TECHNISCHE
UNIVERSITÄT
DARMSTADT

Development of Fast Machine Learning Algorithms for False Discovery Rate Control in Large-Scale High-Dimensional Data

vom Fachbereich Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation
von

Jasin Machkour, M.Sc.

Erstgutachter: Prof. Dr.-Ing. Michael Muma
Zweitgutachter: Prof. Dr. Daniel P. Palomar

Darmstadt, 2024

Machkour, Jasin– Development of Fast Machine Learning Algorithms for False Discovery Rate Control in Large-Scale High-Dimensional Data
Darmstadt, Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2024
URN: urn:nbn:de:tuda-tuprints-282317
Tag der mündlichen Prüfung: 23. August 2024

Entwicklung von schnellen Algorithmen für maschinelles Lernen zur Kontrolle der Falschentdeckungsrate in großen hochdimensionalen Daten

Kurzfassung

In dieser Dissertation werden Algorithmen für maschinelles Lernen zur Kontrolle der Falschentdeckungsrate (FDR) für große hochdimensionale Daten entwickelt. Die Gewährleistung der Reproduzierbarkeit von Entdeckungen, die auf hochdimensionalen Daten basieren, ist für zahlreiche Anwendungen von zentraler Bedeutung. Die entwickelten Algorithmen führen eine schnelle Variablenauswahl in großen hochdimensionalen Daten durch, in denen die Anzahl der Variablen viel größer sein kann als die Anzahl der Stichproben. Dies beinhaltet groß angelegte Daten mit bis zu Millionen von Variablen, wie z. B. genomweite Assoziationsstudien (GWAS). Theoretische FDR-Kontrollgarantien für endliche Stichproben, die auf der Martingaltheorie beruhen, beweisen die Vertrauenswürdigkeit der entwickelten Methoden. Die praktischen Open-Source-R-Softwarepakete TRexSelector und tlars, die die vorgeschlagenen Algorithmen implementieren, wurden im Comprehensive R Archive Network (CRAN) veröffentlicht. Umfangreiche numerische Experimente und reale Probleme in der Biomedizin- und Finanztechnik demonstrieren die Leistungsfähigkeit in anspruchsvollen Anwendungsfällen. Die ersten drei Hauptteile dieser Dissertation präsentieren die methodischen und theoretischen Beiträge, während der vierte Hauptteil die praktischen Beiträge enthält.

Der erste Hauptteil (Kapitel 3) widmet sich dem Terminating-Random Experiments (T-Rex) Selektor, einem neuen schnellen Variablenselektionsverfahren für hochdimensionale Daten. Der T-Rex Selektor kontrolliert eine benutzerdefinierte Ziel-FDR und maximiert gleichzeitig die Anzahl der ausgewählten Variablen. Dies wird durch die Fusionierung der Lösungen mehrerer früh beendeter Zufallsexperimente erreicht. Die Experimente werden mit einer Kombination aus den ursprünglichen Kandidaten-Variablen und mehreren unabhängigen Sätzen von zufällig generierten Dummy-Variablen durchgeführt. Die FDR-Kontrolleigenschaft wird mit Hilfe der Martingaltheorie für endliche Stichproben bewiesen. Die Komplexität des T-Rex Selektors wächst linear mit der Anzahl der Kandidatenvariablen. Darüber hinaus ist seine Berechnungszeit im Vergleich zu modernsten Benchmark-Methoden in großen Datensätzen um mehr als zwei Größenordnungen schneller. Daher skaliert der T-Rex Selektor in einer angemessenen Rechenzeit auf Millionen von Kandidatenvariablen. Ein wichtiger Anwendungsfall des T-Rex Selektors ist die Bestimmung reproduzierbarer Assoziationen zwischen Phänotypen und Genotypen in GWAS, was für die personalisierte Medizin und die Arzneimittelentdeckung unerlässlich ist.

Der zweite Hauptteil (Kapitel 4) beschäftigt sich mit abhängigkeitssensitiven FDR-Kontrollalgorithmen für große hochdimensionale Daten. Die hochdimensionalen Daten in vielen Anwendungen der Biomedizin- und Finanztechnik enthalten oft hochkorrelierte Kandidaten-Variablen (z. B. Genexpressionsdaten und Aktienrenditen). Für solche Anwendungen wurde das abhängigkeitssensitive T-Rex (T-Rex+DA) Framework entwickelt. Es erweitert das gewöhnliche T-Rex Framework um die Berücksichtigung von Abhängigkeitsstrukturen zwischen den Kandidaten-Variablen. Dies wird durch die Integration grafischer Modelle in das T-Rex Framework erreicht. Hierdurch wird es möglich, die Abhängigkeitsstruktur zwischen den Variablen effektiv zu nutzen und Mechanismen zur Penalisierung von Variablen zu entwickeln, die zu einer garantierten FDR-Kontrolle führen.

Im dritten Hauptteil (Kapitel 5) werden Algorithmen für die Auswahl gruppierter Variablen mit gewährleisteter FDR-Kontrolle vorgeschlagen. Dieser Ansatz zur Bewältigung der Herausforderungen, die sich aus dem Vorhandensein von Gruppen hochgradig abhängiger Variablen in den Daten ergeben, unterscheidet sich von dem konservativeren Variablenbestrafungsansatz, der im zweiten Teil dieser Dissertation entwickelt wurde. Das heißt, anstatt die wenigen wirklich aktiven Variablen unter den Gruppen hochkorrelierter Variablen zu finden, besteht das Ziel darin, alle Gruppen hochkorrelierter Variablen auszuwählen, die mindestens eine wirklich aktive Variable enthalten. In der Genomforschung, insbesondere bei GWAS, sind Variablenselektionsverfahren für gruppierte Variablen von großer Bedeutung, da man nicht an der Identifizierung einiger weniger Einzelnukleotid-Polymorphismen (SNPs), die mit einer gewissen Krankheit assoziiert sind, interessiert ist, sondern an den gesamten Gruppen korrelierter SNPs, die auf relevante Stellen im Genom hinweisen.

Der vierte Hauptteil dieser Dissertation (Kapitel 6 und 7) demonstriert die Anwendung der entwickelten Methoden auf praktische Probleme sowohl in der Biomedizintechnik als auch in der Finanztechnik. Zu den biomedizinischen Anwendungen gehören (i) eine halb-reale GWAS, (ii) ein Datensatz des Humanen Immundefizienz-Virus Typ 1 (HIV-1) mit zugehörigen Messungen der Arzneimittelresistenz und (iii) ein Brustkrebs-Datensatz mit zugehörigen Überlebenszeiten der Patienten. Zu den finanztechnischen Anwendungen gehören (i) die genaue Nachverfolgung des S&P 500-Index unter Verwendung eines vierteljährlich aktualisierten und neu ausbalancierten Nachverfolgungsportfolios, das aus wenigen Aktien besteht, und (ii) eine Faktoranalyse der S&P 500-Aktienrenditen. Die gemeinsame Herausforderung aller betrachteten Anwendungen liegt in der Detektion der wenigen aktiven Variablen (d. h. SNPs, Mutationen, Gene, Aktien) unter vielen nicht aktiven Variablen in u. a. großen hochdimensionalen Datensätzen.

Zusammenfassend werden in dieser Dissertation neue schnelle und skalierbare Algorithmen des maschinellen Lernens mit nachweisbaren FDR-Kontrollgarantien für die Variablenselektion in großen hochdimensionalen Daten entwickelt und analysiert. Die entwickelten Algorithmen und Open-Source-Softwarepakete haben reproduzierbare Entdeckungen in verschiedenen Anwendungen ermöglicht, die von der Biomedizin- bis zur Finanztechnik reichen.

Development of Fast Machine Learning Algorithms for False Discovery Rate Control in Large-Scale High-Dimensional Data

Abstract

This dissertation develops false discovery rate (FDR) controlling machine learning algorithms for large-scale high-dimensional data. Ensuring the reproducibility of discoveries based on high-dimensional data is pivotal in numerous applications. The developed algorithms perform fast variable selection tasks in large-scale high-dimensional settings where the number of variables may be much larger than the number of samples. This includes large-scale data with up to millions of variables such as genome-wide association studies (GWAS). Theoretical finite sample FDR-control guarantees based on martingale theory have been established proving the trustworthiness of the developed methods. The practical open-source R software packages TRexSelector and tlars, which implement the proposed algorithms, have been published on the Comprehensive R Archive Network (CRAN). Extensive numerical experiments and real-world problems in biomedical and financial engineering demonstrate the performance in challenging use-cases. The first three main parts of this dissertation present the methodological and theoretical contributions, while the fourth main part contains the practical contributions.

The first main part (Chapter 3) is dedicated to the Terminating-Random Experiments (T-Rex) selector, a new fast variable selection framework for high-dimensional data. The proposed T-Rex selector controls a user-defined target FDR while maximizing the number of selected variables. This is achieved by fusing the solutions of multiple early terminated random experiments. The experiments are conducted on a combination of the candidate variables and multiple independent sets of randomly generated dummy variables. A finite sample proof of the FDR control property is provided using martingale theory. The computational complexity of the T-Rex selector grows linearly with the number of candidate variables. Furthermore, its computation time is more than two orders of magnitude faster compared to state-of-the-art benchmark methods in large-scale data settings. Therefore, the T-Rex selector scales to millions of candidate variables in a reasonable computation time. An important use-case of the T-Rex selector is determining reproducible associations between phenotypes and genotypes in GWAS, which is imperative in personalized medicine and drug discovery.

The second main part (Chapter 4) concerns dependency-aware FDR-controlling algorithms for large-scale high-dimensional data. In many biomedical and financial applications, the high-dimensional data sets often contain highly correlated candidate variables (e.g., gene expression data and stock returns). For such applications, the dependency-aware T-Rex (T-Rex+DA)

framework has been developed. It extends the ordinary T-Rex framework by accounting for dependency structures among the candidate variables. This is achieved by integrating graphical models within the T-Rex framework, which allows to effectively harness the dependency structure among variables and to develop variable penalization mechanisms that guarantee FDR control.

In the third main part (Chapter 5), algorithms for joint grouped variable selection and FDR control are proposed. This approach to tackling the challenges resulting from the presence of groups of highly dependent variables in the data is different to the more conservative variable penalization approach that has been developed in the second part of this dissertation. That is, instead of finding the few true active variables among groups of highly correlated variables, the goal is to select all groups of highly correlated variables that contain at least one true active variable. In genomics research, especially for GWAS, grouped variable selection approaches are highly relevant, since one is not interested in identifying a few single-nucleotide polymorphisms (SNPs) that are associated with a disease of interest but rather the entire groups of correlated SNPs that point to relevant locations on the genome.

The fourth main part of this dissertation (Chapters 6 and 7) demonstrates the application of the developed methods to practical problems in biomedical engineering as well as financial engineering. The biomedical applications include (i) a semi-real-world GWAS, (ii) a human immunodeficiency virus type 1 (HIV-1) data set with associated drug resistance measurements, and (iii) a breast cancer data set with associated survival times of the patients. The financial engineering applications include (i) accurately tracking the S&P 500 index using a quarterly updated and rebalanced tracking portfolio that consists of few stocks and (ii) a factor analysis of S&P 500 stock returns. The common challenge of all considered applications lies in detecting the few true active variables (i.e., SNPs, mutations, genes, stocks) among many non-active variables in, among other things, large-scale high-dimensional settings.

Summarizing, this dissertation develops and analyses new fast and scalable machine learning algorithms with provable FDR-control guarantees for variable selection tasks in large-scale high-dimensional data. The developed algorithms and the associated open-source software packages have enabled making reproducible discoveries in various real-world applications ranging from biomedical to financial engineering.

Publications

The following works have been published during the period of doctoral candidacy.

Internationally Refereed Journal Articles

J. Machkour, M. Muma, and D. P. Palomar. "The Terminating-Random Experiments Selector: Fast High-Dimensional Variable Selection with False Discovery Rate Control". In: *arXiv preprint arXiv:2110.06048* (under review).

J. Machkour, M. Muma, and D. P. Palomar. "High-Dimensional False Discovery Rate Control for Dependent Variables". In: *arXiv preprint arXiv:2401.15796* (under review).

J. Machkour, D. P. Palomar, and M. Muma. "FDR-Controlled Portfolio Optimization for Sparse Financial Index Tracking". In: *arXiv preprint arXiv:2401.15139* (under review).

J. Machkour, M. Muma, B. Alt, and A. M. Zoubir. "A robust adaptive Lasso estimator for the independent contamination model". In: *Signal Process.* 174 (2020), p. 107608.

Internationally Refereed Conference Papers

T. Koka, **J. Machkour**, and M. Muma. "False Discovery Rate Control for Gaussian Graphical Models via Neighborhood Screening". In: *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*. 2024, pp. 2482–2486.

J. Machkour, A. Breloy, M. Muma, D. P. Palomar, and F. Pascal. "Sparse PCA with False Discovery Rate Controlled Variable Selection". In: *Proc. IEEE 49th Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. 2024, pp. 9716–9720.

J. Machkour, M. Muma, and D. P. Palomar. "FDR-Controlled Sparse Index Tracking with Autoregressive Stock Dependency Models". In: *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*. 2024, pp. 2662–2666.

J. Machkour, M. Muma, and D. P. Palomar. "False Discovery Rate Control for Fast Screening of Large-Scale Genomics Biobanks". In: *Proc. 22nd IEEE Statist. Signal Process. Workshop (SSP)*. 2023, pp. 666–670.

J. Machkour, M. Muma, and D. P. Palomar. "The Informed Elastic Net for Fast Grouped Variable Selection and FDR Control in Genomics Research". In: *Proc. IEEE 9th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*. 2023, pp. 466–470.

F. Scheidt, **J. Machkour**, and M. Muma. "Solving FDR-Controlled Sparse Regression Problems with Five Million Variables on a Laptop". In: *Proc. IEEE 9th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*. 2023, pp. 116–120.

J. Machkour, M. Muma, and D. P. Palomar. "False Discovery Rate Control for Grouped Variable Selection in High-Dimensional Linear Models Using the T-Knock Filter". In: *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*. 2022, pp. 892–896.

Internationally Refereed Software Packages

J. Machkour, S. Tien, D. P. Palomar, and M. Muma. *tlars: The T-LARS Algorithm: Early-Terminated Forward Variable Selection*. R package version 1.0.1. 2024. URL: https://CRAN.R-project.org/package=tlars.

J. Machkour, S. Tien, D. P. Palomar, and M. Muma. *TRexSelector: T-Rex Selector: High-Dimensional Variable Selection & FDR Control*. R package version 1.0.0. 2024. URL: https://CRAN.R-project.org/package=TRexSelector.

Acknowledgments

First and foremost, I would like to thank my supervisor, Prof. Dr.-Ing. Michael Muma, for giving me the opportunity, freedom, guidance, and support to pursue my research interests. His support enabled me to collaborate with many brilliant scientists around the world and paved the way for enriching scientific stays in Hong Kong and Paris.

I would also like to thank my co-supervisor, Prof. Dr. Daniel P. Palomar, whom I visited twice at The Hong Kong University of Science and Technology (HKUST) for a total of ten months. His hospitality, guidance, and support made my stays in Hong Kong an invaluable experience from a scientific and personal perspective.

I fondly remember many meetings with my supervisors that lasted for hours, during which we discussed new ideas, algorithms, and applications. Both of my advisors would challenge every idea, argument, and thought, allowing me to sharpen my thinking and grow both scientifically and personally.

With genuine gratitude, I extend my heartfelt thanks to Prof. Dr.-Ing. Abdelhak M. Zoubir, the head of the Signal Processing Group at TU Darmstadt, for sparking my interest and curiosity in the fields of signal processing, machine learning, and data science through his enthusiasm and guidance in the early years of my studies.

I would like to thank Prof. Dr.-Ing. Grace Li Zhang and Prof. Dr.-Ing. Christoph Hoog Antink for being members of my PhD committee and Prof. Dr.-Ing. Dr. h.c. Ralf Steinmetz for chairing the PhD committee.

I wish to thank Prof. Dr. Frédéric Pascal and Prof. Dr. Arnaud Breloy for our fruitful collaboration and for giving me the opportunity to spend three months at Université Paris-Saclay, CentraleSupélec. I would also like to thank Dr. Stefano Fortunati for showing me all the hidden gems of Paris during my stay.

I thank Prof. Dr. Wing-Kin Ma from The Chinese University of Hong Kong and Prof. Dr. Hing Cheung So from City University of Hong Kong for inviting me to give scientific talks at their universities.

I am very grateful to have had the opportunity to be a scientist within the LOEWE center emergenCITY. Many thanks go to the entire emergenCity team and especially the scientific coordinator, Prof. Dr.-Ing. Matthias Hollick, the head of the program area Cyber-Physical Systems, Prof. Dr. rer. nat. Oskar von Stryk, the managing director, Katharina Kleinschnit-

ger, the former managing director, Dr. Anne Hofmeister, the lab engineer, Julian Euler, and the administrative staff, Gloria Philipp, Vanessa Winkels, and Uta Drews.

Many thanks go to the current and former members and visitors of the Robust Data Science Group of Prof. Dr.-Ing. Michael Muma and the Signal Processing Group of Prof. Dr.-Ing. Abdelhak M. Zoubir at Technische Universität Darmstadt, who made my PhD journey very enjoyable: Taulant Koka, Fabian Scheidt, Martin Gölz, Christian Schroth, Mahmoud El-Hindi, Pertami Kunz, Christian Eckrich, Julian Weick, Dr.-Ing. Dominik Reinhard, Dr.-Ing. Aylin Taştan, Dr.-Ing. Huiping Huang, Dr.-Ing. Afief Dias Pambudi, Dr.-Ing. Sergey Sukhanov, Dr.-Ing. Ann-Kathrin Seifert, Dr.-Ing. Di Jin, Dr.-Ing. Tim Schäck, Dr.-Ing. Michael Faußß, Dr.-Ing. Sara Al-Sayed, Dr.-Ing. Christian Debes, Dr. Roy Howard, Andrea Gargano, and Pascal Zhang. I fondly remember our group gatherings and many conferences and workshops that we visited together all around the world.

Many thanks also go to the current and former members and visitors of the Convex Optimization Group of Prof. Dr. Daniel P. Palomar at HKUST, who made my stays in Hong Kong and at many international conferences very enjoyable, especially during our dinner parties and hiking trips: Dr. Xiwen Wang, Dr. Vinícius Cardoso, Dr. Rui Zhou, Dr. Jiaxi Ying, Dr. Junyan Liu, Prof. Dr. Sandeep Kumar, Amirhossein Javaheri, Runhao Shi, Shengjie Xiu, Yifan Yu, Prof. Dr. Gesualdo Scutari, Prof. Dr. Ziping Zhao, Dr. Licheng Zhao, Dr. Tianyu Qiu, Dr. Yiyong Feng, Miguel Cidra Senra, Roger Romero Morral, and Víctor Domínguez Cámara.

