)} \geq -\left(\frac{1}{(1-q\delta)} - 1\right) = \frac{-q\delta}{1-q\delta}.$$

Therefore

$$\begin{aligned} f'(\delta) &= \frac{(1-q)}{2}\left(\log(1-q\delta) + \frac{q\delta(1+\delta)}{(1-\delta)(1-q\delta)}\right) \\ &\geq \frac{(1-q)}{2}\left(\frac{-q\delta}{1-q\delta} + \frac{q\delta(1+\delta)}{(1-\delta)(1-q\delta)}\right) \\ &= \frac{q(1-q)\delta^2}{1-q\delta} \geq 0, \end{aligned}$$

and the lemma is proved. □

Lemma 6. *Consider a random walk $S_t = \sum_{i=1}^{t} \zeta_t$ where $S_0 = 0$ and ζ_t are i.i.d Bernoulli variables with $\mathbb{P}(\zeta_t = 1) = 1 - q$ and $\mathbb{P}(\zeta_t = -1/\alpha) = q$. Let $p = \alpha/(1+\alpha)$ and suppose that $q > p$. Then*

$$\mathbb{P}(\max_{t \geq 1} S_t < 0) \leq \frac{2q(p-q)}{p(1-q)}.$$

Proof. We begin by identifying the martingale associated with the moment generating function of S_t. Let $\psi_0 > 0$ be a positive value that satisfies

$$\mathbb{E}e^{\psi_0 \zeta_1} = qe^{\psi_0} + (1-q)e^{-\psi_0 \frac{1}{\alpha}} = 1. \tag{6.20}$$

We will prove later that such ψ_0 exists and is unique. It follows that $e^{\psi_0 S_t}$ is a martingale, since

$$\mathbb{E}[e^{\psi_0 S_{t+1}} | S_t] = e^{\psi_0 S_t} \mathbb{E}e^{\psi_0 \zeta_(t+1)} = e^{\psi_0 S_t}.$$

For any value $M > 0$, let τ be the first time when the random walk leaves $(-M, 0]$, i.e.

$$\tau = \min_{t \geq 1} : S_\tau > 0 \text{ or } S_\tau \leq -M.$$

Then τ is a stopping time, and $S_{t \wedge \tau}$ is bounded for all $t \geq 1$. Therefore, by the optional stopping theorem,

$$1 = \mathbb{E}e^{\psi_0 S_\tau} = p_0(M)\mathbb{E}[e^{\psi_0 S_\tau} | S_\tau > 0] + (1 - p_0(M))\mathbb{E}[e^{\psi_0 S_\tau} | S_\tau \leq -M],$$

where $p_0(M)$ is the probability that S_t reaches $(0, \infty)$ before it reaches $(-\infty, -M)$. Since S_t can increase no more than 1 at a time, we have $S_\tau \leq 1$. Therefore

$$\mathbb{E}[e^{\psi_0 S_\tau} | S_\tau > 0] \leq e^{\psi_0}.$$

On the other hand, we have

$$\mathbb{E}[e^{\psi_0 S_\tau} | S_\tau \leq -M] \leq e^{-\psi_0 M}.$$

Therefore

$$\begin{aligned} 1 &= p_0(M)\mathbb{E}[e^{\psi_0 S_\tau} | S_\tau > 0] + (1 - p_0(M))\mathbb{E}[e^{\psi_0 S_\tau} | S_\tau \leq -M] \\ &\leq p_0(M)e^{\psi_0} + (1 - p_0(M))e^{-\psi_0 M}, \end{aligned}$$

and it follows that

$$p_0(M) \geq \frac{1 - e^{-\psi_0 M}}{e^{\psi_0} - e^{-\psi_0 M}}.$$

Let p_0 be the probability that $S_t \geq 0$ for some t. Since $\mathbb{E}\zeta_1 < 0$, $S_t \to -\infty$ almost surely, and $\tau < \infty$ almost surely. As such, $p_0 \geq p_0(M)$. Take $M \to \infty$, we obtain

$$p_0 \geq e^{-\psi_0}.$$

Let $\lambda = e^{-\psi_0} \in (0,1)$. Then by Equation 6.20, we know that λ is the solution (in $(0,1)$) to the following equation

$$\lambda = 1 - q + q\lambda^{1+\alpha^{-1}}. \tag{6.21}$$

We pause to prove the above equation of λ has a unique solution on $(0,1)$. To see this, let $g(\lambda) = q\lambda^{1+\alpha^{-1}} - \lambda + 1 - q$. Then $g(0) > 0$ and $g(1) = 0$. In addition, $g'(\lambda) = (1+\alpha)\lambda^{1/\alpha}/q\alpha - 1$. Let $\lambda_0 = \exp\left(\alpha \log \frac{\alpha}{(1+\alpha)q}\right)$. Then $g'(\lambda_0) = 0$. In addition, $g'(\lambda) > 0$ for $\lambda \in (0, \lambda_0)$ and $g'(\lambda) < 0$ for $\lambda \in (\lambda_0, 1)$. As such, the function $g(\lambda)$ is monotonically decreasing from $(0, \lambda_0)$ and monotonically decreasing from $(\lambda_0, 1)$. Thus, $g(\lambda) = 0$ has a unique solution on $(0,1)$.