A heartfelt thank you goes to Sonja Höhn, Renate Koschella, and Rojin Ulusoy for their administrative support and to Hauke Fath for his technical support.

This dissertation would not have been possible without the unconditional love and support of my parents, Fatima and Lahcen, and my siblings, Rachid and Salma. Thank you, Sara, for your love and support.

Darmstadt, August 2024

Contents

1 Introduction **1**
- 1.1 Motivation and Goals . . . 1
- 1.2 State of the Art . . . 3
- 1.3 Contributions and Overview . . . 5

2 Fundamentals **9**
- 2.1 High-Dimensional Statistical Learning . . . 9
  - 2.1.1 Lasso . . . 11
  - 2.1.2 Model Selection Methods . . . 12
  - 2.1.3 LARS Algorithm . . . 14
  - 2.1.4 Elastic Net . . . 17
- 2.2 False Discovery Rate Control . . . 19
  - 2.2.1 FDR and TPR . . . 19
  - 2.2.2 The BH and the BY Method . . . 20
  - 2.2.3 The Fixed-X and the Model-X Knockoff Methods . . . 21
- 2.3 Martingale Theory . . . 23
  - 2.3.1 Definitions . . . 24
  - 2.3.2 Doob's Optional Stopping Theorem . . . 25
  - 2.3.3 FDR Control Proof of the BH Method Using Doob's Optional Stopping Theorem . . . 26

3 High-Dimensional False Discovery Rate Control **29**
- 3.1 The T-Rex Selector in a Nutshell . . . 30
  - 3.1.1 Key Ideas . . . 30
  - 3.1.2 Main Theoretical Results . . . 31
  - 3.1.3 Major Advantages Compared to Existing Methods . . . 31
- 3.2 Methodology . . . 32
- 3.3 Main Ingredients of the T-Rex Selector . . . 35
- 3.4 Problem Statement . . . 37
- 3.5 Main Results . . . 40
  - 3.5.1 FDR Control . . . 40
    - 3.5.1.1 Preliminary Definitions . . . 41
    - 3.5.1.2 The FDR Control Theorem . . . 43
  - 3.5.2 Dummy Generation . . . 45

3.5.3 Optimal Calibration Algorithm . . . 47
3.5.4 Extension to the Calibration Algorithm . . . 49
3.6 Numerical Experiments . . . 50
3.6.1 Setup . . . 52
3.6.2 Results . . . 52
3.7 The Screen-T-Rex Selector . . . 55
3.7.1 Methodology and Main Results . . . 56
3.7.1.1 Ordinary Screen-T-Rex Selector . . . 56
3.7.1.2 Confidence-Based Screen-T-Rex Selector . . . 58
3.7.1.3 Screening Genomics Biobanks . . . 59
3.7.2 Numerical Experiments . . . 59
3.8 Summary . . . 60

4 Dependency-Aware High-Dimensional False Discovery Rate Control 63
4.1 The T-Rex +DA Selector: Motivation and Major Contributions . . . 64
4.1.1 Motivation . . . 64
4.1.2 Major Contributions . . . 65
4.2 Methodology and Main Theoretical Results . . . 66
4.2.1 Preliminaries . . . 66
4.2.2 The Dependency-Aware T-Rex Selector . . . 67
4.2.3 Clustering Variables via Hierarchical Graphical Models . . . 68
4.2.4 Preliminaries for the FDR Control Theorem . . . 71
4.2.5 Dependency-Aware FDR Control . . . 73
4.2.6 General Group Design Principle . . . 73
4.2.7 Optimal Dependency-Aware T-Rex Calibration Algorithm . . . 74
4.3 Numerical Experiments . . . 76
4.3.1 Setup . . . 76
4.3.2 Results . . . 79
4.4 The T-Rex+DA Selector for Sparse Financial Index Tracking . . . 80
4.4.1 Adaptation for Overlapping Groups of Highly Correlated Variables 81
4.4.1.1 Group Design . . . 81
4.4.1.2 FDR Control . . . 82
4.4.2 Adaptation for Autoregressive Variable Dependency Models . . . 83
4.4.2.1 Group Design . . . 84
4.4.2.2 Theoretical Analysis . . . 85
4.4.2.3 Numerical Experiments . . . 86
4.5 Summary . . . 87

5 Joint Grouped Variable Selection and False Discovery Rate Control 89
5.1 The T-Rex+GVS Selector . . . 90
5.1.1 Motivation . . . 90

5.1.2 Major Contributions . . . 90
5.1.3 Grouped Variable Selection . . . 91
5.1.4 Dummy Generation . . . 92
5.2 The Informed Elastic Net for Grouped Variable Selection and FDR Control 93
5.2.1 Motivation . . . 94
5.2.2 The T-Rex+GVS Selector with the Informed Elastic Net: Overview of Major Contributions . . . 94
5.2.3 Methodology and Theoretical Analysis . . . 95
5.2.4 Numerical Experiments . . . 97
5.2.4.1 Grouping Effect and Solution Path . . . 97
5.2.4.2 Relative Computation Time . . . 98
5.3 Sparse PCA with FDR-Controlled Variable Selection . . . 99
5.3.1 Introduction to Sparse PCA . . . 100
5.3.2 Motivation for FDR-Controlled Sparse PCA . . . 101
5.3.3 The T-Rex PCA Method: Overview of Major Contributions . . . 102
5.3.4 Methodology . . . 102
5.3.4.1 T-Rex PCA Algorithm . . . 102
5.3.4.2 Percentage of Explained Variance . . . 104
5.3.5 Numerical Experiments . . . 105
5.4 Summary . . . 106

6 Applications in Biomedical Engineering **109**
6.1 Simulated Genome-Wide Association Study . . . 109
6.1.1 GWAS in a Nutshell . . . 110
6.1.2 Setup and Problem Statement . . . 110
6.1.3 Preprocessing . . . 112
6.1.4 Results . . . 113
6.2 HIV-1 Drug Resistance Analysis . . . 116
6.2.1 Problem Statement and Setup . . . 116
6.2.2 Results . . . 116
6.3 Breast Cancer Survival Analysis . . . 118
6.3.1 TCGA Breast Cancer Data . . . 118
6.3.2 Methods and Results . . . 119

7 Applications in Financial Engineering **121**
7.1 FDR-Controlled Sparse Index Tracking . . . 121
7.1.1 Problem Statement and Stock Returns Data Model . . . 122
7.1.2 Algorithm: FDR-Controlled Index Tracking . . . 123
7.1.3 Real-World S&P 500 Index Tracking . . . 124
7.1.3.1 T-Rex+DA+NN Selector: Setup and Results . . . 125
7.1.3.2 T-Rex+DA+AR1 Selector: Setup and Results . . . 127

7.2 Factor Analysis of S&P 500 Stock Returns . . . 127

8 Conclusion **131**
8.1 Summary . . . 131
8.2 Limitations and Open Challenges for Future Research . . . 134

Appendix A Appendix to Chapter 3 **139**
A.1 Proofs . . . 139
A.1.1 Preliminaries: Technical Corollaries and Lemmas . . . 140
A.1.2 Proof of Lemma 1 (Martingale) . . . 146
A.1.3 Proof of Theorem 4 (Dummy generation) . . . 148
A.1.4 Proof of Corollary 3 . . . 150
A.1.5 Proof of Theorem 5 (Optimality of Algorithm 2) . . . 151
A.2 The Deflated Relative Occurrence . . . 153
A.3 Hyperparameter Choices for the Extended Calibration Algorithm . . . 154
A.4 Computational Complexity . . . 155
A.5 General Assumptions . . . 156
A.6 Exemplary Numerical Verification of A-I, A-II, and A-III . . . 160
A.6.1 Exemplary Numerical Verification of A-I . . . 160
A.6.2 Exemplary Numerical Verification of A-II . . . 160
A.6.3 Exemplary Numerical Verification of A-III . . . 161
A.7 Verification of A-I, A-II, and A-III on HAPGEN2 Genomics Data . . . 162
A.8 Additional Simulation Results . . . 162
A.9 Illustration of Theorem 4 (Dummy Generation) . . . 164
A.10 Robustness of The T-Rex Selector . . . 165

Appendix B Appendix to Chapter 4 **171**
B.1 Proofs and Technical Lemmas . . . 171
B.1.1 Preliminaries for Theorem 8 . . . 171
B.1.2 Proof of Theorem 8 . . . 173
B.1.3 Proof of Theorem 9 . . . 174
B.1.4 Proof of Theorem 10 . . . 177
B.1.5 Proof of Theorem 11 . . . 177
B.1.6 Proof of Theorem 12 . . . 179
B.1.7 Technical Lemmas . . . 180
B.2 Additional Simulations . . . 183

Appendix C Appendix to Chapter 5 **187**
C.1 Proof of Theorem 14 . . . 187
C.2 Proof of Theorem 15 . . . 188
C.3 Proof of Theorem 16 . . . 189

List of Acronyms & Abbreviations 191

List of Operators & Symbols 195

List of Figures 199

List of Tables 203

References 205

To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science.

Albert Einstein

1 Introduction

In this introduction, Section 1.1 presents the motivation and goals of the dissertation. In Section 1.2, a brief overview of the state-of-the-art in high-dimensional variable selection is provided. Section 1.3 summarizes the contributions and provides an overview of the dissertation.

1.1 Motivation and Goals

This dissertation addresses a fundamental problem in modern machine learning and signal processing: the reproducible discovery of the few true active variables (e.g., signals, features) among up to millions of candidate variables. More specifically, determining the set of reproducible active signals or variables is a crucial task in, e.g., genomics research [Hufi8; Sud+15; Bun+19], financial engineering [HLZ15; BFP17; Pal24], detection [Chu+07; CZP20; CSY18], antenna array processing [TEN14], distributed learning [DS12], robust estimation [Zou+12; Zou+18; Mac+17; Mac+20; Yan+19], and many other areas (for an overview, see [FL06]). Any attempt to solve such variable selection problems usually leads to encountering the following four major challenges:

1. High-dimensional nature of the data: The number of samples is usually much smaller compared to the number of candidate variables, which leads to ill-posed problems and under-determined systems of equations [BV11].

2. High demand for computational resources: Storing and processing huge data sets (e.g., genomics data [Sud+15], gene expression data [TCW15]) using state-of-the-art methods [BC15; Can+18] is a challenging task, even for high-performance computers.

3. Requirement for statistical reproducibility guarantees: Without reproducibility guarantees, limited resources are wasted on costly follow-up studies in, e.g., biomedical research [GC18], or false discoveries are published in, e.g., financial economics [HLZ15].

4. Highly dependent candidate variables: The presence of strong and application-specific dependency structures such as groups of highly correlated candidate variables in, e.g., genomics [Bal06], gene expression [SDC03], and stock returns data (see Chapters 6 and 7) may lead to a breakdown of existing FDR-controlling methods.

The goal of this dissertation is to address these challenges by proposing fast machine learning algorithms for high-dimensional data that allow for the selection of reproducible variables. In this context, a reproducible variable or reproducible discovery is a variable that is not a false positive. To enable reproducible discoveries, it is essential that (i) the proportion of false discoveries among all selected variables is low, while (ii) the proportion of true discoveries among all true active variables is high. The expected values of these quantities are referred to as the false discovery rate (FDR) and the true positive rate (TPR), respectively. For example, without FDR control, expensive functional genomics studies and biological laboratory experiments are wasted on researching false positives [Cha+07; Vis+17; Huf18; GC18]. In addition, FDR control helps to achieve the goal of reproducibility in many scientific fields that suffer from false discoveries, such as biomedical research [Ioa05; 13; Bak16a; Mul21] and finance [HLZ15]. To account for the high-dimensional setting, where the number of variables exceeds the number of observations, it is imperative that the FDR control proofs rely on finite-sample (i.e., non-asymptotic) theory.

This dissertation contains the following major methodological, theoretical, and practical contributions (see Section 1.3 for a more detailed outline of the contributions):

1. The development of the Terminating-Random Experiments (*T-Rex*) selector, a new fast and flexible variable selection framework.

2. A theory for finite sample FDR control, which is rooted in martingale theory, and its numerical verification through extensive simulation studies on the Lichtenberg high-performance computer of the Technische Universität Darmstadt.

3. A demonstration of the applicability and efficiency of the proposed methods in biomedical applications (genome-wide association studies (GWAS), survival analysis using gene expression data, human immunodeficiency virus type 1 (HIV-1) drug resistance studies), and financial engineering (sparse financial index tracking, factor analysis of the S&P 500 stock returns).

4. The release of two open-source R software packages with an efficient C++ backend on the Comprehensive R Archive Network (CRAN), which contain the implementations of the proposed methods as well as demos and tutorials.

1.2 State of the Art

The fundamental task of selecting the subset of true active variables in high-dimensional data has attracted the interest of many scientists. Although the underlying selection problem is NP-hard [Nat95], various relaxations have been proposed to deal with the stringent condition of the best subset selection problem. The arguably most adopted approaches are sparsity inducing methods that solve least squares type optimization problems with a sparsity inducing constraint. The sparsity level (i.e., the number of selected variables) usually depends on a sparsity parameter that needs to be chosen by some type of parameter tuning or by applying rules-of-thumb. Some of the most used methods are *Lasso* [Tib96], adaptive *Lasso* [Zou06], and *elastic net* [ZHT06], which can be formulated as quadratic programs (QPs) and solved using off-the-shelf QP-solvers. However, there exist more efficient and widespread algorithms such as the *LARS* algorithm [Efr+04] and the pathwise coordinate descent method [Fri+07] to solve these QPs. The major drawback of sparsity inducing methods is that they require tuning the sparsity parameter, which is often done using model selection approaches such as cross-validation (CV) [Sto74; Sto77; All74], information criteria (e.g., Akaike information criterion (AIC) [Hir73] and Bayesian information criterion (BIC) [Sch78]), and other approaches (see [HTF09] for an overview and discussion). Unfortunately, these approaches do not control the FDR and in many cases may lead to differing models [Dzi+20] or select too many false positives [Abr+06].

Alternatively to combining sparsity inducing methods with model selection approaches, the umbrella term multiple testing refers to the simultaneous testing of many hypotheses [Sha95]. Frequently used multiple testing methods, which control the conservative family-wise error rate (FWER) (i.e., the probability of making one or more type I errors), are p-value based meth-

ods such as the Bonferroni correction [Abd+07], the Holm-Bonferroni method [Hol79], the Hochberg method [Hoc88], and others (see [Sha95] for an overview and a discussion). However, the FWER-metric is a conservative metric that may lead to many missed discoveries in variable selection tasks. This has lead to the development of the more liberal FDR-metric, which usually allows for more discoveries while controlling a user-defined target FDR [BH95; BY01].

Developing FDR-controlling methods for multiple testing in high-dimensional settings is challenging. While there exist established FDR-controlling methods for low-dimensional data, e.g., the Benjamini-Hochberg (*BH*) method [BH95], the Benjamini-Yekutieli (*BY*) method [BY01], and the *fixed-X* knockoff method [BC15], there exist not many and, especially, no computationally fast multivariate approaches for large-scale high-dimensional data. One could resort to single hypothesis testing using marginal p-values. However, this approach to hypothesis testing in high-dimensional settings raises many issues, which concern the validity of marginal p-values in high-dimensional settings, the loss of power due to marginal testing, and many others [Hog+08; Can+18]. Therefore, in recent years, the multivariate *model-X* knockoff method [Can+18] and derandomized versions thereof [RWC21; RB24] have been proposed for the high-dimensional setting. However, they are computationally demanding. In fact, creating knockoff predictors that mimic the covariance structure of the original predictors renders them infeasible for settings beyond a few thousand variables (see Figure 3.1). Moreover, the original derandomized knockoffs approach controls the conservative per family error rate (*PFER*) and the k-family-wise error rate (k-*FWER*) but does not consider the less conservative FDR metric [RWC21]. Only the derandomized approach based on e-values controls the FDR [RB24]. Nevertheless, the need for running the *model-X* knockoff method multiple times renders both derandomized knockoffs approaches practically infeasible for large-scale high-dimensional settings. Alternative FDR-controlling approaches that rely on conditional randomization test (*CRT*) p-values [Can+18] are computationally significantly more demanding than the *model-X* knockoff methods, which renders them infeasible in relatively small settings (see [Can+18] for a discussion).

In recent years, deep learning approaches for FDR control, such as *DeepPink* [Lu+18] and *SurvNet* [SL21], have been proposed. However, the black box nature, the need for large amounts of data to train deep neural networks, or strong assumptions on the feature distribution limit the applicability of existing approaches in real-world large-scale high-dimensional applications where provable finite sample FDR control guarantees are required for trustwor-

thy methods and reproducible discoveries. This is especially important, since cases have been reported in which these methods did not control the FDR [SL21].

Other related lines of research on error-controlled high-dimensional variable selection are centered around stability selection methods [MB10; SS13], data-splitting methods [Cox75; WR09; MMB09; BC19], and post-selection inference [Loc+14; FST14; Lee+16; Tib+16].

1.3 Contributions and Overview

This dissertation addresses the need for fast, scalable, and tractable machine learning approaches with provable FDR control guarantees and demonstrable efficacy and efficiency in real-world data biomedical and financial applications, as outlined in this section.

First, **Chapter 2** establishes and links together the theoretical and methodological foundations of this dissertation, which lie in high-dimensional statistical learning, FDR control, and martingale theory.

In **Chapter 3**, the first main contribution of this dissertation, the Terminating-Random Experiments (*T-Rex*) selector, a fast variable selection framework for high-dimensional data, is presented. The *T-Rex* selector controls a user-defined target FDR while maximizing the number of selected variables. This is achieved by fusing the solutions of multiple early terminated random experiments. The experiments are conducted on a combination of the candidate variables and multiple sets of randomly generated dummy variables. A finite sample proof of the FDR control property is provided using martingale theory. The *T-Rex* selector outperforms state-of-the-art methods for FDR control in numerical experiments and on a simulated genome-wide association study (GWAS), while its sequential computation time is more than two orders of magnitude lower than that of the strongest benchmark methods. Additionally, for screening large-scale genomics biobanks with tens of thousands of phenotypes (e.g., diseases, traits) and millions of single nucleotide polymorphisms (SNPs) for reproducible associations, the *Screen-T-Rex* selector, a fast FDR-controlling method based on the developed *T-Rex* selector, is proposed. The computation time of the *Screen-T-Rex* selector is about an order of magnitude lower than that of the ordinary *T-Rex* selector and it does not require choosing a target FDR level but provides the user with an estimate of the achieved FDR.

Chapter 4 contains the second main contribution of this dissertation, i.e., it proposes a class of dependency-aware FDR-controlling algorithms for large-scale high-dimensional data. In

many biomedical and financial applications, the high-dimensional data sets are characterized by highly correlated candidate variables (e.g., gene expression data and stock returns). For such applications, the dependency-aware *T-Rex* (*T-Rex+DA*) framework has been developed. It extends the ordinary *T-Rex* framework by accounting for dependency structures among the candidate variables. This is achieved by integrating hierarchical graphical models within the *T-Rex* framework, which allows to effectively harness the dependency structure among variables and to develop variable penalization mechanisms that guarantee FDR control. Using martingale theory, the FDR control property has been established for the proposed approach. The *T-Rex+DA* framework has been further generalized by stating and proving a condition that needs to be satisfied by both graphical and non-graphical dependency-capturing models. This allows for specifying the framework for various application-specific dependency models. In this dissertation, the *T-Rex+DA* framework has been specified for all dependency models that are relevant for the considered applications in biomedical and financial engineering, i.e., disjoint groups of highly correlated variables, overlapping groups of highly correlated variables, and autoregressive dependencies.

Chapter 5, which presents the third main contribution, proposes theory and algorithms for joint grouped variable selection and FDR control. This approach to tackling the challenges arising from the presence of groups of highly dependent variables in the data is different from the more conservative variable penalization approach that has been developed in Chapter 4. That is, instead of finding the few true active variables among groups of highly correlated variables, the goal is to select all groups of highly correlated variables that contain at least one true active variable. In genomics research, especially for GWAS, grouped variable selection approaches are highly relevant, since one is not interested in identifying a few SNPs that are associated with a disease of interest but rather the entire groups of correlated SNPs that point to relevant locations on the genome. Therefore, the *T-Rex* selector for grouped variable selection (*T-Rex+GVS*), a version of the *T-Rex* selector that uses the *elastic net (EN)* as a base selector to perform grouped variable election, is proposed. To further reduce the computation time and, thus, increase the scalability of the *T-Rex+GVS* selector, the *EN* base selector has been replaced by the proposed *informed elastic net (IEN)*, a new base selector that significantly reduces the computation time while retaining the grouped variable selection property. Leveraging the developed FDR-controlling grouped variable selection algorithms and the fact that sparse principal component analysis (PCA) can be cast as a series of *EN* optimization problems, an alternative formulation of sparse PCA driven by the FDR has been proposed.

This formulation allows to use the *T-Rex+GVS* selector to automatically determine an FDR-controlled support of the loading vectors. A major advantage of the resulting *T-Rex* PCA is that compared to ordinary sparse PCA no sparsity parameter tuning is required.

Chapters 6 and 7, the fourth main part of this dissertation, contain the practical contributions. The considered applications lie in two different areas: Biomedical engineering (Chapter 6) and financial engineering (Chapter 7). The biomedical applications include (i) a semi-real-world GWAS, (ii) a human immunodeficiency virus type 1 (HIV-1) data set with associated drug resistance measurements, and (iii) a breast cancer data set with associated survival times of the patients. The financial engineering applications include (i) accurately tracking the S&P 500 index using a quarterly updated and rebalanced tracking portfolio that consists of few stocks and (ii) a factor analysis of S&P 500 stock returns. The common challenge of all considered applications lies in detecting the few true active variables (i.e., SNPs, mutations, genes, stocks) among many non-active variables in, among other things, large-scale high-dimensional settings that consist of up to tens of thousands of variables and much fewer samples.

Finally, **Chapter 8** concludes this dissertation and provides interesting open problems for future theoretical and practical research and software development projects.

Facts are the air of scientists. Without them you can never fly.

Linus Pauling

2 Fundamentals

This chapter introduces the fundamentals of high-dimensional statistical learning, false discovery rate (FDR) control, and martingale theory, which are essential throughout this dissertation. Section 2.1 describes the high-dimensional learning problem and introduces methods that are often used to solve it. Section 2.2 provides the mathematical definitions of the FDR and the true positive rate (TPR) and discusses classical as well as recent FDR-controlling methods. Section 2.3 introduces the fundamentals of martingale theory, which is the foundation of many finite sample FDR control proofs.

2.1 High-Dimensional Statistical Learning

A supervised statistical learning problem is said to be high-dimensional if the number of variables p exceeds the number of samples n [BV11]. One of the most prominent and well-studied problems in high-dimensional learning arises in the context of linear regression. It considers the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{X} \coloneqq [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_p]$ with $\boldsymbol{x}_j \coloneqq [x_{1j}\ x_{2j}\ \cdots\ x_{nj}]^\top \in \mathbb{R}^n$, $j = 1, \ldots, p$, is the fixed predictor matrix containing p predictors and n observations, $\boldsymbol{y} \coloneqq [y_1\ y_2\ \cdots\ y_n]^\top \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\beta} \coloneqq [\beta_1\ \beta_2\ \cdots\ \beta_p]^\top \in \mathbb{R}^p$ is the parameter vector, and $\boldsymbol{\epsilon} \coloneqq$

$[\epsilon_1\ \epsilon_2\ \cdots\ \epsilon_n]^\top \in \mathbb{R}^n$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, with $\boldsymbol{I}$ being the identity matrix, is an additive Gaussian noise vector with standard deviation σ. In the low-dimensional setting, where $n \leq p$, and when $\boldsymbol{X}$ is full rank, the unique closed-form solution of the linear regression problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\, \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \tag{2.2}$$

i.e., the minimization of the sum of squared residuals (SSR), is given by the ordinary least squares (*OLS*) estimator

$$\hat{\boldsymbol{\beta}} \coloneqq (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.3}$$

However, this solution cannot even be computed in the high-dimensional setting ($p > n$) because the ranks of $\boldsymbol{X}^\top$ and $\boldsymbol{X}$ are not greater than n, i.e.,

$$\text{rank}(\boldsymbol{X}^\top) = \text{rank}(\boldsymbol{X}) \leq \min\{n, p\} = n \tag{2.4}$$

and, therefore, the $(p \times p)$-matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is rank deficient and not invertible, i.e.,

$$\text{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = \min\{\text{rank}(\boldsymbol{X}^\top), \text{rank}(\boldsymbol{X})\} \leq n < p. \tag{2.5}$$

Furthermore, the least squares problem in the high-dimensional setting is ill-posed because the optimization problem in (2.2) does not have a unique solution, as it does in the low-dimensional setting, but instead has infinitely many solutions. Moreover, the least squares optimization problem generally does not allow for sparse solutions, i.e., solutions with zero entries in $\hat{\boldsymbol{\beta}}$. Therefore, it does not allow to recover the support of $\boldsymbol{\beta}$ when the true underlying model is sparse (i.e., zero entries in $\boldsymbol{\beta}$). In the remainder of this dissertation, variables whose associated coefficients in $\boldsymbol{\beta}$ are non-zero (zero) are referred to as actives or active variables (nulls or null variables). The number of actives is denoted by p_1, while the number of nulls is denoted by p_0.

The aim in sparse high-dimensional statistical learning is to minimize the SSR subject to a constraint on the support of $\boldsymbol{\beta}$, i.e.,

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\, \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq t, \tag{2.6}$$

where $\|\boldsymbol{\beta}\|_0 \coloneqq |\{\beta_j \in \{1, \ldots, p\} : \beta_j \neq 0\}|$, with $|\cdot|$ denoting the cardinality operator, is the ℓ_0-"norm" of the coefficient vector $\boldsymbol{\beta}$ and $t > 0$ is a tuning parameter that controls the

sparsity of the solution vector $\hat{\boldsymbol{\beta}}$. Note that the ℓ_0-"norm" is not a norm in the mathematical sense, but it has been used to denote the number of non-zero entries of a vector $\boldsymbol{\beta}$. Unfortunately, the ℓ_0-"norm" is non-convex and, therefore, the optimization problem in (2.6) is a non-convex optimization problem that cannot be easily solved using an off-the-shelf solver. In fact, (2.6) is an NP-hard problem [Nat95].

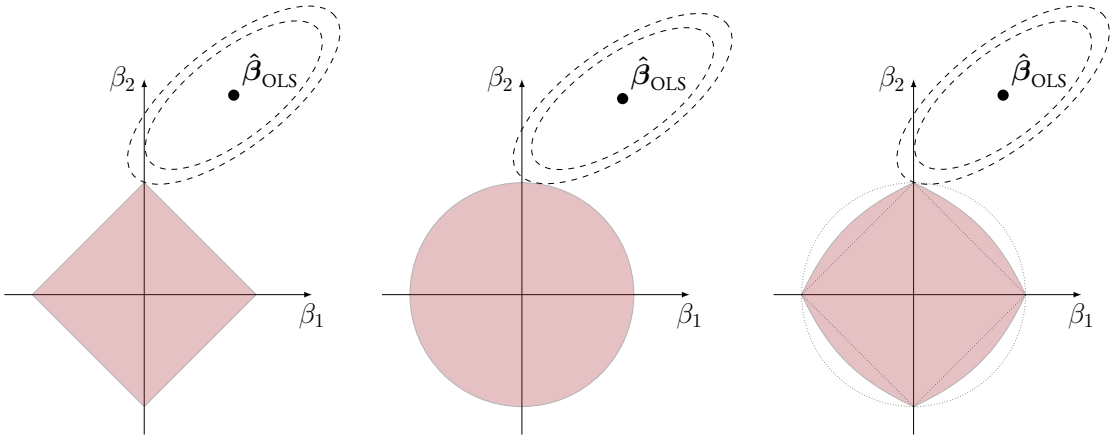
In order to alleviate the above described shortcomings of ordinary linear regression and the NP-hard optimization problem in (2.6), various regularized regression methods, which relax the ℓ_0-"norm" constraint in (2.6), have been proposed in the last few decades. The following commonly used regularized high-dimensional regression methods are introduced and discussed in the upcoming sections: *Lasso* [Tib96], *LARS* [Efr+04], and *elastic net* [ZH05]. These methods will be relevant throughout this dissertation as building blocks of the proposed FDR-controlling methods. Other closely related regularized regression methods are the fused *Lasso* [Tib+05], group *Lasso* [YL06], adaptive *Lasso* [Zou06], and regularized Cox's proportional hazards model [Tib97; Sim+11]. For more details on high-dimensional variable selection and regularized regression methods, the interested reader is referred to standard textbooks [HTF09; BV11; HTW15].

2.1.1 Lasso

The least absolute shrinkage and selection operator (*Lasso*) [Tib96] is often used to obtain a sparse estimate $\hat{\boldsymbol{\beta}}$ of the true coefficient vector $\boldsymbol{\beta}$ in the high-dimensional setting. It minimizes the SSR subject to a budget constraint on the ℓ_1-norm of $\boldsymbol{\beta}$, i.e.,

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\, \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_1 \leq t, \tag{2.7}$$

where $t > 0$ is the tuning parameter that controls the sparsity of $\hat{\boldsymbol{\beta}}$. That is, the non-convex ℓ_0-norm constraint in (2.6) is replaced by the convex ℓ_1-norm constraint and, thus, (2.7) is a convex quadratic program (QP). The *Lasso* in (2.7) yields sparse estimates of $\boldsymbol{\beta}$, i.e., it performs variable selection. Loosely speaking, the reason why the *Lasso* promotes sparse estimates of $\boldsymbol{\beta}$ is that the feasible set has "sharp" corners. This is illustrated in Figure 2.1a, where the dense (i.e., non-sparse) *OLS* estimate appears to be shrunk towards a "sharp" corner of the feasible set. For more details on the properties of the *Lasso*, the interested reader is referred to [BV11].



(a) Lasso **(b)** Ridge regression **(c)** Elastic net

Figure 2.1: Illustration of the (a) *Lasso*, (b) ridge regression, and (c) *elastic net* optimization problems for two variables with associated coefficients β_1 and β_2: The *OLS* solution $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, which is obtained by minimizing the sum of squared residuals (SSR) (contour lines of the SSR are indicated by the dashed lines around $\hat{\boldsymbol{\beta}}_{\text{OLS}}$), is shrinked towards the feasible sets (i.e., the areas in red) that are determined by the penalty terms of the three regularized estimators. Loosely speaking, due to the "sharp corners" of their feasible sets, the *Lasso* and the *elastic net* allow for sparse solutions, in which the estimate of β_1 is exactly zero while ridge regression does not allow for such sparse solutions.

The Lagrangian formulation of the constraint *Lasso* optimization problem in (2.7) is given by

$$\hat{\boldsymbol{\beta}}(\lambda_1) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1, \tag{2.8}$$

where $\lambda_1 > 0$ is the tuning parameter that controls the sparsity of $\hat{\boldsymbol{\beta}}$. Note that there exists a one-to-one relationship between t in the constraint formulation and λ_1 in the Lagrangian formulation of the *Lasso*. For both formulations of the *Lasso*, the choices of the sparsity tuning parameters t and λ_1 are crucial because they determine how many variables are included in the final model.

2.1.2 Model Selection Methods

In the following, since there exists a one-to-one relationship between t and λ_1, we introduce existing model selection approaches for choosing the sparsity parameter only from the perspective of λ_1. Some of the most widely used model selection methods are M-fold cross validation (M-fold CV) [HTF09], the Akaike information criterion (AIC) [Aka98] and the Bayesian information criterion (BIC) [Sch78].

In M-fold CV, the data set is divided into M disjoint sets of approximately the same size, where M is usually set to 5 or 10. Each of these sets is used once for validation while the others are merged and used as training data for the *Lasso* problem in (2.8). Thus, we need to solve M *Lasso* problems given by

$$\hat{\boldsymbol{\beta}}_{-m}(\lambda_1) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y}_{-m} - \boldsymbol{X}_{-m}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1, \quad m = 1, \ldots, M, \tag{2.9}$$

where the subscript $-m$ in $\boldsymbol{y}_{-m}$, $\boldsymbol{X}_{-m}$ and $\hat{\boldsymbol{\beta}}_{-m}$ refers to the response and predictor matrix without the data belonging to the mth validation set and the corresponding estimator of the parameter vector, respectively. Based on (2.9), the CV-error (CVE) curve can be computed by

$$\mathrm{CVE}(\lambda_1) \coloneqq \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{y}_m - \boldsymbol{X}_m\hat{\boldsymbol{\beta}}_{-m}(\lambda_1)\|_2^2, \tag{2.10}$$

where $\boldsymbol{y}_m$ and $\boldsymbol{X}_m$ are the response and predictor matrix containing only the mth set of validation data. The sparsity parameter λ_1 is chosen by minimizing (2.10) with respect to λ_1. Generally, the obtained λ_1-value leads to a model with too many active variables, which is not desired in sparse regression. Therefore, the one-standard-error rule [HTF09], a rule of thumb, is often used to select a sparser model. It replaces the value of λ_1 that minimizes (2.10) by a larger value that corresponds to the CVE-value that deviates by one standard error from the minimum value of (2.10).

In contrast to M-fold CV, the AIC and BIC minimize a different objective function that takes into account the SSR and the model size. For the *Lasso*, the objective functions of both information criteria are defined by

$$\mathrm{IC}(\lambda_1) \coloneqq \frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_1)\|_2^2}{n\sigma^2} + \frac{c_n}{n} \cdot \mathrm{df}(\lambda_1), \tag{2.11}$$

where σ^2 is the noise variance in (2.1), which is usually estimated based on the largest model, $c_n = 2$ for the AIC, $c_n = \ln(n)$ for the BIC and $\mathrm{df}(\lambda_1)$ is the number of degrees of freedom as a function of the sparsity parameter, which is equivalent to the number of selected variables [ZHT07]. The BIC usually selects a sparser model than the AIC. The reason for this is that $\ln(n) > 2$ for $n \geq 8$, that is, the BIC penalizes larger models stronger than the AIC when there are at least eight observations in the data set, which is usually the case.

2.1.3 LARS Algorithm

Algorithm 1 *LARS* algorithm.

1. **Input**: $\boldsymbol{X}, \boldsymbol{y}$ ($\boldsymbol{X}$ standardized, $\boldsymbol{y}$ centered).
2. **Initialize** the solution vector $\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$, the selected active set $\widehat{\mathcal{A}}_0 = \varnothing$, the current residual vector $\hat{\boldsymbol{r}} = \boldsymbol{y}$, the current correlation vector $\hat{\boldsymbol{c}} = [\hat{c}_1 \ \hat{c}_2 \ \cdots \hat{c}_p]^\top = \boldsymbol{X}^\top \hat{\boldsymbol{r}}$, and the iteration number $\kappa = 1$.
3. **While** $\kappa \leq \min\{n, p\}$ **do**:
   - 3.1. **Determine** the predictor that is not contained in the selected active set $\widehat{\mathcal{A}}_{\kappa-1}$ and has the highest correlation with the current residual and **add** it to the selected active set, i.e.,
   $$\widehat{\mathcal{A}}_\kappa = \widehat{\mathcal{A}}_{\kappa-1} \cup \left\{ \underset{j \notin \widehat{\mathcal{A}}_{\kappa-1}}{\arg\max} |\hat{c}_j| \right\}. \tag{2.12}$$
   - 3.2. **Compute** the least squares direction vector
   $$\boldsymbol{d} = \left(\boldsymbol{X}_{\widehat{\mathcal{A}}_\kappa}^\top \boldsymbol{X}_{\widehat{\mathcal{A}}_\kappa} \right)^{-1} \boldsymbol{X}_{\widehat{\mathcal{A}}_\kappa}^\top \hat{\boldsymbol{r}}, \tag{2.13}$$
   where $\boldsymbol{X}_{\widehat{A}_\kappa}$ contains only the predictors in $\widehat{A}_\kappa$.
   - 3.3. **Compute** the exact step size γ that is required to reach the next change point at which another variable has the same correlation with the current residual and, therefore, enters the solution path as detailed in [Efr+04].
   - 3.4. **Update** the solution vector, current residual vector, and current correlation vector:
   $$\hat{\boldsymbol{\beta}}_\kappa^{(\mathcal{A}_\kappa)} = \hat{\boldsymbol{\beta}}_{\kappa-1}^{(\widehat{\mathcal{A}}_{\kappa-1})} + \gamma \boldsymbol{d}, \tag{2.14}$$
   $$\hat{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\kappa, \tag{2.15}$$
   $$\hat{\boldsymbol{c}} = \boldsymbol{X}^\top \hat{\boldsymbol{r}}. \tag{2.16}$$
   The superscript $(\widehat{\mathcal{A}}_\kappa)$ indicates that only the entries corresponding to the index set $\widehat{\mathcal{A}}_\kappa$ are updated, while all other entries in $\widehat{\boldsymbol{\beta}}_\kappa$ remain zero.
   - 3.5. **Set** $\kappa \leftarrow \kappa + 1$.
4. **Output**: $\widehat{\mathcal{A}}_\kappa, \hat{\boldsymbol{\beta}}_\kappa, \kappa = 1, \ldots, \min\{n, p\}$.