We now return to the proof of the lemma. Since by definition $p_0 \geq \lambda$, we have

$$\mathbb{P}(\max_{t \geq 1} S_t < 0) = 1 - p_0 \leq 1 - \lambda.$$

Thus, the proof boils down to bounding the difference between λ and 1. Let $1 - \lambda = q\delta$. Then Equation 6.21 can be expressed as

$$1 - \delta = (1 - q\delta)^{\alpha^{-1}+1}. \tag{6.22}$$

Taking the log on both sides, we have

$$(\alpha^{-1} + 1)q = \frac{q \log(1-\delta)}{\log(1 - q\delta)}.$$

We are mainly interested in the case where q is close to $1/(\alpha^{-1}+1)$, in which case the left hand side of the above equation would be close to 1. On the other hand, the right hand side approaches 1 as $\delta \to 0$ and diverges to infinity as $\delta \to 1$. As such, in order for the above equation to hold, δ can not be too large. Using Lemma 5, we get

$$(\alpha^{-1} + 1)q - 1 \geq \frac{(1-q)}{2}\delta.$$

Therefore, $\delta \leq \frac{2((\alpha^{-1}+1)q-1)}{1-q}$, and we conclude that

$$\mathbb{P}(\max_{t \geq 1} S_t < 0) \leq 1 - \lambda \leq \frac{2q((\alpha^{-1}+1)q-1)}{1-q} = \frac{2q(q-p)}{p(1-q)}.$$

□

Lemma 7. *Consider the following random walk defined in the proof of Proposition 6:*

$$\tilde{S}_k = \sum_{j=1}^{k} (p - \tilde{Z}_j), \quad \tilde{Z}_j \overset{i.i.d.}{\sim} \textit{Bern}\left(q_\delta\right),$$

where $q_\delta > p$. Define

$$t_* = \frac{3(p - q_\delta)}{4} < 0, \quad \textit{and} \quad c_h = \frac{1}{2} C(p, q_\delta) t_*^2,$$

where

$$C(p, q_\delta) = 1 + \max\left\{0, \min\left\{\frac{5}{3p} - \frac{2q_\delta}{3p^2}, \frac{1}{q_\delta}\right\}\right\} \geq 1.$$

Then for any

$$r > 4k^* \left(\frac{q_\delta}{p} - 1\right)^{-1},$$

we have

$$p(r) \triangleq \mathbb{P}\left(\max\left\{k : \tilde{S}_k \geq -pk^*\right\} \;=\; r\right) \leq \frac{q_\delta}{p(1-q_\delta)} \frac{e}{\pi\sqrt{rp(1-p)}} (q_\delta - p) e^{-c_h r}$$

Proof. Define events

$$A_1 = \left\{\sum_{j=1}^{r} \tilde{Z}_j = \lfloor pr + pk^* \rfloor\right\},$$

and

$$A_2 = \left\{\max_{k\geq 1} \sum_{j=r+1}^{r+k} (p - \tilde{Z}_j) \leq 0\right\}.$$

We will show that

$$\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} \;=\; r\right\} \subset A_1 \cap A_2.$$

First, note that $\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r$ implies the following three conditions:

$$(1)\tilde{S}_r \geq -pk^*, \quad (2)\tilde{S}_{r+1} \leq -pk^*, \quad \text{and} \quad (3)\max_{k\geq 1}(\tilde{S}_{r+k} - \tilde{S}_r) \leq 0.$$

Recalling the definition of $\tilde{S}_r$, conditions (1) and (2) are equivalent to

$$pr + pk^* + p - 1 \;\leq\; \sum_{j=1}^{r} \tilde{Z}_j \;\leq\; pr + pk^*.$$

Since $\sum_{j=1}^{r} \tilde{Z}_j$ is an integer, one of the following two events must happen: (a) there exists no integer between $pr + pk^*$ and $pr + pk^* + p - 1$. In this case, $\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} \;=\; r\right\}$ is an empty set; (b) there exists exactly one integer between $pr + pk^*$ and $pr + pk^* + p - 1$. Then it follows that $\sum_{j=1}^{r} \tilde{Z}_j = \lfloor pr + pk^* \rfloor$. In either case, we have shown that

$$\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} \;=\; r\right\} \subset A_1.$$

In addition, condition (3) is equivalent to event A_2. Therefore, we have shown that

$$\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} \;=\; r\right\} \subset A_1 \cap A_2.$$