Since the *Lasso* optimization problem in (2.7) (or (2.8)) is a QP, it can be solved using an off-the-shelf QP-solver. This approach is effective to compute the solution for a single value of λ_1, but it is not efficient to compute the entire solution path for all values of λ_1. However,

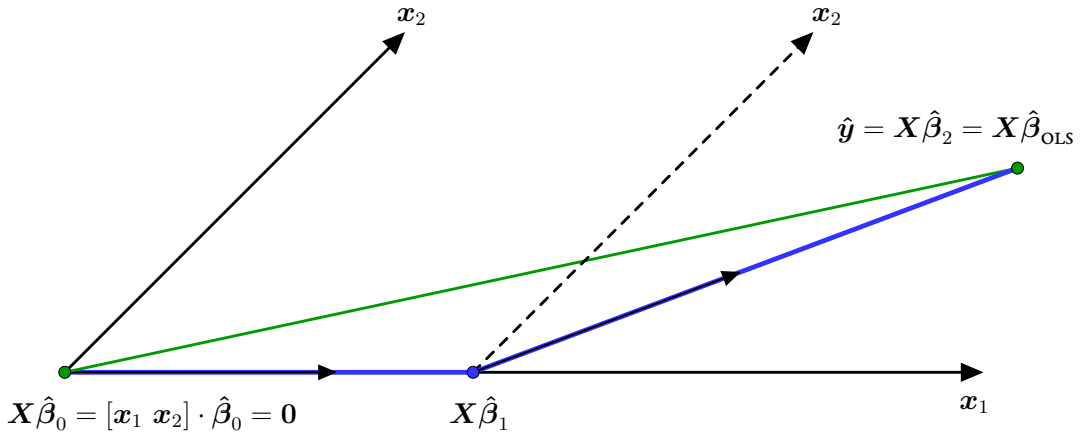


Figure 2.2: Illustration of the *LARS* algorithm as in [Efr+04] for two predictors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$: The correlations with the current residual are represented as angles between the current residual vector and the predictors. In the first iteration, $\boldsymbol{x}_1$ is selected because it has the smallest angle ("least angle") with the current residual vector. Its coefficient is increased along the least squares direction vector, which is simply the direction of $\boldsymbol{x}_1$ in the first iteration. As soon as the angles of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with the current residual become equal, $\boldsymbol{x}_2$ is selected and the coefficients of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are both moved along the least squares direction vector until the *OLS* solution is reached.

the entire solution path is required to perform model selection as described in Section 2.1.2, i.e., to determine the λ_1-value that minimizes $\text{CVE}(\lambda_1)$ in (2.10) or $\text{IC}(\lambda_1)$ in (2.11). A suitable algorithm to efficiently compute the entire solution path of the *Lasso* is the least angle regression (*LARS*) algorithm [Efr+04]. As detailed in the simplified pseudo-code in Algorithm 1 and illustrated in Figure 2.2 for two predictors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the *LARS* algorithm is a forward variable selection procedure that selects one variable in each iteration. It takes as inputs the standardized predictor matrix $\boldsymbol{X}$ and the centered response vector $\boldsymbol{y}$. That is, the means of the predictors and the response are equal to zero and the variances of the predictors are equal to one, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0, \quad \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \ldots, p. \tag{2.17}$$

In each iteration, the variable with the highest correlation with the current residual is selected and added to the active set. The solution vector $\hat{\boldsymbol{\beta}}_\kappa$ in the κth iteration of the algorithm is updated by linearly moving the coefficients of all selected active variables along the least squares direction vector up to the change point at which another variable has the same correlation with the current residual and, therefore, enters the solution path. The *LARS* algorithm com-

putes the exact step size along the least squares direction vector to arrive at the next change point, which corresponds to decreasing λ_1 in (2.9) to a certain value that allows for exactly one more variable to enter the solution path. This procedure is continued until $\min\{n, p\}$ variables have been selected.

In order to obtain the entire solution path for any value of λ_1, it suffices to compute and store the selected active sets and the solution vectors (i.e., $\widehat{\mathcal{A}}_\kappa, \hat{\boldsymbol{\beta}}_\kappa, \kappa = 1, \ldots, \min\{n, p\}$) at the change points where a new variable enters the solution path. The reason for this is that the solution path of the *LARS* algorithm is piecewise linear [OPT00a; OPT00b; Efr+04]. That is, the solution vector between the λ_1-values corresponding to any two consecutive solution vectors $\hat{\boldsymbol{\beta}}_\kappa$ and $\hat{\boldsymbol{\beta}}_{\kappa+1}$ lies on the line segment that connects $\hat{\boldsymbol{\beta}}_\kappa$ and $\hat{\boldsymbol{\beta}}_{\kappa+1}$.

The name "least angle regression" is inspired by an appealing geometrical interpretation of the *LARS* algorithm. The geometrical illustration of the *LARS* algorithm in Figure 2.2 represents correlations with the current residual as angles between the current residual vector and the predictors. In this illustrating example, only two predictors (i.e., $p = 2$) are considered. In the first iteration, $\boldsymbol{x}_1$ is selected because it has the smallest angle ("least angle") with the current residual vector. Its coefficient is increased along the least squares direction vector, which is simply the direction of $\boldsymbol{x}_1$ in the first iteration. As soon as the angles of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with the current residual become equal, $\boldsymbol{x}_2$ is selected and the coefficients of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are both moved along the least squares direction vector until the *OLS* solution is reached.

In fact, the ordinary *LARS* algorithm in Algorithm 1 requires a slight modification to efficiently solve the *Lasso* optimization problem in (2.8). The modification is that instead of adding one variable at a time based on the highest correlation with the current residual, the *Lasso* modification requires the removal of previously added variables when the associated coefficients change their sign [Efr+04]. However, removed variables can enter the solution path again in later steps.

The computational complexity of the *LARS* algorithm is $\mathcal{O}(p^2 n)$ in low-dimensional settings (i.e., $p \leq n$), which is the same as that of *OLS*, and $\mathcal{O}(p^3)$ in high-dimensional settings (i.e., $p > n$).

An alternative approach to obtain the solution path of the *Lasso* is to apply the pathwise coordinate descent method [Fri+07]. However, the proposed algorithms in this dissertation rely on forward variable selection methods and the pathwise coordinate descent method is not of that type. Therefore, we omit a detailed description of it and refer the interested reader

to [Fri+07; HTF09].

2.1.4 Elastic Net

The *elastic net* [ZH05] is a sparsity promoting variable selection and estimation method that addresses some well-known drawbacks of the *Lasso* [ZH05]:

1. The *Lasso* cannot select more than $\min\{n, p\}$ variables as indicated in Step 3 of Algorithm 1. This can be an issue in high-dimensional settings, where the number of true active variables p_1 is larger than the sample size n. In such scenarios, the *Lasso* is incapable of selecting all true active variables.

2. In the presence of groups of highly correlated variables in the data, the variable selection performance of the *Lasso* deteriorates. More specifically, instead of selecting entire active groups of highly correlated variables and increasing their coefficients in a correlated fashion, the *Lasso* is generally not capable of capturing such dependency structures. Instead, the *Lasso* tends to produce solutions in which only a single variable of a highly correlated group of variables is selected.

The *elastic net* alleviates these issues by combining the *Lasso* with ridge regression (also called Tikhonov regularization) [HK70; TA77]. Ridge regression minimizes the SSR plus an ℓ_2-norm penalty on the coefficient vector, i.e.,

$$\hat{\boldsymbol{\beta}}(\lambda_2) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2, \tag{2.18}$$

where $\lambda_2 > 0$ is a tuning parameter. The closed form solution of (2.18) is given by

$$\hat{\boldsymbol{\beta}}(\lambda_2) := (\boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{2.19}$$

where $\boldsymbol{I}$ is a $(p \times p)$ identity matrix. Note that (2.19) is very similar to the *OLS* solution in (2.3). The only difference is that a scaled identity matrix is added to $\boldsymbol{X}^\top \boldsymbol{X}$. However, this seemingly minor difference gives rise to the following desirable properties and advantages of ridge regression over *OLS* regression:

1. In contrast to *OLS* regression, ridge regression is feasible in high-dimensional settings. The reason for this is that $\boldsymbol{X}^\top \boldsymbol{X}$ in (2.3) is rank-deficient in high-dimensional settings and, therefore, it's inverse does not exist, while $\boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \boldsymbol{I}$ in (2.19) is not rank defi-

cient thanks to the addition of λ_2 to all diagonal elements of $\boldsymbol{X}^\top \boldsymbol{X}$ [HK70; TA77].

2. In contrast to *OLS* regression, ridge regression shrinks the coefficients towards zero and, thereby, achieves a reduced variance that comes at the cost of only a slightly increased bias of the ridge estimator. This desirable property leads to more stable coefficient estimates with lower variance and a better prediction accuracy, especially in the presence of multicollinearity (i.e., correlated predictors).

However, as illustrated in Figure 2.1b, the feasible set of the ridge regression optimization problem is a p-dimensional ball (i.e., 2-dimensional disc in Figure 2.1b) that has no "sharp" corners. Therefore, ridge regression generally does not shrink any coefficient to exactly zero. Thus, unlike the *Lasso*, ridge regression does not perform variable selection.

The *elastic net* combines the ℓ_2-norm and the ℓ_1-norm penalties of *Lasso* and ridge regression, i.e.,

$$\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1. \tag{2.20}$$

Thereby, the *elastic net* combines the advantages and remedies the drawbacks of ridge regression and the *Lasso*. Figure 2.1c illustrates the *elastic net* optimization problem. It can be observed that the shape of the feasible set of the *elastic net* is a mixture of the feasible sets of ridge regression and the *Lasso*. The "round" edges of ridge regression and the "sharp" corners of the *Lasso* are preserved, which leads to the following desirable properties of the *elastic net* [ZH05]:

1. In contrast to ridge regression, the *elastic net* performs variable selection by shrinking many coefficients to exactly zero.

2. In contrast to the *Lasso* that cannot select more than $\min\{n, p\}$ variables, the *elastic net* has the capability of selecting all p variables in high-dimensional settings.

3. In contrast to the *Lasso*, the *elastic net* is capable of selecting entire groups of highly correlated variables and increasing/decreasing their coefficients in a correlated fashion along the solution path.

The *elastic net* solution path can also be obtained using the *LARS* algorithm. This is achieved by merging the SSR and the ℓ_2-norm penalty term in the following reformulation of (2.20) [ZH05]:

$$\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y}' - \boldsymbol{X}'\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1. \tag{2.21}$$

The augmented predictor matrix $\boldsymbol{X}' \in \mathbb{R}^{(n+p)\times p}$ and the augmented response vector $\boldsymbol{y}'$ are defined by

$$\boldsymbol{X}' := \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}_{\boldsymbol{p}\times\boldsymbol{p}} \end{pmatrix}, \qquad \boldsymbol{y}' := \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0}_p \end{pmatrix}, \tag{2.22}$$

where $\boldsymbol{I}_{p\times p}$ and $\boldsymbol{0}_p$ are the $(p \times p)$ identity matrix and the p-dimensional zero vector. Since (2.21) is simply the *Lasso* problem in (2.8) with augmented data, the *elastic net* solution path can be obtained by inputting the augmented data $\boldsymbol{X}'$ and $\boldsymbol{y}'$ into the *LARS* algorithm.

In [ZH05], it is suggested to choose the ridge penalty parameter λ_2 and the sparsity parameter λ_1 by performing 10-fold CV as described in Section 2.1.2 but not only on λ_1 but on a two-dimensional grid of λ_1 and λ_2. More specifically, it is suggested to generate a small grid for λ_2 (e.g., $0, 0.01, 0.1, 1, 10, 100$) and to perform 10-fold CV for every fixed value of λ_2 and to choose the λ_1, λ_2 combination that yields the minimum CVE on the two-dimensional grid.

2.2 False Discovery Rate Control

In this section, the terms false discovery rate (FDR) and true positive rate (TPR) are mathematically defined. Then, existing and often used FDR-controlling variable selection methods are introduced. For the low-dimensional settings (i.e., $p \leq n$), there exist the Benjamini-Hochberg (*BH*) method [BH95], the Benjamini-Yekutieli (*BY*) method [BY01], and the *fixed-X* knockoff+ method [BC15], while the *model-X* knockoff+ method [Can+18] is applicable in high-dimensional settings (i.e., $p > n$). These methods will be used as benchmark methods throughout this dissertation.

2.2.1 FDR and TPR

The FDR and TPR are expressed mathematically as follows: Given the index set of the active variables $\mathcal{A} \subseteq \{1, \ldots, p\}$, where p is the number of candidate variables, and the index set of the selected active variables $\widehat{\mathcal{A}} \subseteq \{1, \ldots, p\}$, the FDR and the TPR are defined by

$$\mathrm{FDR} := \mathbb{E}\big[\,\mathrm{FDP}\,\big] := \mathbb{E}\left[\frac{|\widehat{\mathcal{A}}\backslash\mathcal{A}|}{1 \vee |\widehat{\mathcal{A}}|}\right] \tag{2.23}$$

and

$$\mathrm{TPR} := \mathbb{E}\big[\,\mathrm{TPP}\,\big] := \mathbb{E}\left[\frac{|\mathcal{A} \cap \widehat{\mathcal{A}}|}{1 \vee |\mathcal{A}|}\right], \tag{2.24}$$

respectively, where $\mathbb{E}[\cdot]$ is the expectation operator, $|\cdot|$ denotes the cardinality operator, and the symbol $\vee$ stands for the maximum operator, i.e., $a \vee b = \max\{a, b\}$, $a, b \in \mathbb{R}$. In words, the FDR is the expected value of the false discovery proportion (FDP), i.e., the expected percentage of false discoveries among all discoveries and the TPR is the expected value of the true positive proportion (TPP), i.e., the expected percentage of true discoveries among all true active variables. Note that by definition the FDR and TPR are zero when $|\widehat{\mathcal{A}}| = 0$ and $|\mathcal{A}| = 0$, respectively.[1]

While the FDR and the TPR of an oracle variable selection procedure are 0% and 100%, respectively, in practice, a tradeoff must be accomplished. In fact, existing FDR-controlling methods allow the user to set a target FDR value α and then select variables such that the FDR is controlled at the target level, i.e.,

$$\text{FDR} \leq \alpha, \quad \alpha \in [0, 1], \tag{2.25}$$

while maximizing the number of selected variables and, thus, implicitly maximizing the TPR.

2.2.2 The BH and the BY Method

The *BH* method [BH95] and the *BY* method [BY01] are often used FDR-controlling multiple hypothesis testing methods for the low-dimensional setting. Both methods consider the null hypotheses

$$H_j : \beta_j = 0, \quad j = 1, \ldots, p, \tag{2.26}$$

with associated p-values $P_1, \ldots, P_p$. A variable is considered to be selected if the corresponding null hypothesis is rejected. More specifically, for all candidate variables p-values are computed and sorted in an ascending order. Then, an estimate of the number of active variables $\widehat{p}_1(\alpha)$ is determined by finding the largest p-value that does not exceed a threshold depending on the target FDR α by solving

$$\widehat{p}_1(\alpha) = \max\left\{m : P_m \leq \frac{m}{p \cdot c(p)} \cdot \alpha\right\}, \tag{2.27}$$

[1] Throughout this dissertation, the original definition of the FDR in [BH95] is used. Other definitions of the FDR, such as the positive FDR [Sto03], exist. The interested reader is referred to both papers for discussions on different potential definitions of the FDR.

where $c(p) = 1$ for the *BH* method and $c(p) = \sum_{j=1}^{p} 1/j \approx \ln(p) + \gamma$ for the *BY* method with $\gamma \approx 0.577$ being the Euler-Mascheroni constant. If no such $\widehat{p}_1(\alpha)$ exists, then no hypothesis is rejected, i.e., no variable is selected. Otherwise, the variables corresponding to the $\widehat{p}_1(\alpha)$ smallest p-values are selected. The *BH* method requires independent hypotheses or, at least, a so-called positive regression dependency among the candidate variables to guarantee FDR control at the target level. In contrast, the *BY* method provably controls the FDR at the target level and does not require independent hypotheses or any assumptions regarding the dependency among the hypotheses. However, the *BY* method is more conservative than the *BH* method, i.e., it usually achieves a considerably lower TPR than the *BH* method at the same target FDR level.

2.2.3 The Fixed-X and the Model-X Knockoff Methods

The *fixed-X* knockoff method [BC15] is a relatively new method for controlling the FDR in sparse linear regression settings. Since it requires $n \geq 2p$ observations, it is not suitable for high-dimensional settings. The method generates a knockoff matrix $\mathring{\boldsymbol{X}}$ consisting of p knockoff variables and appends it to the original predictor matrix. The knockoff variables must be designed to mimic the usually unknown covariance structure of $\boldsymbol{X}$. Further, they must be designed to be, conditional on the original variables, independent of the response. Designing such knockoffs is difficult and especially computationally demanding. However, if the knockoffs are designed correctly, they act as a control group and when a knockoff variable enters the active set before its original counterpart it provides some evidence against this variable being a true positive.

The predictor matrix $\boldsymbol{X}$ of, e.g., the *Lasso* optimization problem in (2.8) is then replaced by $[\boldsymbol{X} \;\; \mathring{\boldsymbol{X}}]$ and the λ_1-values corresponding to the first entry points of the original and knockoff variables are extracted from the solution path resulting in

$$Z_j = \sup\{\lambda_1 : \hat{\beta}_j \neq 0 \text{ first time}\}, \quad j = 1, \ldots, p \tag{2.28}$$

and

$$\mathring{Z}_j = \sup\{\lambda_1 : \hat{\beta}_{j+p} \neq 0 \text{ first time}\}, \quad j = 1, \ldots, p. \tag{2.29}$$

The authors suggest to use the test statistics

$$W_j = (Z_j \vee \mathring{Z}_j) \cdot \operatorname{sign}(Z_j - \mathring{Z}_j), \quad j = 1, \ldots, p, \tag{2.30}$$

and to determine the threshold

$$\tau = \min\left\{\tau' \in \mathcal{W} : \frac{b + \left|\{j : W_j \leq -\tau'\}\right|}{\left|\{j : W_j \geq \tau'\}\right| \vee 1} \leq \alpha\right\}, \tag{2.31}$$

where $\mathcal{W} = \{|W_j| : j = 1, \ldots, p\} \backslash \{0\}$. Note that this is only one of the test statistics that were proposed by the authors. In general, many other test statistics obeying a certain sufficiency and anti-symmetry property are suitable for the knockoff method. In (2.31), $b = 0$ yields the knockoff method and $b = 1$ the more conservative (i.e., higher threshold τ) knockoff+ method. Finally, only those variables whose test statistics exceed the threshold are selected, which gives us the selected active set

$$\widehat{\mathcal{A}} = \{j : W_j \geq \tau\}. \tag{2.32}$$

The knockoff+ method controls the FDR at the target level α. The advantage of the knockoff over the knockoff+ method is that it is less conservative. But the knockoff method does not control the FDR.

The *model-X* knockoff method [Can+18] was proposed as an extension to the *fixed-X* knockoff method for high-dimensional settings. It does not require any knowledge about the conditional distribution of the response given the explanatory variables

$$Y|X_1, \ldots, X_p \tag{2.33}$$

but needs to know the distribution of the covariates

$$(X_{i1} \ \cdots \ X_{ip}), \quad i = 1, \ldots, n. \tag{2.34}$$

The difference to the deterministic design of *fixed-X* knockoffs is that *model-X* knockoffs need to be designed probabilistically by sequentially sampling each knockoff predictor

$$\mathring{\boldsymbol{x}}_j, \quad j = 1, \ldots p, \tag{2.35}$$

from the conditional distribution of

$$X_j|X_{-j} \text{ for } j = 1 \text{ and } X_j|X_{-j}, \mathring{X}_{1:j-1} \text{ for } j = 2, \ldots, p, \tag{2.36}$$

where X_{-j} is the set of all explanatory variables except for X_j and $\mathring{X}_{1:j-1} \coloneqq \{\mathring{X}_1, \ldots, \mathring{X}_{j-1}\}$ is the set of generated knockoffs. However, the authors state that determining a new conditional distribution for each knockoff predictor and sampling from it turned out to be complicated and computationally very expensive [Can+18]. The only case in which *model-X* knockoffs can be easily constructed by sampling from the Gaussian distribution with a certain mean vector and covariance matrix is when the covariates follow the Gaussian distribution. For all other distributions of the covariates, especially when p is large, the authors consider an approximate construction of *model-X* knockoffs which yields the so-called second-order *model-X* knockoffs. Unfortunately, however, there is no proof that FDR control is achieved with second-order *model-X* knockoffs. Moreover, for $p > 500$ the authors consider an approximate semidefinite program (asdp) instead of the original semidefinite program that needs to be solved to construct second-order *model-X* knockoffs. This is the default choice in the R package accompanying the *fixed-X* and *model-X* papers.[2]

2.3 Martingale Theory

Martingale theory [Wil91] is the backbone of the finite sample FDR control proofs in this dissertation. It has also been used to prove the FDR control property of existing methods (see, e.g., [STS04; BC15; Can+18]). Therefore, this section briefly revisits the fundamentals of martingale theory that are relevant for the FDR control proofs in this dissertation. For common terms and concepts of probability theory, such as σ-algebra, probability space, measurable function, random variable, almost sure (a.s.), the reader is referred to standard textbooks on probability theory, e.g., [Wil91; Kle13].

First, the terms filtered probability space, adapted process, martingale, sub-martingale, super-martingale, and stopping time are defined. Second, Doob's optional stopping theorem is introduced. Finally, the FDR control proof of the *BH* method from [STS04] is presented to illustrate the usefulness of martingale theory. All (slightly modified) definitions and theorems in this section are taken from the standard textbook on martingale theory by David Williams [Wil91].

[2]The R package containing the implementations of the *fixed-X* and the *model-X* methods is available at `https://CRAN.R-project.org/package=knockoff` (last access: June 26, 2024).

2.3.1 Definitions

For all subsequent definitions and the formulation of Doob's optional stopping theorem, the general probability space needs to be extended by adding an increasing collection of sub-σ-algebras, which is called filtration.

Definition 1 (Filtered probability space)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with sample space Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$. The increasing collection of sub-σ-algebras $\{\mathcal{F}\} \coloneqq \{\mathcal{F}_t : t \geq 0\}$, $t \in \{\mathbb{N}; \infty\}$, where*

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F} \tag{2.37}$$

is called a filtration and

$$(\Omega, \mathcal{F}, \{\mathcal{F}\}, \mathbb{P}) \tag{2.38}$$

is called a filtered probability space.

A discrete time stochastic process $X = (X_t : t \geq 0)$ can be defined on a filtered probability space. If the random variables that constitute the stochastic process are measurable with respect to the filtration, then the stochastic process is called adapted.

Definition 2 (Adapted stochastic process)**.** *A stochastic process $X = (X_t : t \geq 0)$ is called adapted (to the filtration $\{\mathcal{F}\}$) if for each t, X_t is $\mathcal{F}_t$-measurable.*

Definitions 1 and 2 can be intuitively understood as follows: The sub-σ-algebra $\mathcal{F}_t$ belonging to a filtration $\{\mathcal{F}\}$ contains all the information about $\omega \in \Omega$ up to time t (including t). Thus, the value $X_t(\omega)$ of a stochastic process $X = (X_t : t \geq 0)$, which is adapted to the filtration $\{\mathcal{F}\}$, is known at time t.

With these definitions in place, martingales, sub-martingales, and super-martingales, which are stochastic processes with a certain property, can be defined as follows:

Definition 3 (Martingale, sub-martingale, super-martingale)**.** *Let $(\Omega, \mathcal{F}, \{\mathcal{F}\}, \mathbb{P})$ be a filtered probability space. A stochastic process $X = (X_t : t \geq 0)$ that is adapted and satisfies $\mathbb{E}[|X_t|] < \infty$ for all $t \geq 0$ is called (relative to $(\{\mathcal{F}\}, \mathbb{P})$) a*

(i) martingale if $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = X_{t-1}$, a.s. ($t \geq 1$),

(ii) sub-martingale if $\mathbb{E}[X_t|\mathcal{F}_{t-1}] \geq X_{t-1}$, a.s. ($t \geq 1$),

(iii) super-martingale if $\mathbb{E}[X_t|\mathcal{F}_{t-1}] \leq X_{t-1}$, a.s. ($t \geq 1$).

Loosely speaking, martingales, super-martingales, and sub-martingales are stochastic processes that, on average, remain at the same level, decrease, and increase over time, respectively.

Martingales, super-martingales, and sub-martingales can be stopped by a stopping rule at a certain point in time. Loosely speaking, if that stopping rule is a random variable and does not use any future information (i.e., if at any time t the stopping rule is $\mathcal{F}_t$-measurable), then it is called a stopping time.

Definition 4 (Stopping time)**.** *Let $(\Omega, \mathcal{F}, \{\mathcal{F}\}, \mathbb{P})$ be a filtered probability space. A map $U : \Omega \to \{0, 1, 2, \ldots; \infty\}$ is called a stopping time if*

$$\{U \leq t\} \coloneqq \{\omega : U(\omega) \leq t\} \in \mathcal{F}_t, \forall t \leq \infty, \tag{2.39}$$

or, equivalently,

$$\{U = t\} \coloneqq \{\omega : U(\omega) = t\} \in \mathcal{F}_t, \forall t \leq \infty. \tag{2.40}$$

Stopping times and stopped martingales are essential building blocks of Doob's optional stopping theorem that is introduced in the next section.

2.3.2 Doob's Optional Stopping Theorem

Doob's optional stopping theorem is an essential result in martingale theory. It allows, under certain conditions, to replace the expected value of a stopped martingale by the expected value of that martingale at it's starting point. For super-martingales, it allows to upper-bound and for sub-martingales to lower-bound the expected values of the stopped stochastic processes with the expected values at the starting points of these stochastic processes.

Theorem 1 (Optional stopping)**.** *Let U be a stopping time. Let the stochastic process $X = (X_t : t \geq 0)$ be a super-martingale. Then, X_U is integrable and*

$$\mathbb{E}[X_U] \leq \mathbb{E}[X_0] \tag{2.41}$$

in each of the following situations

(i) U is bounded (for some $N \in \mathbb{N}$, $U(\omega) \leq N$, $\forall \omega \in \Omega$),

(ii) X is bounded (for some $R \in \mathbb{R}_+$, $|X_t(\omega)| \leq R$ for every t and every ω),

(iii) $\mathbb{E}[U] < \infty$ *and, for some* $R \in \mathbb{R}_+$,

$$|X_t(\omega) - X_{t-1}(\omega)| \leq R, \, \forall (t, \omega). \tag{2.42}$$

If any of the conditions (i) - (iii) holds and X *is a martingale, then*

$$\mathbb{E}[X_U] = \mathbb{E}[X_0]. \tag{2.43}$$

If any of the conditions (i) - (iii) holds and X *is a sub-martingale, then*

$$\mathbb{E}[X_U] \geq \mathbb{E}[X_0]. \tag{2.44}$$

2.3.3 FDR Control Proof of the BH Method Using Doob's Optional Stopping Theorem

Doob's optional stopping theorem is essential for many finite sample FDR control proofs (e.g., [STS04; BC15]). Although the FDR control proofs in this dissertation are different from the existing proofs, the theorem will be used as a major element of our FDR control proofs throughout this dissertation. In the following, the usage of the theorem is showcased by sketching the FDR control proof of the *BH* method from [STS04]. For this purpose, and as in Section 2.2.2, let $P_1, \ldots, P_p$ be sorted p-values corresponding to the null hypotheses $H_1, \ldots, H_p$, where $H_j : \beta_j = 0$ and $P_1 \leq P_2 \leq \ldots \leq P_p$. Let $v \in [0, 1]$ and let

$$V(v) \coloneqq |\{\text{null } j : P_j \leq v\}| \tag{2.45}$$

and

$$R(v) \coloneqq |\{j : P_j \leq v\}| \tag{2.46}$$

be the number of selected null variables (i.e., number of false positives) and the total number of selected variables (i.e., number of rejected null hypotheses), respectively. Then, the FDP and FDR of the *BH* method are specified as follows:

$$\text{FDR}(v) \coloneqq \mathbb{E}[\text{FDP}(v)] \coloneqq \mathbb{E}\left[\frac{V(v)}{R(v) \vee 1}\right]. \tag{2.47}$$

The threshold v that corresponds to the decision rule of the *BH* method in (2.27) is a stopping time that is given by

$$v \coloneqq \sup\{\nu \in [0,1] : \widehat{\mathrm{FDP}}(\nu) \leq \alpha\}, \tag{2.48}$$

where

$$\widehat{\mathrm{FDP}}(\nu) \coloneqq \frac{\nu \cdot p}{R(\nu) \vee 1}. \tag{2.49}$$

Noting that

$$R(P_m) = m \quad \text{and} \quad \widehat{\mathrm{FDP}}(P_m) = \frac{P_m \cdot p}{m \vee 1}, \tag{2.50}$$

it can be easily verified that the decision rule in (2.27) with $c(p) = 1$ is equivalent to rejecting all null hypotheses for which $P_j \leq v$ holds.

With these preliminaries in place, the FDR control theorem of the *BH* method is stated and the proof is sketched. Note that this is only a proof sketch that serves the purpose of showcasing the usage of Doob's optional stopping theorem in the context of FDR control. Therefore, we omit proving the intermediate result that $V(v)/v$ constitutes a backward-running martingale with respect to v. A backward-running martingale with respect to v is simply a martingale that has its starting point at the largest value of v, i.e., $v = 1$. For ease of readability, we do not distinguish between the argument v of $V(v)$ and $R(v)$ and the stopping time v and whenever it is not clear from the context whether we are referring to the stopped process or not, it is stated explicitly.

Theorem 2 (*BH* method - FDR control)**.** *Let the p-values corresponding to the p_0 true null hypotheses be independent and identically distributed (i.i.d.), uniformly distributed on the interval $[0,1]$, and independent of the non-null p-values. Let v be as defined in* (2.48)*. Then, the BH method controls the FDR exactly at $(p_0/p)\alpha$ and conservatively at the level $\alpha \in [0,1]$, i.e.,*

$$\mathrm{FDR}(v) = \frac{p_0}{p} \cdot \alpha \leq \alpha. \tag{2.51}$$

Proof. The process $R(v)$ that is stopped at v, as defined in (2.48), is upper semi-continuous and, therefore, $\widehat{\mathrm{FDP}}(v) = \alpha$ and, equivalently, $R(v) = vp/\alpha$. After plugging $R(v) = vp/\alpha$ into (2.47) it remains to show that

$$\mathrm{FDR}(v) = \frac{\alpha}{p} \cdot \mathbb{E}\left[\frac{V(v)}{v}\right] \tag{2.52}$$

is equal to $(p_0/p)\alpha$. Based on the assumptions on the null p-values stated in the theorem,

it can be shown that $V(v)/v$ constitutes a backward-running martingale with respect to v. Moreover, $V(v)/v$ that is stopped at v, as defined in (2.48), is bounded, i.e., $V(v)/v \leq R(v)/v = p/\alpha$. Thus, Doob's optional stopping theorem is applicable and yields

$$\mathbb{E}\left[\frac{V(v)}{v}\right] = \mathbb{E}\left[\frac{V(1)}{1}\right] = p_0. \tag{2.53}$$

The last equation follows from the fact that $V(v)$ follows a binomial distribution with p_0 trials, success probability v, and expected value $\mathbb{E}[V(v)] = p_0 v$. Finally,

$$\text{FDR}(v) = \frac{p_0}{p} \cdot \alpha \leq \alpha. \tag{2.54}$$

□

The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.

Marcel Proust

3

High-Dimensional False Discovery Rate Control

In this chapter, the proposed framework for high-dimensional variable selection with FDR control at the user-defined target level is presented. Section 3.1 presents the proposed Terminating-Random Experiments (*T-Rex*) selector in a nutshell. Section 3.2 details the general methodology of the *T-Rex* framework and Section 3.3 highlights its main ingredients. Section 3.4 formulates the optimization problem that governs the optimal calibration process of the *T-Rex* selector. In Section 3.5, the main theoretical results and calibration algorithms are presented. In Section 3.6 the theoretical properties of the *T-Rex* selector are numerically verified and its performance is compared against state-of-the-art benchmark methods via numerical experiments. Section 3.7 introduces the *Screen-T-Rex* selector, an extension of the *T-Rex* selector, which is designed to efficiently perform multiple reproducible GWAS by screening through large-scale genomics biobanks and providing the user with self-estimated achieved FDR levels. Section 3.8 summarizes this chapter.

The proposed methods are used in Chapter 6 to solve challenging real-world data problems in biomedical engineering. Technical proofs, numerical verifications, additional simulations, and other appendices are deferred to Appendix A.

The main content of this chapter is based on the publications [MMPewa] and [MMP23a]. The implementations of the developed methods are included in the open source R software packages (with C++ backend) TRexSelector [Mac+24c] and tlars [Mac+24b] on CRAN.

3.1 The T-Rex Selector in a Nutshell

The *T-Rex* selector is a scalable framework that transforms forward variable selection methods into FDR-controlling methods. This section briefly summarizes the key ideas, the main theoretical results, and the major advantages of the proposed *T-Rex* selector compared to existing methods.

3.1.1 Key Ideas

Intuitively, the *T-Rex* selector builds upon three key ideas:

Key idea 1: "Random experiments with dummy variables"
Let L computer-generated dummy variables compete with the p candidate variables that are contained in a given data set to be included by a forward selection method. For this, the forward selection method is provided with a response vector $\boldsymbol{y} \in \mathbb{R}^n$, where n is the number of samples, and an enlarged predictor matrix $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n \times (p+L)}$ that contains the p candidate variables and L dummy variables.

Repeat this procedure K times, where each iteration consists of an independent random experiment yielding a candidate set of included variables. Since the dummies act as flagged null variables, one can loosely imagine this as placebo controlled trials. And the active variables will have to succeed in sufficiently many placebo controlled trials.

Key idea 2: "Urn model"
Mathematically, the false positive variable selection process is modeled as an urn model. Since the response neither depends on the null variables, nor on the dummy variables, the false positive forward variable selection process can be modeled as picking nulls and dummies one by one without replacement from an urn containing the nulls and dummies until some number of dummies (i.e., T) are selected.

A variable is added to the final selected active set if its relative frequency over all random experiments exceeds a certain voting level. This voting level is determined such that the cardinality

of the selected active set is maximized, while a conservative estimator of the FDR remains below the target FDR level.

Key idea 3: "Early termination"
The random experiments are terminated early, as soon as the target FDR is exceeded. Not computing the full solution path provides a massive speedup for high-dimensional data. It is possible because, particularly for sparse problems, at some point along the forward variable selection process, the probability of selecting an active variable will be very low.

Summarizing, the *T-Rex* selector fuses the solutions of K early terminated random experiments, in which original and computer-generated dummy variables compete to be selected in a forward variable selection process. The *T-Rex* calibration algorithm automatically determines its parameters, i.e., (i) the number of generated dummies L, (ii) the number of included dummies before terminating the random experiments T, and (iii) the voting level in the fusion process, such that the FDR is controlled at the target level.

3.1.2 Main Theoretical Results

Our main theoretical results are summarized as follows:

1. Using martingale theory (see Section 2.3), we provide a finite sample FDR control proof (Theorem 3) that applies to low- ($p \leq n$) and high-dimensional ($p > n$) settings.
2. We prove that, for the *T-Rex* selector, the dummies can be sampled from any univariate distribution with finite mean and finite non-zero variance (Theorem 4). This is a fundamentally new result, and it does not hold for, e.g., knockoff methods (see Section 2.2.3) that require mimicking the covariance structure of the predictors, which is computationally expensive (see Figure 3.4).
3. We also prove that the proposed calibration algorithm is optimal in the sense that it maximizes the number of selected variables while controlling the FDR at the target level (Theorem 5).

3.1.3 Major Advantages Compared to Existing Methods

The major advantages compared to existing methods are:

1. The computation time of the *T-Rex* selector is multiple orders of magnitude lower compared to that of the current benchmark method (see Figure 3.1). Its complexity stems

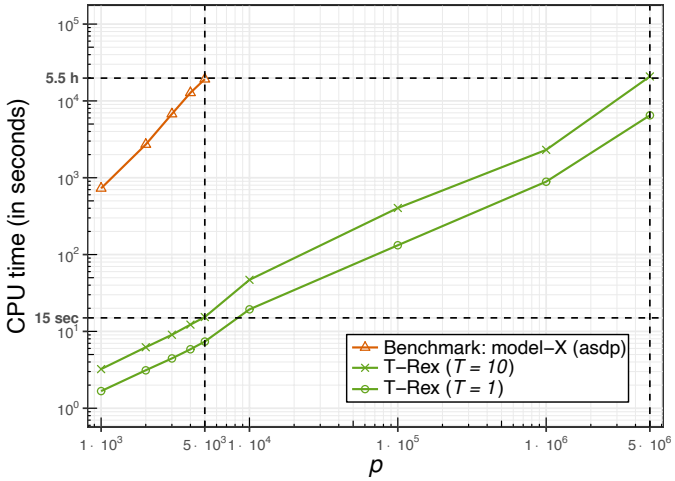


Figure 3.1: The sequential computation time of the *T-Rex* selector is multiple orders of magnitude lower than that of the *model-X* knockoff method [Can+18]. Note that, e.g., for $p = 5{,}000$ variables the absolute sequential computation time of the *T-Rex* selector for $T = 10$ included dummies is only 15 seconds as compared to more than 5.5 hours for the *model-X* knockoff method. Moreover, the sequential computation time of the *T-Rex* selector for 5,000,000 variables is comparable to that of the *model-X* knockoff method for only 5,000 variables. Note that both axes are scaled logarithmically. Setup: $n = 300$ (observations), $p_1 = 10$ (true active variables), $L = p$ (generated dummies), $K = 20$ (random experiments), SNR $= 1$, $MC = 955$ (Monte Carlo replications) for $p \leq 5{,}000$ and $MC = 100$ for $p > 5{,}000$.

from the computation of K terminated random experiments with expected complexity $\mathcal{O}(np)$ (see Appendix A.4).