Note that A_1 only depends on $\tilde{Z}_j, j \leq r$ and A_2 only depends on $\tilde{Z}_j, j > r$. Since $\{\tilde{Z}_j\}$ is a sequence of i.i.d Bernoulli variables, A_1 and A_2 are independent. Therefore

$$p(r) \leq P(A_1)P(A_2).$$

We now bound $P(A_1)$ and $P(A_2)$ separately. First,

$$P(A_1) = \text{Binom}(r, q_\delta; \lfloor pr + pk^* \rfloor).$$

where $\text{Binom}(r, q_\delta; \lfloor pr + pk^* \rfloor)$ is the probability of the binomial distribution $\text{Binom}(r, q_\delta; \cdot)$ at $\lfloor pr + pk^* \rfloor$. Let $m = \lfloor pr + pk^* \rfloor$, we have

$$\text{Binom}(r, q_\delta; \lfloor pr + pk^* \rfloor) = \frac{r!}{m!(r-m)!} e^{m \log q_\delta} e^{(r-m)\log(1-q_\delta)}.$$

Using Sterling's lemma, we obtain

$$\frac{r!}{m!(r-m)!} \leq \frac{e}{2\pi} \sqrt{\frac{r}{m(r-m)}} e^{r\log r - m\log m - (r-m)\log(r-m)}$$

Since $m \geq pr$, we know that

$$\sqrt{\frac{r}{m(r-m)}} \leq \sqrt{\frac{1}{rp(1-p)}}$$

Therefore

$$\text{Binom}(r, q_\delta; m) \leq \frac{e}{2\pi} \sqrt{\frac{r}{m(r-m)}} \exp\left(m \log \frac{rq_\delta}{m} + (r-m) \log \frac{r(1-q_\delta)}{r-m} \right).$$

First, for any $x < 0$, $\log(1+x) \leq x - \frac{1}{2}x^2$. Therefore,

$$\begin{aligned}
&(r-m) \log \frac{r(1-q_\delta)}{r-m} \\
&\leq r(1-q_\delta) - (r-m) - \frac{1}{2} \frac{(m - rq_\delta)^2}{r-m} \\
&\leq r(1-q_\delta) - (r-m) - \frac{1}{2} \frac{(m - rq_\delta)^2}{r}.
\end{aligned}$$

Next, since $\log(1+x) \leq x$ and $\log(1+x) \leq x - \frac{1}{2}x^2 + \frac{1}{3}x^3$ for any $x > -1$, we have

$$m \log \frac{rq_\delta}{m} \leq rq_\delta - m + \min\left\{ 0, -\frac{1}{2} \frac{(rq_\delta - m)^2}{m} + \frac{1}{3} \frac{(rq_\delta - m)^3}{m^2} \right\}$$

Combining the above two equations, we have

$$m \log \frac{rq_\delta}{m} + (r-m) \log \frac{r(1-q_\delta)}{r-m} \leq -\frac{1}{2} \frac{(m - rq_\delta)^2}{r} \left(1 + \max\left\{ 0, \frac{5r}{3m} - \frac{2r^2 q_\delta}{3m^2} \right\} \right).$$

Since $rp \leq m \leq rq_\delta$, we must have

$$\frac{5r}{3m} - \frac{2r^2 q_\delta}{3m^2} \geq \min\left\{\frac{5}{3p} - \frac{2q_\delta}{3p^2}, \frac{1}{q}\right\},$$

where the two terms on the right hand side is obtained by taking $m = rp$ and $m = rq_\delta$ respectively. Therefore,

$$\left(1 + \max\left\{0, \frac{5r}{3m} - \frac{2r^2 q_\delta}{3m^2}\right\}\right) \geq 1 + \max\left\{0, \min\left\{\frac{5}{3p} - \frac{2q_\delta}{3p^2}, \frac{1}{q_\delta}\right\}\right\} := C(p, q_\delta).$$

Note that for

$$r > 4k^* \left(\frac{q_\delta}{p} - 1\right)^{-1},$$

we have

$$rq_\delta - m > t_* r.$$

Therefore, taking the above Equations together, we have

$$-\frac{1}{2}\frac{(m - rq_\delta)^2}{r}\left(1 + \max\left\{0, \frac{5r}{3m} - \frac{2r^2 q_\delta}{3m^2}\right\}\right) \leq -\frac{1}{2} t_*^2 C(p, q) r.$$

Therefore

$$P(A_1) \leq \frac{e}{2\pi\sqrt{rp(1-p)}} e^{-c_h r}.$$

Now we bound the probability of event A_2. Using Lemma 6 with $N = 1/\alpha$, we get

$$P(A_2) \leq \frac{2q_\delta}{p(1-q_\delta)}(q_\delta - p).$$

Therefore

$$p(r) \leq P(A_1)P(A_2) \leq \frac{q_\delta}{p(1-q_\delta)} \frac{e}{\pi\sqrt{rp(1-p)}}(q_\delta - p) e^{-c_h r}.$$

The proof is now complete. □

Bibliography

[1] François Aguet et al. “Advances in analysis of low signal-to-noise images link dynamin and AP2 to the functions of an endocytic checkpoint”. In: *Developmental cell* 26.3 (2013), pp. 279–291.

[2] François Aguet et al. “Membrane dynamics of dividing cells imaged by lattice light-sheet microscopy”. In: *Molecular biology of the cell* 27.22 (2016), pp. 3418–3435.