2. As inputs, the *T-Rex* selector requires only the data and the target FDR level. The tuning of the sparsity parameter for *Lasso*-type methods [Tib96; Efr+04; ZH05; Zou06] is no longer required when incorporating them into the *T-Rex* selector framework.

In summary the *T-Rex* selector is, to the best of our knowledge, the first multivariate high-dimensional FDR-controlling method that scales to millions of variables in a reasonable amount of computation time (see Figure 3.1), which makes it a suitable method for large-scale GWAS, i.e., our major use-case.

3.2 Methodology

In this section, we will introduce the framework and the notation, which will be crucial for understanding why the *T-Rex* selector efficiently controls the FDR at the target level. A simplified overview of the *T-Rex* selector framework is provided in Figure 3.2. The general methodology underpinning the *T-Rex* selector consists of several steps that are detailed in the follow-

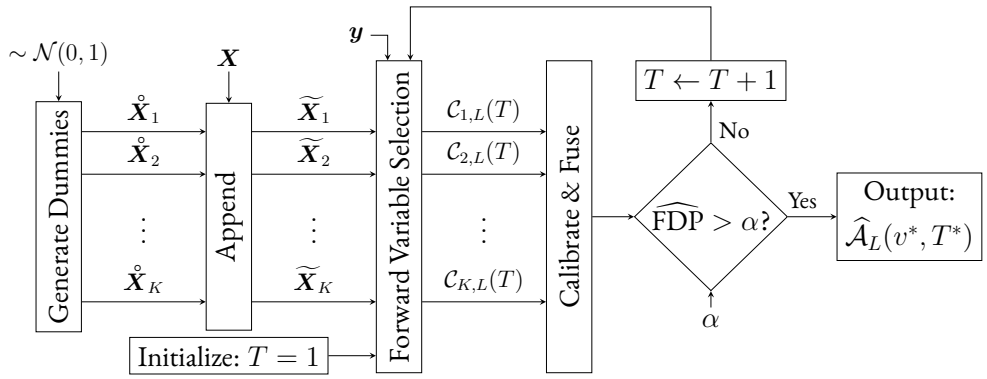


Figure 3.2: Simplified overview of the *T-Rex* selector framework: For each random experiment $k \in \{1, \dots, K\}$, the *T-Rex* selector generates a dummy matrix $\mathring{\boldsymbol{X}}_k$ containing L dummies and appends it to $\boldsymbol{X}$ to obtain the enlarged predictor matrix $\widetilde{\boldsymbol{X}}_k = \big[\boldsymbol{X}\ \mathring{\boldsymbol{X}}_k\big]$. With $\widetilde{\boldsymbol{X}}_k$ and the response $\boldsymbol{y}$ as inputs, a forward variable selection method is applied to obtain the candidate sets $\mathcal{C}_{1,L}(T), \dots, \mathcal{C}_{K,L}(T)$, where T is iteratively increased from one until $\widehat{\text{FDP}}$ (i.e., an estimate of the proportion of false discoveries among all selected variables that is determined by the calibration process) exceeds the target FDR level $\alpha \in [0, 1]$. Finally, a fusion procedure determines the selected active set $\widehat{\mathcal{A}}_L(v^*, T^*)$ for which the calibration procedure provides the optimal parameters v^* and T^*, such that the FDR is controlled at the target level α while maximizing the number of selected variables.

ing:

Step 1: Generate $K > 1$ dummy matrices $\mathring{\boldsymbol{X}}_k$, $k = 1, \dots, K$, each containing $L \geq 1$ dummy predictors that are sampled from a standard normal distribution.

Step 2: Append each dummy matrix to the original predictor matrix $\boldsymbol{X}$, resulting in the enlarged predictor matrices

$$\widetilde{\boldsymbol{X}}_k \coloneqq \big[\boldsymbol{X}\ \mathring{\boldsymbol{X}}_k\big] \tag{3.1}$$

$$= \big[\boldsymbol{x}_1\ \cdots\ \boldsymbol{x}_p\ \mathring{\boldsymbol{x}}_{k,1}\ \cdots\ \mathring{\boldsymbol{x}}_{k,L}\big], \quad k = 1, \dots, K, \tag{3.2}$$

where $\mathring{\boldsymbol{x}}_{k,1}, \dots, \mathring{\boldsymbol{x}}_{k,L}$ are the dummies. Figure 3.3 illustrates the enlarged predictor matrix, where $\mathcal{A}$ and $\mathcal{Z}$ denote the index sets of the true active variables and the null variables, respectively. Their respective cardinalities are denoted by $p_1 \coloneqq |\mathcal{A}|$ and $p_0 \coloneqq |\mathcal{Z}|$.

Step 3: Apply a forward variable selection procedure to $\big\{\widetilde{\boldsymbol{X}}_k, \boldsymbol{y}\big\}$, $k = 1, \dots, K$. For each random experiment, terminate the forward selection process after $T \geq 1$ dummy variables are included. This results in the candidate active sets $\mathcal{C}_{k,L}(T)$,

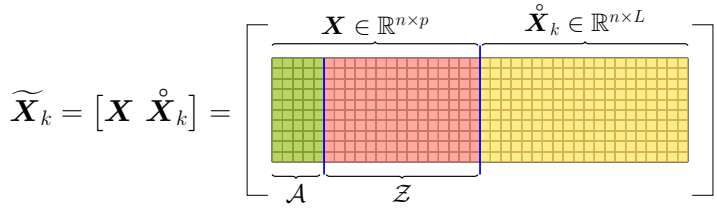


Figure 3.3: The enlarged predictor matrices $\widetilde{\boldsymbol{X}}_k$, $k = 1, \ldots, K$, replace the original predictor matrix $\boldsymbol{X}$ in each random experiment within the *T-Rex* selector framework. They contain the original and the dummy predictors. The index set of the true active variables and the index set of the null variables are denoted by $\mathcal{A}$ and $\mathcal{Z}$, respectively. The number of active variables and the number of null variables are denoted by $p_1 := |\mathcal{A}|$ and $p_0 := |\mathcal{Z}|$, respectively.

$k = 1, \ldots, K$. After terminating the forward selection process remove all dummies from the candidate active sets.[1]

Step 4: Iteratively increase T and carry out Step 3 until $\widehat{\text{FDP}}$ (i.e., a conservative estimate of the proportion of false discoveries among all selected variables) exceeds the target FDR level $\alpha \in [0, 1]$. The calibration process for determining $\widehat{\text{FDP}}$ and the optimal values v^* and T^* such that the FDR is controlled at the target level $\alpha \in [0, 1]$ while maximizing the number of selected variables is derived in Section 3.5.

Step 5: Fuse the candidate active sets to determine the estimate of the active set $\widehat{\mathcal{A}}_L(v^*, T^*)$. The fusion step is based on the *relative occurrences* $\Phi_{T,L}(j)$, $j = 1, \ldots, p$, of the original variables. For a mathematically rigorous definition of $\Phi_{T,L}(j)$, see Definition 5 in Section 3.4.

All variables whose relative occurrences at $T = T^*$ exceed the voting level $v^* \in [0.5, 1)$ are selected and the estimator of the active set is defined by

$$\widehat{\mathcal{A}}_L(v^*, T^*) := \{j : \Phi_{T^*,L}(j) > v^*\}. \tag{3.3}$$

The details of how the calibration process determines T^* and v^* such that, for any choice of L, the *T-Rex* selector controls the FDR at the target level while maximizing the number

[1] Since we use the *LARS* method throughout this chapter, variables can only be included but not dropped along the solution paths. Nevertheless, the *T-Rex* selector can also incorporate forward selection methods that remove some previously included variables from the candidate set along the solution path (e.g., *Lasso*). For such methods, the number of currently active dummies can decrease along the solution path. However, because the solution paths are terminated after T dummies are included for the first time, there is no ambiguity regarding the step in which the forward selection process ends.

of selected variables are deferred to Section 3.5.3. Moreover, an extension to the calibration process to jointly determine T^*, v^*, and L is also proposed in Section 3.5.4. The number of random experiments K is not subject to optimization. However, choosing $K \geq 20$ provides excellent empirical results and we never observed notable improvements for $K \geq 100$.[2]

3.3 Main Ingredients of the T-Rex Selector

The following example helps to develop an intuition for the three main ingredients of the *T-Rex* selector, which are

1. sampling dummies from the univariate standard normal distribution (see Figure 3.4),

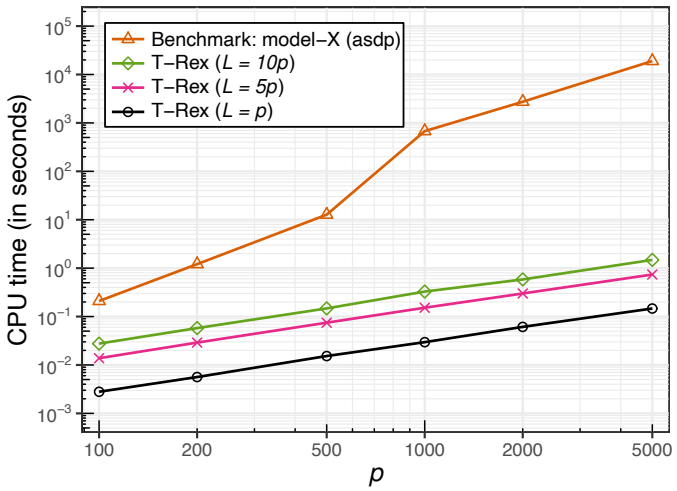
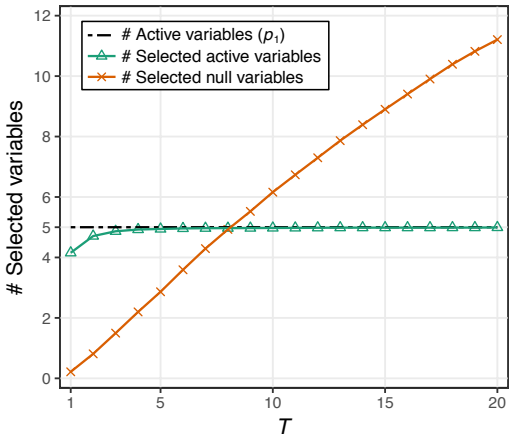


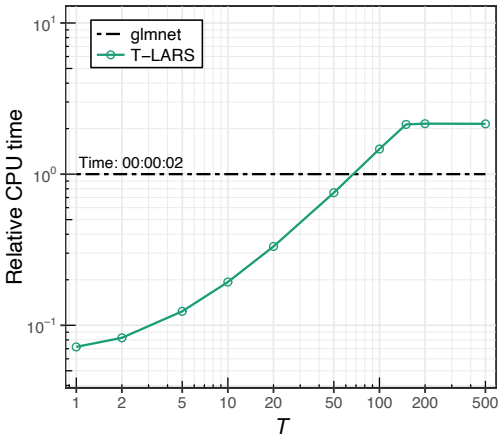
Figure 3.4: Ingredient 1 - sampling dummies from the univariate standard normal distribution. The sequential computation time of generating one dummy matrix for the proposed *T-Rex* selector is multiple orders of magnitude lower than the computation time of generating a knockoff matrix for the *model-X* knockoff method, which is a current benchmark. For example, for $p = 5{,}000$ and $L = p$, the *T-Rex* dummy generation process requires less than a second as compared to more than five hours for the *model-X* knockoff method. Even taking into account that the *T-Rex* selector requires, e.g., $K = 20$ of such dummy matrices, its sequential computation time is still multiple orders of magnitude lower than that of the *model-X* knockoff method. The jump in computation time for the *model-X* knockoff method between $p = 500$ and $p = 1{,}000$ is due to the suggestion of the authors to solve their proposed approximate semi-definite program (asdp) instead of their original semi-definite program for $p > 500$ in order to reduce the computation time required to generate *model-X* knockoffs.[3]Note that both axes are scaled logarithmically. Setup: $n = 300$, $MC = 955$.

[2]Instead of fixing the number of random experiments, it could be increased until the relative occurrences $\Phi_{T,L}(j)$, $j = 1, \ldots, p$, converge. However, a significant reduction of computation time is achieved by executing the independent random experiments in parallel on multicore computers or high-performance clusters. Therefore, in practice, fixing K to a multiple of the number of available CPUs is preferable.

2. early terminating the solution paths of the random experiments (see Figure 3.5), and



(a) Number of selected variables vs. T. Setup: $n = 150$, $p = 300$, $p_1 = 5$, $v = 0.8$, $L = p$, $K = 20$, SNR $= 1$, $MC = 500$.



(b) The pathwise coordinate descent algorithm from the R package "glmnet" [FHT10] (used to compute the *Lasso* on a λ-grid with 500 values) and the Terminating-*LARS* (*T-LARS*) algorithm from the R package "tlars" [Mac+24b]. Note that both axes are scaled logarithmically. Setup: $n = 300$, $p = 5{,}000$, $p_1 = 5$, $L = p$, SNR $= 1$, $MC = 955$.

Figure 3.5: Ingredient 2 - early terminating the solution paths of the random experiments. Figure (a) exemplifies that, on average, the number of selected active variables quickly increases towards the sparsity level p_1 (i.e., the number of active variables) and already for three included dummies almost all active variables are selected on average. However, the number of selected null variables also increases with increasing T. This example shows that terminating the forward selection process after selecting a few dummies is a reasonable strategy to select many active variables while keeping the number of selected null variables low. Therefore, the *T-Rex* selector terminates the solution paths early such that the number of selected variables is maximized while the FDR is controlled at the user-defined target level. Figure (b) illustrates that for $p = 5{,}000$ and $L = p$, when terminated early, the Terminating-*LARS* (*T-LARS*) algorithm (a fundamental building block of the *T-Rex* selector) is substantially faster than fitting the entire *Lasso* solution path using the pathwise coordinate descent algorithm for $2p$ variables. The pathwise coordinate descent algorithm is the state-of-the-art solver for Lasso-type problems and it is used by other FDR-controlling methods, such as the the *fixed-X* and *model-X* knockoff methods. Although the *T-Rex* selector needs to run the *T-LARS* algorithm for, e.g., $K = 20$ random experiments within the *T-Rex* selector, its sequential computation time is still comparable to that of a single run of "glmnet" in high-dimensional settings where p is much larger than n. Moreover, the independent random experiments can be run in parallel on multicore computers to achieve a substantial reduction in computation time. The "glmnet" computation time is used as the reference computation time and its absolute value is given above the reference line (format: hh:mm:ss). Note that after $T = 150$ dummies are included the computation time of the *T-LARS* algorithm does not increase further because the *T-LARS* algorithm includes at most $\min\{n, p+L\} = n = 300$ variables and with $T = 150$ we can expect that, on average, also 150 null variables plus the 5 active variables are included.

[3]See the default parameters in the R package implementing the *fixed-X* method and the *model-X* method, which is available at `https://CRAN.R-project.org/package=knockoff` (last access: June 26, 2024).

3. fusing the candidate sets based on their relative occurrences and a voting level $v \in [0.5, 1)$ (see Figure 3.6).

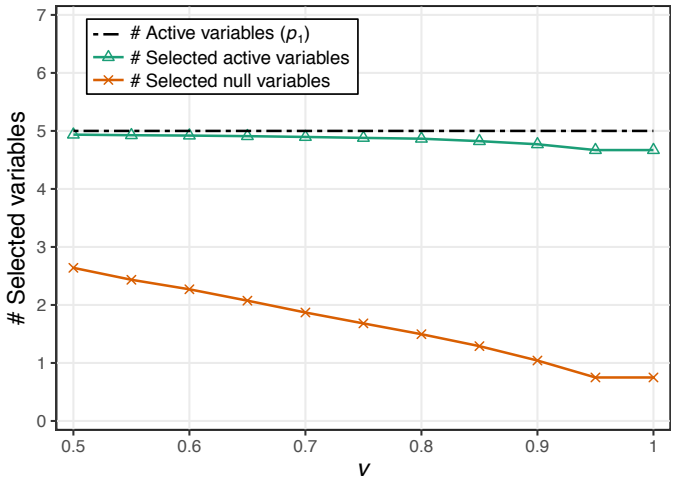


Figure 3.6: Ingredient 3 - fusing the candidate sets based on their relative occurrences and a voting level $v \in [0.5, 1)$. The number of selected active variables remains high when increasing the voting level, while the number of selected null variables decreases faster with increasing v. Setup: $n = 150, p = 300, p_1 = 5, T = 3, L = p, K = 20$, SNR $= 1, MC = 500$.

In the example, we generate sparse high-dimensional data sets with n observations and p predictors and a response that is generated by the linear model in (2.1). Further, $\beta_j = 1$ for active variables and $\beta_j = 0$ for null variables. The predictors are sampled from the standard normal distribution. The standard deviation σ is chosen such that the signal-to-noise ratio (SNR), which is given by $\text{Var}[\boldsymbol{X}\boldsymbol{\beta}] \,/\, \sigma^2$, is equal to one.[4] The specific values of n, p, p_1 (i.e., the number of active variables), v, T, L, K, SNR, and MC (i.e., the number of Monte Carlo realizations that the results are averaged over) are reported along with the discussion of the results in Figures 3.4, 3.5, and 3.6.

3.4 Problem Statement

Now that the notation and general steps of the *T-Rex* selector have been introduced, we are ready to formulate an optimization problem formalizing the task of selecting as many true positives as possible while controlling the FDR at the target level. We start with some remarks on notation followed by a mathematically rigorous definition of the relative occurrence of

[4] Note that, in this case, Var denotes the sample variance operator.

a candidate variable, which is crucial for the fusion step of the *T-Rex* selector (i.e., Step 5 in Section 3.2). Then, the definitions of the FDR and the TPR, which particularize the generic definitions in (2.23) and (2.24) for the *T-Rex* selector, are provided. Finally, the main constrained optimization problem is formulated.

For better readability, the arguments T and L of the estimator of the active set are dropped, i.e., $\widehat{\mathcal{A}}(v) := \widehat{\mathcal{A}}_L(v, T)$, except when referring specifically to the set in (3.3) for which the values v^* and T^* result from the calibration that will be discussed in Section 3.5. Note that the term "included candidates" refers to the variables that were picked (and not dropped) along the solution path of each random experiment while the term "selected variables" refers to the variables whose relative occurrences exceed the voting level $v \in [0.5, 1)$.

Definition 5 (Relative occurrence)**.** *Let $K \in \mathbb{N}_+ \setminus \{1\}$ be the number of random experiments, $L \in \mathbb{N}_+$ the number of dummies, and $T \in \{1, \dots, L\}$ the number of included dummies after which the forward variable selection process in each random experiment is terminated. The relative occurrence of variable $j \in \{1, \dots, p\}$ is defined by*

$$\Phi_{T,L}(j) := \begin{cases} \dfrac{1}{K} \sum\limits_{k=1}^{K} \mathbb{1}_k(j, T, L), & T \geq 1 \\ 0, & T = 0 \end{cases}, \tag{3.4}$$

where $\mathbb{1}_k(j, T, L)$ is the indicator function for which

$$\mathbb{1}_k(j, T, L) = \begin{cases} 1, & j \in \mathcal{C}_{k,L}(T) \\ 0, & \textit{otherwise} \end{cases}. \tag{3.5}$$

Definition 6 ($V_{T,L}(v)$, $S_{T,L}(v)$ and $R_{T,L}(v)$)**.** *The number of selected null variables $V_{T,L}(v)$, the number of selected active variables $S_{T,L}(v)$, and the number of selected variables $R_{T,L}(v)$ are defined, respectively, by*

$$V_{T,L}(v) := \left|\widehat{\mathcal{A}}^0(v)\right| := \left|\{\textit{null } j : \Phi_{T,L}(j) > v\}\right|, \tag{3.6}$$

$$S_{T,L}(v) := \left|\widehat{\mathcal{A}}^1(v)\right| := \left|\{\textit{active } j : \Phi_{T,L}(j) > v\}\right|, \textit{ and} \tag{3.7}$$

$$R_{T,L}(v) := V_{T,L}(v) + S_{T,L}(v) = \left|\widehat{\mathcal{A}}(v)\right|. \tag{3.8}$$

Note that $R_{T,L}(v)$ is observable, while $V_{T,L}(v)$ and $S_{T,L}(v)$ are not, since the set of true active variables is unknown.

To state the optimization problem for the calibration of the *T-Rex* selector, the FDR and TPR expressions in (2.23) and (2.24) are rewritten using Definition 6 as follows:

Definition 7 (FDP and FDR)**.** *The false discovery proportion (FDP) is defined by*

$$\mathrm{FDP}(v, T, L) \coloneqq \frac{V_{T,L}(v)}{R_{T,L}(v) \vee 1} \tag{3.9}$$

and the FDR is defined by

$$\mathrm{FDR}(v, T, L) \coloneqq \mathbb{E}\big[\,\mathrm{FDP}(v, T, L)\big], \tag{3.10}$$

where the expectation is taken with respect to the noise in (2.1)*.*

Definition 8 (TPP and TPR)**.** *The true positive proportion (TPP) is defined by*

$$\mathrm{TPP}(v, T, L) \coloneqq \frac{S_{T,L}(v)}{p_1 \vee 1}, \tag{3.11}$$

where p_1 denotes the unknown number of true active variables, and the TPR is defined by

$$\mathrm{TPR}(v, T, L) \coloneqq \mathbb{E}\big[\,\mathrm{TPP}(v, T, L)\big], \tag{3.12}$$

where the expectation is taken with respect to the noise in (2.1)*.*

Remark 1. *Note that if $R_{T,L}(v)$ is equal to zero, then $V_{T,L}(v)$ is zero as well. In this case, the denominator in the expression for the FDP is set to one and, thus, the FDP becomes zero. This is a reasonable solution to the "$0/0$" case, because when no variables are selected there exist no false discoveries. Similarly, when there exist no true active variables among the candidates, i.e. $p_1 = S_{T,L}(v) = 0$, the TPP equals zero.*

A major result of this work is to determine T^* and v^*, such that, for any fixed $L \in \mathbb{N}_+$, the *T-Rex* selector maximizes $\mathrm{TPR}(v, T, L)$ while provably controlling $\mathrm{FDR}(v, T, L)$ at any given target level $\alpha \in [0, 1]$. In practice, this amounts to finding the solution of the optimization problem

$$\underset{v,T}{\text{maximize}}\ \mathrm{TPP}(v, T, L) \quad \text{subject to} \quad \widehat{\mathrm{FDP}}(v, T, L) \leq \alpha, \tag{3.13}$$

which is equivalent to

$$\underset{v,T}{\text{maximize}}\ S_{T,L}(v) \quad \text{subject to} \quad \widehat{\mathrm{FDP}}(v, T, L) \leq \alpha \tag{3.14}$$

because p_1 is a constant. Note that $\widehat{\text{FDP}}(v, T, L)$ is a conservative estimator of $\text{FDP}(v, T, L)$, i.e., it holds that $\text{FDR}(v, T, L) = \mathbb{E}\big[\,\text{FDP}(v, T, L)\big] \leq \mathbb{E}\big[\widehat{\text{FDP}}(v, T, L)\big] = \widehat{\text{FDR}}(v, T, L)$. The details of the conservative FDP estimator are discussed in Section 3.5. Since we cannot observe $S_{T,L}(v)$, it is replaced by $R_{T,L}(v)$. This results in the final optimization problem:

$$\underset{v,T}{\text{maximize}}\; R_{T,L}(v) \quad \text{subject to} \quad \widehat{\text{FDP}}(v, T, L) \leq \alpha. \tag{3.15}$$

In words: *The T-Rex selector maximizes the number of selected variables while controlling a conservative estimator of the FDP at the target level* α.

In Section 3.5, it is shown that the *T-Rex* selector efficiently solves (3.15) and that any solution of (3.15) is a feasible solution of (3.13) and (3.14).

3.5 Main Results

This section contains the main results about the proposed *T-Rex* selector, which concern: FDR-control (Theorem 3), dummy generation (Theorem 4), and the optimal calibration algorithm (Theorem 5). We use martingale theory (see Section 2.3) to prove the FDR control property of the *T-Rex* selector. The developed FDR control theory relies on standard assumptions that are extensively verified especially for GWAS, i.e., the main use-case of the proposed methods (see Appendices A.5, A.6, and A.7). For a numerical evaluation of the *T-Rex* selector on a simulated GWAS, see Section 6.1. Additionally, the computational complexity of the *T-Rex* selector, which stems from the computation of K terminated random experiments with expected complexity $\mathcal{O}(np)$, is derived in Appendix A.4.

3.5.1 FDR Control

This section first defines the deflated relative occurrence, an FDP estimator, and the voting level, which are the essential building blocks for the FDR control theorem and its proof. Then, an important martingale lemma is formulated, followed by the FDR control theorem and its proof.

3.5.1.1 Preliminary Definitions

In Definition 5, the relative occurrence $\Phi_{T,L}(j)$ of the jth candidate variable has been introduced. It can be decomposed into the changes in relative occurrence, i.e.,

$$\Phi_{T,L}(j) = \sum_{t=1}^{T} \Delta\Phi_{t,L}(j), \quad j = 1, \ldots, p, \tag{3.16}$$

where $\Delta\Phi_{t,L}(j) := \Phi_{t,L}(j) - \Phi_{t-1,L}(j)$ is the change in relative occurrence from step $t-1$ to t for variable j.[5] Since the active and the null variables are interspersed in the solution paths of the random experiments, some null variables might appear earlier on the solution paths than some active variables.[6] Therefore, it is unavoidable that the $\Delta\Phi_{t,L}(j)$'s of the null variables are inflated along the solution paths of the random experiments. Moreover, we observe interspersion not only for active and null variables but also for dummies, which is expected since dummies can be interpreted as flagged null variables.

The above considerations motivate the definition of the *deflated relative occurrence* to harness the information about the fraction of included dummies in each step along the solution paths in order to deflate the $\Delta\Phi_{t,L}(j)$'s of the null variables and, thus, account for the interspersion effect.

Definition 9 (Deflated relative occurrence)**.** *The deflated relative occurrence of variable j is defined by*

$$\Phi'_{T,L}(j) := \sum_{t=1}^{T} \left(1 - \frac{p - \sum_{q=1}^{p} \Phi_{t,L}(q)}{L - (t-1)} \frac{1}{\sum_{q \in \widehat{\mathcal{A}}(0.5)} \Delta\Phi_{t,L}(q)}\right) \Delta\Phi_{t,L}(j), \tag{3.17}$$

$j = 1, \ldots, p$.

In words: *The deflated relative occurrence is the sum over the deflated $\Delta\Phi_{t,L}(j)$'s from step $t = 1$ until step $t = T$.* As detailed and intuitively explained in Appendix A.2, the $\Delta\Phi_{t,L}(j)$'s are multiplied by a deflation factor that takes into account the ratio between the fraction of

[5] When using a forward selection method within the *T-Rex* selector framework that does not drop variables along the solution path (e.g. *LARS*), all $\Phi_{t,L}(j)$'s are non-decreasing in t and, therefore, $\Delta\Phi_{t,L}(j) \geq 0$ for all j. In contrast, when using forward selection methods that might drop variables along the solution path (e.g. *Lasso*), the $\Phi_{t,L}(j)$'s might decrease in t and, therefore, the $\Delta\Phi_{t,L}(j)$'s can be negative. Nevertheless, the relative occurrence $\Phi_{T,L}(j)$ is non-negative for all j and any forward selection method.

[6] Many researchers have observed that active and null variables are interspersed in solution paths obtained from sparsity-inducing methods, such as the *LARS* algorithm or the *Lasso* [SBC17; BC15].

selected dummies and the fraction of selected candidate variables in each step $t \in \{1, \dots, T\}$.

Using the deflated relative occurrences, the estimator of $V_{T,L}(v)$, i.e., the number of selected null variables (see Definition 6), and the corresponding FDP estimator are defined as follows:

Definition 10 (FDP estimator)**.** *The estimator of $V_{T,L}(v)$ is defined by*

$$\widehat{V}_{T,L}(v) \coloneqq \sum_{j \in \widehat{\mathcal{A}}(v)} \left(1 - \Phi'_{T,L}(j)\right) \tag{3.18}$$

and the corresponding estimator of $\mathrm{FDP}(v, T, L)$ *is defined by*

$$\widehat{\mathrm{FDP}}(v, T, L) = \frac{\widehat{V}_{T,L}(v)}{R_{T,L}(v) \vee 1} \tag{3.19}$$

with

$$\widehat{\mathrm{FDR}}(v, T, L) \coloneqq \mathbb{E}\big[\widehat{\mathrm{FDP}}(v, T, L)\big] \tag{3.20}$$

being its expected value.

The main idea behind FDR control for the *T-Rex* selector is that controlling $\widehat{\mathrm{FDP}}(v, T, L)$ at the target level $\alpha \in [0, 1]$ guarantees that $\mathrm{FDR}(v, T, L)$ is controlled at the target level as well. To achieve this, we define $v \in [0.5, 1)$ as the voting level at which $\widehat{\mathrm{FDP}}(v, T, L)$ is controlled at the target level.

Definition 11 (Voting level)**.** *Let $T \in \{1, \dots, L\}$ and $L \in \mathbb{N}_+$ be fixed. Then, the voting level is defined by*

$$v \coloneqq \inf\{\nu \in [0.5, 1) : \widehat{\mathrm{FDP}}(\nu, T, L) \leq \alpha\} \tag{3.21}$$

with the convention that $v = 1$ if the infimum does not exist.

Remark 2. *Note that v has to be at least 50% to ensure that all selected variables occur in at least more than the majority of the candidate sets within the T-Rex selector. Further, the convention of setting $v = 1$ if the infimum does not exist ensures that no variables are selected when there exists no triple (T, L, v) that satisfies Equation* (3.21)*.*

Remark 3. *Recall that the aim that is stated in the optimization problem in* (3.15) *is to select as many variables as possible while controlling $\widehat{\mathrm{FDP}}(v, T, L)$ at the target level. For fixed T and L, this is achieved by the smallest voting level that satisfies the constraint on $\widehat{\mathrm{FDP}}(v, T, L)$. We can easily see that for any fixed T and L, the voting level in* (3.21) *solves the optimization problem in* (3.15)*. The reason is that for any two voting levels $v_1,\ v_2 \in [0.5, 1)$ with $v_2 \geq v_1$*

satisfying the $\widehat{\text{FDP}}$*-constraint in* (3.21)*, it holds that* $R_{T,L}(v_1) \geq R_{T,L}(v_2)$*.*

Remark 4. *If* v*,* T*, and* L *satisfy Equation* (3.21)*, then the FDP from Definition 7 can be upper-bounded using* (3.19) *as follows:*

$$\text{FDP}(v, T, L) = \frac{V_{T,L}(v)}{R_{T,L}(v) \vee 1} = \widehat{\text{FDP}}(v, T, L) \cdot \frac{V_{T,L}(v)}{\widehat{V}_{T,L}(v)} \tag{3.22}$$

$$\leq \alpha \cdot \frac{V_{T,L}(v)}{\widehat{V}_{T,L}(v)} \leq \alpha \cdot \frac{V_{T,L}(v)}{\widehat{V}'_{T,L}(v)}, \tag{3.23}$$

where $\widehat{V}'_{T,L}(v)$*, which is supposed to be greater than zero, is defined by*

$$\widehat{V}'_{T,L}(v) \coloneqq \widehat{V}_{T,L}(v) - \sum_{j \in \widehat{\mathcal{A}}(v)} \big(1 - \Phi_{T,L}(j)\big). \tag{3.24}$$

This upper bound on the FDP will be particularly useful in proving that the T-Rex selector has the FDR control property.

Remark 5. *The voting level* v *in* (3.21) *can be interpreted as a stopping time with respect to some still to be defined filtration within Lemma 1 in Section 3.5.1.2. See Definitions 1 and 4 in Section 2.3.1 for details on stopping times and filtrations.*

3.5.1.2 The FDR Control Theorem

Before the FDR control theorem is formulated, we introduce a lemma that contains the backbone of our FDR control theorem, which is rooted in martingale theory (see Section 2.3). That is, we state and prove that $V_{T,L}(v)/\widehat{V}'_{T,L}(v)$ in (3.23) is a backward-running super-martingale with respect to the voting level v and some still to be defined filtration. This will be an essential element of the FDR control proof because it allows to use Doob's optional stopping theorem (i.e., Theorem 1 in Section 2.3.2) to upper bound the expected value of $V_{T,L}(v)/\widehat{V}'_{T,L}(v)$.

Lemma 1. *Define* $\mathcal{V} \coloneqq \{\Phi_{T,L}(j) : \Phi_{T,L}(j) > 0.5,\ j = 1, \ldots, p\}$ *and*

$$H_{T,L}(v) \coloneqq \frac{V_{T,L}(v)}{\widehat{V}'_{T,L}(v)}. \tag{3.25}$$

Let $\mathcal{F}_v \coloneqq \sigma\big(\{R_{T,L}(u)\}_{u \geq v}, \{V_{T,L}(u)\}_{u \geq v}, \{\widehat{V}'_{T,L}(u)\}_{u \geq v}\big)$ *be a backward-filtration with respect to* v*. Then, for all tuples* $(T, L) \in \{1, \ldots, L\} \times \mathbb{N}_+$*,* $\{H_{T,L}(v)\}_{v \in \mathcal{V}}$ *is a backward-*

running super-martingale with respect to $\mathcal{F}_v$. That is,

$$\mathbb{E}\big[H_{T,L}\big(v - \epsilon^*_{T,L}(v)\big) \;\big|\; \mathcal{F}_v\big] \geq H_{T,L}(v), \tag{3.26}$$

where

$$\epsilon^*_{T,L}(v) \coloneqq \inf\{\epsilon \in (0, v) : R_{T,L}(v - \epsilon) - R_{T,L}(v) = 1\} \tag{3.27}$$

with $v \in [0.5, 1)$ and the convention that $\epsilon^*_{T,L}(v) = 0$ if the infimum does not exist.

Proof. The proof is deferred to Appendix A.1. □

Theorem 3 (FDR control - *T-Rex* selector)**.** *Suppose that $\widehat{V}'_{T,L}(v) > 0$. Then, for all triples $(T, L, v) \in \{1, \ldots, L\} \times \mathbb{N}_+ \times [0.5, 1)$ that satisfy Equation (3.21) and as $K \to \infty$, the T-Rex selector controls the FDR at any fixed target level $\alpha \in [0, 1]$, i.e.,*

$$\mathrm{FDR}(v, T, L) = \mathbb{E}\big[\,\mathrm{FDP}(v, T, L)\big] \leq \alpha. \tag{3.28}$$

Proof. With Lemma 1 and since the stopping time in (3.21) is adapted to the filtration, i.e., it is $\mathcal{F}_v$-measurable, and $H_{T,L}(v)$ is bounded, the optional stopping theorem (i.e., Theorem 1 in Section 2.3.2) can be applied to upper bound $\mathbb{E}\big[H_{T,L}(v)\big]$. This yields, as $K \to \infty$,

$$\mathbb{E}\big[H_{T,L}(v)\big] \leq \mathbb{E}\big[H_{T,L}(0.5)\big] = \frac{1}{\widehat{V}'_{T,L}(0.5)} \cdot \mathbb{E}\big[V_{T,L}(0.5)\big] \tag{3.29}$$

$$\leq \frac{1}{\widehat{V}'_{T,L}(0.5)} \cdot \frac{T}{L+1} \cdot p_0 \tag{3.30}$$

$$= \frac{1}{\dfrac{T}{L+1} \cdot p_0} \cdot \frac{T}{L+1} \cdot p_0 = 1. \tag{3.31}$$

The first inequality is a consequence of the optional stopping theorem and Lemma 1 and the equation in the first line follows from $\widehat{V}'_{T,L}(0.5)$ being deterministic as $K \to \infty$. The second line follows from $\mathbb{E}[\mathrm{NHG}(p_0 + L, p_0, T)] = T \cdot p_0/(L+1)$ and $V_{T,L}(v) \overset{d}{\leq} \mathrm{NHG}(p_0 + L, p_0, T)$, $v \in [0.5, 1)$, i.e. $V_{T,L}(v)$ is stochastically dominated by the negative hypergeometric distribution (NHG) with $p_0 + L$ total elements, p_0 success elements, and T failures after which a random experiment is terminated (for more details, see Appendix A.5). The third line

holds since

$$\widehat{V}'_{T,L}(0.5) = \sum_{t=1}^{T} \frac{p_0 - \sum_{q \in \mathcal{Z}} \Phi_{t,L}(q)}{L - (t-1)} = \sum_{t=1}^{T} \frac{p_0 - \frac{t}{L+1} \cdot p_0}{L - (t-1)} \tag{3.32}$$

$$= \frac{p_0}{L+1} \cdot \sum_{t=1}^{T} \frac{L-t+1}{L-t+1} = \frac{T}{L+1} \cdot p_0, \tag{3.33}$$

where the second equation follows from Lemma 4 in Appendix A.1. Finally, it follows that

$$\mathrm{FDR}(v, T, L) = \mathbb{E}\big[\,\mathrm{FDP}(v, T, L)\big] \leq \alpha \cdot \mathbb{E}\big[H_{T,L}(v)\big] \leq \alpha, \tag{3.34}$$

i.e., FDR control at the target level α is achieved. □

3.5.2 Dummy Generation

The *T-Rex* selector is not the first method to use dummies to perform variable selection. However, it utilizes dummies in a fundamentally different manner than existing variable selection methods (e.g., [Mil84; Mil02; WBS07]), which do not guarantee FDR control. As shown in Figure 3.2, the *T-Rex* selector generates L i.i.d. dummies for each random experiment by sampling each element of the dummy vectors from the standard normal distribution, i.e.,

$$\mathring{\boldsymbol{x}}_l = [\mathring{x}_{1,l} \cdots \mathring{x}_{n,l}]^\top, \text{ where } \mathring{x}_{i,l} \sim \mathcal{N}(0,1), \tag{3.35}$$

$i = 1, \ldots, n, l = 1, \ldots, L$. This raises the question whether dummies can be sampled from other distributions, as well, to serve as flagged null variables. From an asymptotic point of view, i.e., $n \to \infty$, and if some mild conditions are satisfied, the perhaps at first glance surprising answer to this question is that *dummies can be sampled from any univariate probability distribution with finite expectation and finite non-zero variance in order to serve as flagged null variables within the T-Rex selector*.