[3] Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. “Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism”. In: *The Annals of Statistics* (2011), pp. 2533–2556.

[4] Ery Arias-Castro and Andrew Ying. “Detection of sparse mixtures: higher criticism and scan statistic”. In: *Electronic Journal of Statistics* 13.1 (2019), pp. 208–230.

[5] Zohar Attias-Geva et al. “Insulin-like growth factor-I receptor inhibition by specific tyrosine kinase inhibitor NVP-AEW541 in endometrioid and serous papillary endometrial cancer cell lines”. In: *Gynecologic oncology* 121.2 (2011), pp. 383–389.

[6] Francisco Azuaje. “Computational models for predicting drug responses in cancer research”. In: *Briefings in bioinformatics* 18.5 (2016), pp. 820–829.

[7] Kathryn Balmanno et al. “Intrinsic resistance to the MEK1/2 inhibitor AZD6244 (ARRY-142886) is associated with weak ERK1/2 signalling and/or strong PI3K signalling in colorectal cancer cell lines”. In: *International journal of cancer* 125.10 (2009), pp. 2332–2341.

[8] Rina Foygel Barber, Emmanuel J Candès, et al. “A knockoff filter for high-dimensional selective inference”. In: *The Annals of Statistics* 47.5 (2019), pp. 2504–2537.

[9] Rina Foygel Barber, Emmanuel J Candès, et al. “Controlling the false discovery rate via knockoffs”. In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[10] Ian Barnett, Rajarshi Mukherjee, and Xihong Lin. “The generalized higher criticism for testing SNP-set effects in genetic association studies”. In: *Journal of the American Statistical Association* 112.517 (2017), pp. 64–76.

[11] Jordi Barretina et al. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. In: *Nature* 483.7391 (2012), p. 603.

[12] Sumanta Basu et al. “Iterative random forests to discover predictive and stable high-order interactions”. In: *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1943–1948.

[13] Sumanta Basu et al. “Iterative random forests to discover predictive and stable high-order interactions”. In: *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1943–1948. ISSN: 0027-8424. DOI: 10.1073/pnas.1711236115. arXiv: 1706.08457. URL: http://arxiv.org/abs/1706.08457%20http://www.pnas.org/content/early/2018/01/17/1711236115.

[14] Mohsen Bayati and Andrea Montanari. “The LASSO risk for Gaussian matrices”. In: *IEEE Transactions on Information Theory* 58.4 (2011), pp. 1997–2017.

[15] Merle Behr et al. “Learning epistatic polygenic phenotypes with boolean interactions”. In: (2020).

[16] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[17] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*. Vol. 27. Cambridge university press, 1989.

[18] Leo Breiman. “Random Forests”. In: *Machine Learning* 45 (2001), pp. 1–33. DOI: 10.1023/A:1010933404324.

[19] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[20] Leo Breiman et al. *Classification and regression trees*. Chapman and Hall/CRC, 1984. ISBN: 9781351460491. DOI: 10.1201/9781315139470.

[21] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.

[22] T Cai, X Jessie Jeng, and Jiashun Jin. “Optimal detection of heterogeneous and heteroscedastic mixtures”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5 (2011), pp. 629–662.

[23] T Tony Cai, Yonina C Eldar, and Xiaodong Li. “Global testing against sparse alternatives in time-frequency analysis”. In: *The Annals of Statistics* 44.4 (2016), pp. 1438–1466.

[24] Tony T Cai and Yihong Wu. “Optimal detection of sparse mixtures against a given null distribution”. In: *IEEE Transactions on Information Theory* 60.4 (2014), pp. 2217–2232.

[25] D Ross Camidge et al. *Efficacy and safety of crizotinib in patients with advanced c-MET-amplified non-small cell lung cancer (NSCLC)*. 2014.

[26] Giordano Caponigro and William R Sellers. “Advances in the preclinical testing of cancer therapeutic hypotheses”. In: *Nature reviews Drug discovery* 10.3 (2011), p. 179.

[27] Strobl Carolin, Hothorn Torsten, and Zeileis Achim. “Party on! A New, Conditional Variable-Importance Measure for Random Forests Available in the party Package”. In: *the R journal* 1/2 (2009), pp. 14–17.

[28] Susan E Celniker et al. “Unlocking the secrets of the genome”. In: *Nature* 459.7249 (2009), p. 927.

[29] Hock Peng Chan, Guenther Walther, et al. “Optimal detection of multi-sample aligned sparse signals”. In: *The Annals of Statistics* 43.5 (2015), pp. 1865–1895.

[30] Liang Chen and Jingxuan Pan. “Dual cyclin-dependent kinase 4/6 inhibition by PD-0332991 induces apoptosis and senescence in oesophageal squamous cell carcinoma cells”. In: *British journal of pharmacology* 174.15 (2017), pp. 2427–2443.

[31] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754. URL: http://arxiv.org/abs/1603.02754%7B%5C%%7D0Ahttp://dx.doi.org/10.1145/2939672.2939785.

[32] James C Costello et al. “A community effort to assess and improve drug sensitivity prediction algorithms”. In: *Nature biotechnology* 32.12 (2014), p. 1202.

[33] Joshua C Denny et al. “PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations”. In: *Bioinformatics* 26.9 (2010), pp. 1205–1210.

[34] R Diaz-Uriarte and S de Andrés. “Gene Selection and Classification of Microarray Data Using Random Forest”. In: *BMC Bioinformatics* 7 (2006). DOI: 10.1186/1471-2105-7-3. URL: http://dx.doi.org/10.1186/1471-2105-7-3.

[35] David Donoho and Jiashun Jin. “Higher criticism for detecting sparse heterogeneous mixtures”. In: *Annals of Statistics* (2004), pp. 962–994.

[36] David Donoho and Jiashun Jin. “Higher criticism for large-scale inference, especially for rare and weak effects”. In: *Statistical Science* 30.1 (2015), pp. 1–25.

[37] E Drakos et al. “Activation of the p53 pathway by the MDM2 inhibitor nutlin-3a overcomes BCL2 overexpression in a preclinical model of diffuse large B-cell lymphoma associated with t (14; 18)(q32; q21)”. In: *Leukemia* 25.5 (2011), pp. 856–867.

[38] Raaz Dwivedi et al. “Stable discovery of interpretable subgroups via calibration in causal studies”. In: *arXiv preprint arXiv:2008.10109* (2020).

[39] Marcelo Ehrlich et al. “Endocytosis by random initiation and stabilization of clathrin-coated pits”. In: *Cell* 118.5 (2004), pp. 591–605.

[40] Caroline M Emery et al. “MEK1 mutations confer resistance to MEK and B-RAF inhibition”. In: *Proceedings of the National Academy of Sciences* 106.48 (2009), pp. 20411–20416.

[41] Francisco J Esteva et al. “Molecular predictors of response to trastuzumab and lapatinib in breast cancer”. In: *Nature reviews Clinical oncology* 7.2 (2010), p. 98.

[42] Yingying Fan et al. “RANK: large-scale inference with graphical nonlinear knockoffs”. In: *Journal of the American Statistical Association* 115.529 (2020), pp. 362–379.

[43] William Fithian and Lihua Lei. "Conditional calibration for false discovery rate control under dependence". In: *arXiv preprint arXiv:2007.10438* (2020).

[44] William Fithian, Dennis Sun, and Jonathan Taylor. "Optimal inference after model selection". In: *arXiv preprint arXiv:1410.2597* (2014).

[45] J Galambos and E Seneta. "Regularly varying sequences". In: *Proceedings of the American Mathematical Society* 41.1 (1973), pp. 110–116.

[46] Gene H Golub, Michael Heath, and Grace Wahba. "Generalized cross-validation as a method for choosing a good ridge parameter". In: *Technometrics* 21.2 (1979), pp. 215–223.

[47] Alexandre Grassart et al. "Actin and dynamin2 dynamics and interplay during clathrin-mediated endocytosis". In: *Journal of Cell Biology* 205.5 (2014), pp. 721–735.

[48] Betül Güvenç Paltun, Hiroshi Mamitsuka, and Samuel Kaski. "Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches". In: *Briefings in Bioinformatics* (2019).

[49] Peter Hall, Jiashun Jin, et al. "Innovated higher criticism for detecting sparse signals in correlated noise". In: *The Annals of Statistics* 38.3 (2010), pp. 1686–1732.

[50] Peter Hall and Jiashun Jin. "Properties of higher criticism under strong dependence". In: *The Annals of Statistics* (2008), pp. 381–402.

[51] Kan He et al. "Hsp90 inhibitors promote p53-dependent apoptosis through PUMA and Bax". In: *Molecular cancer therapeutics* 12.11 (2013), pp. 2559–2568.

[52] Kangmin He et al. "Dynamics of Auxilin 1 and GAK in clathrin-mediated traffic". In: *Journal of Cell Biology* 219.3 (2020).

[53] Ying C Henderson et al. "MEK inhibitor PD0325901 significantly reduces the growth of papillary thyroid carcinoma cells in vitro and in vivo". In: *Molecular cancer therapeutics* 9.7 (2010), pp. 1968–1976.

[54] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[55] Sture Holm. "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics* (1979), pp. 65–70.

[56] Sun Hae Hong, Christa L Cortesio, and David G Drubin. "Machine-learning-based analysis in genome-edited cells reveals the efficiency of clathrin-mediated endocytosis". In: *Cell reports* 12.12 (2015), pp. 2121–2130.

[57] T Hothorn, K Hornik, and A Zeileis. "Unbiased Recursive Partitioning: A Conditional Inference Framework". In: *Journal of Computational and Graphical Statistics* 15 (2006). DOI: 10.1198/106186006X133933. URL: http://dx.doi.org/10.1198/106186006X133933.

[58] Vân Anh Huynh-Thu et al. "Inferring regulatory networks from expression data using tree-based methods". In: *PLoS ONE* 5.9 (2010). ISSN: 19326203. DOI: 10.1371/journal.pone.0012776.

[59] Dankwart Jaeschke. "The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals". In: *The Annals of Statistics* (1979), pp. 108–115.

[60] In Sock Jang et al. "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data". In: *Biocomputing 2014*. World Scientific, 2014, pp. 63–74.

[61] Silke Janitza, Ender Celik, and Anne Laure Boulesteix. "A computationally fast variable importance test for random forests for high-dimensional data". In: *Advances in Data Analysis and Classification* 12.4 (2016), pp. 1–31. ISSN: 18625355. DOI: 10.1007/s11634-016-0270-x.

[62] Zuzana Kadlecova et al. "Regulation of clathrin-mediated endocytosis by hierarchical allosteric activation of AP2". In: *Journal of Cell Biology* 216.1 (2017), pp. 167–179.

[63] Marko Kaksonen and Aurélien Roux. "Mechanisms of clathrin-mediated endocytosis". In: *Nature Reviews Molecular Cell Biology* 19.5 (2018), p. 313.

[64] Jalil Kazemitabar et al. "Variable importance using decision trees". In: *Advances in Neural Information Processing Systems*. 2017, pp. 426–435.

[65] Zheng Tracy Ke, Jun S Liu, and Yucong Ma. "Power of FDR Control Methods: The Impact of Ranking Algorithm, Tampered Design, and Symmetric Statistic". In: *arXiv preprint arXiv:2010.08132* (2020).

[66] Tom Kirchhausen, David Owen, and Stephen C Harrison. "Molecular structure, function, and dynamics of clathrin-mediated membrane traffic". In: *Cold Spring Harbor perspectives in biology* 6.5 (2014), a016725.

[67] Isaac S Kohane. "Ten things we have to do to achieve precision medicine". In: *Science* 349.6243 (2015), pp. 37–38.

[68] J. B. Kruskall. "The Symmetric Time-Warping Problem : From Continuous to Discrete". In: *Time Warps, String Edits, and Macromolecules*. 1983.

[69] Karl Kumbier et al. "Refining interaction search through signed iterative Random Forests". In: *arXiv preprint arXiv:1810.07287* (2018).

[70] Xiao-Feng Le and Robert C Bast Jr. "Src family kinases and paclitaxel sensitivity". In: *Cancer biology & therapy* 12.4 (2011), pp. 260–269.

[71] Jung Bok Jae Won Lee et al. "An extensive comparison of recent classification tools applied to microarray data". In: *Computational Statistics and Data Analysis* 48.4 (2005), pp. 869–885. DOI: 10.1016/j.csda.2004.03.017.

[72] Seunggeung Lee et al. "Rare-variant association analysis: study designs and statistical tests". In: *The American Journal of Human Genetics* 95.1 (2014), pp. 5–23.

[73] Jian Li and David Siegmund. "Higher criticism: p-values and criticism". In: *The Annals of Statistics* 43.3 (2015), pp. 1323–1350.

[74] Chinghway Lim and Bin Yu. "Estimation stability with cross-validation (ESCV)". In: *Journal of Computational and Graphical Statistics* 25.2 (2016), pp. 464–492.

[75] Xiang Ling et al. "FL118 induces p53-dependent senescence in colorectal cancer cells by promoting degradation of MdmX". In: *Cancer research* 74.24 (2014), pp. 7487–7497.

[76] Jingbo Liu and Philippe Rigollet. "Power analysis of knockoff filters for correlated designs". In: *Advances in Neural Information Processing Systems*. 2019, pp. 15420–15429.

[77] Yuqing Liu et al. "Inhibition of PDGF, TGF-β, and Abl signaling and reduction of liver fibrosis by the small molecule Bcr-Abl tyrosine kinase antagonist Nilotinib". In: *Journal of hepatology* 55.3 (2011), pp. 612–625.

[78] Dinah Loerke et al. "Cargo and dynamin regulate clathrin-coated pit maturation". In: *PLoS biology* 7.3 (2009).

[79] Wei-Yin Loh. "Fifty years of classification and regression trees". In: *International Statistical Review* 82.3 (2014), pp. 329–348. ISSN: 17515823. DOI: 10.1111/insr.12016.

[80] Gilles Louppe. "Understanding random forests: From theory to practice". In: *arXiv preprint arXiv:1407.7502* (2014).

[81] Gilles Louppe et al. "Understanding variable importances in forests of randomized trees". In: *Advances in Neural Information Processing Systems 26*. 2013, pp. 431–439. URL: http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf.

[82] Scott Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874* (2017).

[83] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. "Consistent Individualized Feature Attribution for Tree Ensembles". In: *ArXiv e-prints arXiv:1802.03888* (2018). arXiv: 1802.03888. URL: http://arxiv.org/abs/1802.03888.

[84] Stewart MacArthur et al. "Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions". In: *Genome biology* 10.7 (2009), p. 1.

[85] Ultan McDermott et al. "Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling". In: *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 19936–19941.

[86] Harvey T McMahon and Emmanuel Boucrot. "Molecular mechanism and physiological functions of clathrin-mediated endocytosis". In: *Nature reviews Molecular cell biology* 12.8 (2011), p. 517.

[87] Pierre Mordant et al. "Dependence on phosphoinositide 3-kinase and RAS-RAF pathways drive the activity of RAF265, a novel RAF/VEGFR2 inhibitor, and RAD001 (Everolimus) in combination". In: *Molecular cancer therapeutics* 9.2 (2010), pp. 358–368.

[88] Amit Moscovich and Boaz Nadler. "Fast calculation of boundary crossing probabilities for Poisson processes". In: *Statistics & Probability Letters* 123 (2017), pp. 177–182.

[89] Amit Moscovich, Boaz Nadler, and Clifford Spiegelman. "On the exact Berk-Jones statistics and their p-value calculation". In: *Electronic Journal of Statistics* 10.2 (2016), pp. 2329–2354.

[90] W James Murdoch, Peter J Liu, and Bin Yu. "Beyond word importance: Contextual decomposition to extract interactions from LSTMs". In: *arXiv preprint arXiv:1801.05453* (2018).

[91] W. James Murdoch et al. "Interpretable machine learning: definitions, methods, and applications". In: *ArXiv e-prints* (2019), pp. 1–11. arXiv: 1901.04592. URL: http://arxiv.org/abs/1901.04592.

[92] Atsuhiko T Naito et al. "Promotion of CHIP-mediated p53 degradation protects the heart from ischemic injury". In: *Circulation research* 106.11 (2010), p. 1692.

[93] David L Nelson, Albert L Lehninger, and Michael M Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[94] Stefano Nembrini, Inke R. König, and Marvin N. Wright. "The revival of the Gini importance?" In: *Bioinformatics* 34.21 (2018), pp. 3711–3718. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty373.

[95] Jerzy Neyman and Egon Sharpe Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[96] C Nishioka et al. "ZD6474 induces growth arrest and apoptosis of human leukemia cells, which is enhanced by concomitant use of a novel MEK inhibitor, AZD6244". In: *Leukemia* 21.6 (2007), pp. 1308–1310.

[97] Marc Noé. "The calculation of distributions of two-sided Kolmogorov-Smirnov type statistics". In: *The Annals of Mathematical Statistics* (1972), pp. 58–64.

[98] Art B Owen. "Nonparametric likelihood confidence bands for a distribution function". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 516–521.

[99] E Patterson and M Sesia. "knockoff: The Knockoff Filter for Controlled Variable Selection". In: *R package version 0.3. 0* (2017).

[100] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[101] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[102] Gregory Riddick et al. "Predicting in vitro drug sensitivity using Random Forests". In: *Bioinformatics* 27.2 (2011), pp. 220–224.

[103] Wendy Rodenburg et al. "A Framework to Identify Physiological Responses in Microarray Based Gene Expression Studies: Selection and Interpretation of Biologically Relevant Genes". In: *Physiological Genomics* 33 (2008). DOI: 10.1152/physiolgenomics.00167.2007. URL: http://dx.doi.org/10.1152/physiolgenomics.00167.2007.

[104] Mark A Rubin. "Health: Make precision medicine work for cancer care". In: *Nature News* 520.7547 (2015), p. 290.

[105] Ando Saabas. *Interpreting random forests*. 2014. URL: http://blog.datadive.net/interpreting-random-forests/%20http://blog.datadive.net/random-forest-interpretation-with-scikit-learn/.

[106] Stan Salvador and Philip Chan. "Toward accurate dynamic time warping in linear time and space". In: *Intelligent Data Analysis* 11.5 (2007), pp. 561–580.

[107] Marco Sandri and Paola Zuccolotto. "A bias correction algorithm for the gini variable importance measure in classification trees". In: *Journal of Computational and Graphical Statistics* 17.3 (2008), pp. 611–628. ISSN: 10618600.

[108] Sanat K Sarkar and Cheng Yong Tang. "Adjusting the Benjamini-Hochberg method for controlling the false discovery rate in knockoff assisted variable selection". In: *arXiv preprint arXiv:2102.09080* (2021).

[109] Erwan Scornet, Gerard Biau, and Jean Philippe Vert. "Consistency of random forests". In: *Annals of Statistics* 43.4 (2015), pp. 1716–1741. ISSN: 00905364. arXiv: 1405.2881.

[110] Andrew W Senior et al. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792 (2020), pp. 706–710.

[111] Jinjin Shao et al. "Gefitinib synergizes with irinotecan to suppress hepatocellular carcinoma via antagonizing Rad51-mediated DNA-repair". In: *PLoS One* 11.1 (2016), e0146968.

[112] Zbynek Šidák. "On multivariate normal probabilities of rectangles: their dependence on correlations". In: *The Annals of Mathematical Statistics* 39.5 (1968), pp. 1425–1434.

[113] Nikola Simidjievski et al. "Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice". In: *BioRxiv* (2019), p. 719542.

[114] Chandan Singh, W James Murdoch, and Bin Yu. "Hierarchical interpretations for neural network predictions". In: *arXiv preprint arXiv:1806.05337* (2018).

[115] Asher Spector and Lucas Janson. "Powerful Knockoffs via Minimizing Reconstructability". In: *arXiv preprint arXiv:2011.14625* (2020).

[116] John H Strickler et al. "Phase I study of bevacizumab, everolimus, and panobinostat (LBH-589) in advanced solid tumors". In: *Cancer chemotherapy and pharmacology* 70.2 (2012), pp. 251–258.

[117] C Strobl, A L Boulesteix, and T Augustin. "Unbiased Split Selection for Classification Trees Based on the Gini Index". In: *Computational Statistics {&} Data Analysis* 52 (2007). DOI: 10.1016/j.csda.2006.12.030. URL: http://dx.doi.org/10.1016/j.csda.2006.12.030.

[118] Carolin Strobl et al. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution". In: *BMC Bioinformatics* 8 (2007). DOI: 10.1186/1471-2105-8-25. URL: http://dx.doi.org/10.1186/1471-2105-8-25.

[119] Carolin Strobl et al. "Conditional variable importance for random forests". In: *BMC Bioinformatics* 9.1 (2008), p. 307. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-307. URL: http://dx.doi.org/10.1186/1471-2105-9-307.

[120] Erik Štrumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: *Knowledge and Information Systems* 41.3 (2014), pp. 647–665. ISSN: 02193116. DOI: 10.1007/s10115-013-0679-x.

[121] Weijie Su, Małgorzata Bogdan, Emmanuel Candes, et al. "False discoveries occur early on the lasso path". In: *The Annals of statistics* 45.5 (2017), pp. 2133–2150.

[122] Marcus J Taylor, David Perrais, and Christien J Merrifield. "A high precision survey of the molecular dynamics of mammalian clathrin-mediated endocytosis". In: *PLoS biology* 9.3 (2011).

[123] Xiaoying Tian, Jonathan Taylor, et al. "Selective inference with a randomized response". In: *The Annals of Statistics* 46.2 (2018), pp. 679–710.

[124] Anna Tutusaus et al. "Antiapoptotic BCL-2 proteins determine sorafenib/regorafenib resistance and BH3-mimetic efficacy in hepatocellular carcinoma". In: *Oncotarget* 9.24 (2018), p. 16701.

[125] Stefan Wager and Susan Athey. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: *Journal of the American Statistical Association* 1459 (2018), pp. 1–15. ISSN: 1537274X. DOI: 10.1080/01621459.2017.1319839. arXiv: 1510.04342. URL: https://doi.org/10.1080/01621459.2017.1319839%20http://arxiv.org/abs/1510.04342.

[126] Xinxin Wang et al. "DASC, a sensitive classifier for measuring discrete early stages in clathrin-mediated endocytosis". In: *eLife* 9 (2020), e53686.

[127] Wolfgang Warsch et al. "High STAT5 levels mediate imatinib resistance and indicate disease progression in chronic myeloid leukemia". In: *Blood, The Journal of the American Society of Hematology* 117.12 (2011), pp. 3409–3420.

[128] Pengfei Wei, Zhenzhou Lu, and Jingwen Song. "Variable importance analysis: A comprehensive review". In: *Reliability Engineering and System Safety* 142 (2015), pp. 399–432. ISSN: 09518320. DOI: 10.1016/j.ress.2015.05.018.

[129] Asaf Weinstein, Rina Barber, and Emmanuel Candes. "A power and prediction analysis for knockoffs with lasso statistics". In: *arXiv preprint arXiv:1712.06465* (2017).

[130] Marvin Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software, Articles* 77.1 (2017), pp. 1–17. ISSN: 1548-7660. DOI: 10.18637/jss.v077.i01. URL: https://www.jstatsoft.org/v077/i01.

[131] Yi-Ju Wu et al. "Involvement of 14-3-3 proteins in regulating tumor progression of hepatocellular carcinoma". In: *Cancers* 7.2 (2015), pp. 1022–1036.

[132] Siqi Wu et al. "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks". In: *Proceedings of the National Academy of Sciences* 113.16 (2016), pp. 4290–4295.

[133] Joy C Yang et al. "Effect of the specific Src family kinase inhibitor saracatinib on osteolytic lesions using the PC-3 bone model". In: *Molecular cancer therapeutics* 9.6 (2010), pp. 1629–1637.

[134] Yang Yang et al. "ZD6474 induces growth arrest and apoptosis of GIST-T1 cells, which is enhanced by concomitant use of sunitinib". In: *Cancer science* 97.12 (2006), pp. 1404–1409.

[135] Bin Yu and Karl Kumbier. "Veridical data science". In: *Proceedings of the National Academy of Sciences* 117.8 (2020), pp. 3920–3929.

[136] Ju Zhang et al. "DeepCLA: A Hybrid Deep Learning Approach for the Identification of Clathrin". In: *Journal of Chemical Information and Modeling* (2020).

[137] Zhengze Zhou and Giles Hooker. "Unbiased Measurement of Feature Importance in Tree-Based Methods". In: *arXiv preprint arXiv:1903.05179* (2019).