We will prove the above statement for any forward selection procedure that uses sample correlations of the predictors with the response or with the current residuals in each forward selection step to determine which variable is included next. Thus, the statement is true, e.g., for the *LARS* algorithm, *Lasso*, adaptive *Lasso*, and *elastic net*.

Recall that null variables and dummies are not related to the response. For null variables this

holds by definition and for dummies this holds because dummies are generated without using any information about the response.[7] Moreover, the sample correlations of the dummies with the response are random. Thus, the higher the number of generated dummies, the higher the probability of including a dummy instead of a null or even a true active variable in the next step of a random experiment. These considerations suggest that only the number of dummies within the enlarged predictor matrices is relevant for the behavior of the forward selection process in each random experiment. That is, for $n \to \infty$, the distribution from which the dummies are sampled has no influence on the distribution of the correlation variables

$$\mathring{G}_{l,m,k} \coloneqq \sum_{i=1}^{n} \gamma_{i,m,k} \cdot \mathring{X}_{i,l,k}, \tag{3.36}$$

$l \in \mathcal{D}_{m,k}$, $m \geq 1$, $k = 1, \ldots, K$, where $\gamma_{i,m,k}$ is the ith element of $\boldsymbol{\gamma}_{m,k} \coloneqq \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{m,k}$ (i.e., the residual vector in the mth forward selection step of the kth random experiment) with $\hat{\boldsymbol{\beta}}_{m,k}$ and $\mathcal{D}_{m,k}$ being the estimator of the parameter vector and the index set of the non-included dummies in the mth forward selection step of the kth random experiment, respectively.[8] The random variable $\mathring{X}_{i,l,k}$ represents the ith element of the lth dummy within the kth random experiment. Therefore, $\mathring{G}_{l,m,k}$ is simply a weighted sum of the i.i.d. random variables $\mathring{X}_{1,l,k}, \ldots, \mathring{X}_{n,l,k}$ with fixed weights $\gamma_{1,m,k}, \ldots, \gamma_{n,m,k}$. With these preliminaries in place, the dummy generation theorem is formulated as follows:

Theorem 4 (Dummy generation)**.** *Let $\mathring{X}_{i,l,k}$, $i = 1, \ldots, n$, $l \in \mathcal{D}_{m,k}$, $m \geq 1$, $k = 1, \ldots, K$, be standardized i.i.d. dummy random variables (i.e., $\mathbb{E}\big[\mathring{X}_{i,l,k}\big] = 0$ and $\mathrm{Var}\big[\mathring{X}_{i,l,k}\big] = 1$ for all i, l, m, k) following any probability distribution with finite expectation and finite non-zero variance. Define*

$$D_{n,l,m,k} \coloneqq \frac{1}{\Gamma_{n,m,k}} \cdot \mathring{G}_{l,m,k}, \tag{3.37}$$

where $\Gamma^2_{n,m,k} \coloneqq \sum_{i=1}^{n} \gamma^2_{i,m,k}$ with $\Gamma_{n,m,k} > 0$ for all n, m, k and with fixed $\gamma_{i,m,k} \in \mathbb{R}$ for

[7] Note that the knockoff generation processes of the *fixed-X* and the *model-X* knockoff method, i.e., the benchmark methods, are fundamentally different from our approach that uses dummies. Although these methods also do not use any information about the response to generate the knockoffs, unlike the proposed *T-Rex* selector, they must incorporate the covariance structure of the predictor matrix, which leads to a large computation time, especially for high dimensions (see Figures 3.1 and 3.4).

[8] Note that $\boldsymbol{\gamma}_{1,k} = \boldsymbol{y}$ for all k, since $\hat{\boldsymbol{\beta}}_{1,k} = \boldsymbol{0}$ for all k, i.e., the residual vector in the first step of the forward selection process is simply the response vector $\boldsymbol{y}$.

all i, m, k. *Suppose that*

$$\lim_{n\to\infty} \frac{\gamma_{i,m,k}}{\Gamma_{n,m,k}} = 0, \quad i = 1, \ldots, n, \tag{3.38}$$

for all m, k. *Then, as* $n \to \infty$,

$$D_{n,l,m,k} \xrightarrow{d} D, \quad D \sim \mathcal{N}(0,1), \tag{3.39}$$

for all l, m, k.

Proof sketch. The Lindeberg-Feller central limit theorem is applicable because $\mathring{X}_{i,l,k}$, $i = 1, \ldots, n$, $l \in \mathcal{D}_{m,k}$, $m \geq 1$, $k = 1, \ldots, K$, are i.i.d. random variables and it holds that $\mathbb{E}\big[D_{n,l,m,k}\big] = 0$ and $\mathrm{Var}\left[D_{n,l,m,k}\right] = 1$. Moreover, since $\mathring{Q}_{i,l,m,k} \coloneqq \gamma_{i,m,k} \cdot \mathring{X}_{i,l,k} \,/\, \Gamma_{n,m,k}$ satisfies the Lindeberg condition for all l, m, k, the theorem follows. □

The details of the proof and illustrative examples with non-Gaussian dummies are deferred to Appendix A.1 and Appendix A.9, respectively.

Remark 6. *Note that sampling dummies from any univariate probability distribution with finite expectation and finite non-zero variance to serve as flagged null variables is only reasonable in combination with multiple random experiments as conducted by the proposed T-Rex selector. We emphasize that Theorem 4 is not applicable to knockoff generation procedures of, e.g., fixed-X and model-X knockoffs.*

3.5.3 Optimal Calibration Algorithm

This section describes the proposed *T-Rex* calibration algorithm, which efficiently solves the optimization problem in (3.15) and provides feasible solutions for (3.13) and (3.14). The pseudocode of the *T-Rex* calibration method is provided in Algorithm 2. To ensure clarity, we provide the following comprehensive summary of the algorithm flow: First, the number of dummies L and the number of random experiments K are set (usually $L = p$ and $K = 20$).[9] Then, setting $v = 1 - \Delta v$ and starting at $T = 1$, the number of included dummies is iteratively increased until reaching the value of T for which the FDP estimate at a voting level of $v = 1 - \Delta v$ exceeds the target level for the first time. In each iteration, before the target level

[9] As already mentioned in Section 3.2, K is not subject to optimization. In practice, choosing $K = 20$ already provides excellent results (see Section 3.6) and only incremental improvements are achieved with larger values of K.

Algorithm 2 *T-Rex* Calibration.

1. **Input**: $\alpha \in [0,1]$, K, L, $\boldsymbol{X}$, $\boldsymbol{y}$.
2. **Set** $T = 1, \Delta v = \dfrac{1}{K}, \widehat{\text{FDP}}(v = 1 - \Delta v, T, L) = 0.$
3. **While** $\widehat{\text{FDP}}(v = 1 - \Delta v, T, L) \leq \alpha$ and $T \leq L$ **do**:
   - 3.1. **For** $v = 0.5, 0.5 + \Delta v, 0.5 + 2 \cdot \Delta v, \ldots, 1 - \Delta v$ **do**:
     - i. **Compute** $\widehat{\text{FDP}}(v, T, L)$ as in (3.19).
     - ii. **If** $\widehat{\text{FDP}}(v, T, L) \leq \alpha$

       Compute $\widehat{\mathcal{A}}_L(v, T)$ as in (3.3).

       Else

       Set $\widehat{\mathcal{A}}_L(v, T) = \varnothing$.
   - 3.2. **Set** $T \leftarrow T + 1$.
4. **Solve**

$$\begin{aligned} &\max_{v', T'} \left| \widehat{\mathcal{A}}_L(v', T') \right| \\ &\text{s.t.} \quad T' \in \{1, \ldots, T-1\} \\ &\qquad v' \in \{0.5, 0.5 + \Delta v, 0.5 + 2 \cdot \Delta v, \ldots, 1 - \Delta v\} \end{aligned}$$

   and let (v^*, T^*) be a solution.
5. **Output**: (v^*, T^*) and $\widehat{\mathcal{A}}_L(v^*, T^*)$.

is exceeded, $\widehat{\mathcal{A}}_L(v, T)$ is computed as in (3.3) on a grid for v, while for values of v for which $\widehat{\text{FDP}}(v, T, L)$ exceeds the target level $\widehat{\mathcal{A}}_L(v, T)$ is equal to the empty set. Picking the v' and T' that maximize the number of selected variables yields the final solution.[10]

The reason for exiting the "while"-loop in Step 3 of Algorithm 2 when the FDP estimate at a voting level of $1 - \Delta v$ exceeds the target level for the first time is based on two key observations from our still to be presented simulation results (see Figure 3.7):

[10] In case of multiple solutions, we recommend to choose the solution with the largest v because such a solution provides the variables that were selected most frequently. Nevertheless, all solutions to the calibration problem that are computed using Algorithm 2 provide FDR control while maximizing the number of selected variables.

1. For any fixed T and L the average value of $\widehat{\text{FDP}}(v, T, L)$ decreases as v increases.

2. For any fixed v and L the average value of $\widehat{\text{FDP}}(v, T, L)$ increases as T increases.

Remark 7. *To foster the intuition behind these observations, we note that Equation* (3.19) *can be written as follows:*

$$\widehat{\text{FDP}}(v, T, L) = \frac{\widehat{V}_{T,L}(v)}{\big(V_{T,L}(v) + S_{T,L}(v)\big) \vee 1}. \tag{3.40}$$

Taking Definition 6, Definition 10, and the reformulation of Equation (3.19) *into account, we see that the observations suggest that we can expect the rather conservative estimate $\widehat{V}_{T,L}(v)$ of $V_{T,L}(v)$ in the numerator to decrease faster than the total number of selected variables $V_{T,L}(v) + S_{T,L}(v)$ in the denominator when increasing the voting level v. This is something that can be expected since, in general, assuming any variable selection method that on average performs better than random selection, active variables are expected to have higher relative occurrences than null variables and, therefore, remain selected even for large values of the voting level v. A similar reasoning can be applied to intuitively understand the monotonical increase of $\widehat{\text{FDP}}(v, T, L)$ with respect to T.*

With these preliminaries in place, the optimal calibration theorem can be formulated:

Theorem 5 (Optimality of Algorithm 2)**.** *Let (v^*, T^*) be a solution determined by Algorithm 2 and suppose that, ceteris paribus, $\widehat{\text{FDP}}(v, T, L)$ is monotonically decreasing in v and monotonically increasing in T. Then, (v^*, T^*) is an optimal solution of* (3.15) *and a feasible solution of* (3.13) *and* (3.14)*.*

Proof sketch. Since the objective functions of the optimization problems in Step 4 of Algorithm 2 and in (3.15) are equivalent, i.e., $\big|\widehat{\mathcal{A}}_L(v, T)\big| = R_{T,L}(v)$, it only needs to be shown that the feasible set in Step 4 of the algorithm contains the feasible set of (3.15). Since the conditions of the optimization problems in (3.13), (3.14), and (3.15) are equivalent, this also proves that (v^*, T^*) is a feasible solution of (3.13) and (3.14). □

The details of the proof are deferred to Appendix A.1.

3.5.4 Extension to the Calibration Algorithm

In Theorem 3, we have also established that the *T-Rex* selector controls the FDR at the target level for any choice of the number of dummies L. However, the choice of L has an influence

on how tightly the FDR is controlled at the target level (see Figure 3.7). Since controlling the FDR more tightly usually increases the TPR (i.e., power), it is desirable to choose the parameters of the *T-Rex* selector accordingly. We will see in the simulations in Section 3.6 that with increasing L, the FDR can be more tightly controlled at low target levels. In order to harness the positive effects that come with larger values of L while limiting the increased memory requirement for high values of L, we propose an extended version of the calibration algorithm that jointly determines v, T, and L such that the FDR is more tightly controlled at the target FDR level while not running out of memory.[11] The major difference to Algorithm 2 is that the number of dummies L is iteratively increased until the estimate of the FDP falls below the target FDR level α. The pseudocode of the extended *T-Rex* calibration algorithm is provided in Algorithm 3.[12]

Note that the extension to Algorithm 2 lies in Step 2 and Step 3. Additionally, and in contrast to Algorithm 2, the input to the algorithm is extended by a reference voting level $\tilde{v} \in [0.5, 1)$ and the maximum values of L and T, namely L_{max} and T_{max}. To ensure clarity, the algorithm flow is briefly summarized as follows: First L and T are set to be $L = p$ and $T = 1$. Then, starting at $L = p$, the number of dummies L is iteratively increased in steps of p until the estimate of the FDP at the voting level $\tilde{v}$ falls below the target FDR level α or L exceeds L_{max}. The rest of the algorithm is as in Algorithm 2 except that the "while"-loop in Step 5 of Algorithm 3 is exited when T exceeds T_{max}.

What remains to be discussed are the choices of the hyperparameters $\tilde{v}$, L_{max}, and T_{max}. Throughout this dissertation, we have set $\tilde{v} = 0.75$, $L_{\text{max}} = 10p$, and $T_{\text{max}} = \lceil n/2 \rceil$, where $\lceil n/2 \rceil$ denotes the smallest integer that is equal to or larger than $n/2$. An explanation and a discussion of these choices are deferred to Appendix A.3.

3.6 Numerical Experiments

In this section, the performances of the proposed *T-Rex* selector and the benchmark methods are compared in a simulation study. As discussed in Section 2.2, the benchmark meth-

[11] The reader might raise the question whether also the computation time increases with increasing L. There is no definite answer to this question. On the one hand, for very large values of L the computation time might increase. On the other hand, with increasing L the solution paths of the experiments are terminated earlier because the probability of selecting dummies grows with increasing L. Thus, increasing L might increase or decrease the computation time depending on whether the first or the second effect dominates.

[12] The R package *TRexSelector* [Mac+24c] contains the implementation of the extended calibration algorithm in Algorithm 3.

Algorithm 3 Extended *T-Rex* Calibration.

1. **Input**: $\alpha \in [0,1]$, K, $\boldsymbol{X}$, $\boldsymbol{y}$, $\tilde{v}$, $L_{\max}$, $T_{\max}$.
2. **Set** $L = p, T = 1$.
3. **While** $\widehat{\text{FDP}}(v = \tilde{v}, T, L) > \alpha$ and $L \leq L_{\max}$ **do**:

   Set $L \leftarrow L + p$.
4. **Set** $\Delta v = \dfrac{1}{K}$, $\widehat{\text{FDP}}(v = 1 - \Delta v, T, L) = 0$.
5. **While** $\widehat{\text{FDP}}(v = 1 - \Delta v, T, L) \leq \alpha$ and $T \leq T_{\max}$ **do**:

   5.1. **For** $v = 0.5, 0.5 + \Delta v, 0.5 + 2 \cdot \Delta v, \ldots, 1 - \Delta v$ **do**:

      i. **Compute** $\widehat{\text{FDP}}(v, T, L)$ as in (3.19).

      ii. **If** $\widehat{\text{FDP}}(v, T, L) \leq \alpha$

         Compute $\widehat{\mathcal{A}}_L(v, T)$ as in (3.3).

      Else

         Set $\widehat{\mathcal{A}}_L(v, T) = \varnothing$.

   5.2. **Set** $T \leftarrow T + 1$.
6. **Solve**

$$\begin{aligned} &\max_{v', T'} \left| \widehat{\mathcal{A}}_L(v', T') \right| \\ &\text{s.t.} \quad T' \in \{1, \ldots, T-1\} \\ &\qquad v' \in \{0.5, 0.5 + \Delta v, 0.5 + 2 \cdot \Delta v, \ldots, 1 - \Delta v\} \end{aligned}$$

   and let (v^*, T^*) be a solution.
7. **Output**: (v^*, T^*) and $\widehat{\mathcal{A}}_L(v^*, T^*)$.

ods in low-dimensional settings (i.e., $p \leq n$) are the well-known Benjamini-Hochberg (*BH*) method, the Benjamini-Yekutieli (*BY*) method, and the *fixed-X* knockoff methods, while the *model-X* knockoff methods are the benchmarks in high-dimensional settings (i.e., $p > n$). Knockoff methods come in two variations called "knockoff" and "knockoff+". Only the "knockoff+" version is an FDR-controlling method.

3.6.1 Setup

We generate a sparse high-dimensional setting[13] with n observations, p predictors, and a response given by the linear model in (2.1). Further, $\beta_j = 1$ for p_1 randomly selected j's while $\beta_j = 0$ for the others. The predictors are (i) sampled independently from the standard normal distribution (Figures 3.7 and 3.8) and (ii) sampled from an autoregressive model of order one with autocorrelation coefficient $\rho = 0.5$ (Figure 3.9). The standard deviation of the noise σ is chosen such that the signal-to-noise ratio (SNR), which is given by $\text{Var}[\boldsymbol{X}\boldsymbol{\beta}] \,/\, \sigma^2$, is equal to the desired value. In Appendices A.9 and A.10, we show results for non-Gaussian predictors and heavy-tailed noise settings. The specific values of the above described simulation setting and the parameters of the *T-Rex* selector, i.e., the values of n, p, p_1, SNR, K, L, T, v, are specified in the figure captions. The results are averaged over $MC = 955$ Monte Carlo replications.[14]

First, in order to assess the FDR control performance and the achieved power of the *T-Rex* selector, respectively, the average FDP, $\widehat{\text{FDP}}$, and TPP are computed over a two-dimensional grid for v and T for different values of L. Then, leaving all other parameters in this setup fixed, we compare the performance of the proposed *T-Rex* selector in combination with the proposed extended calibration algorithm in Algorithm 3 with the benchmark methods for different values of p_1 and the SNR at a target FDR level of 10%.

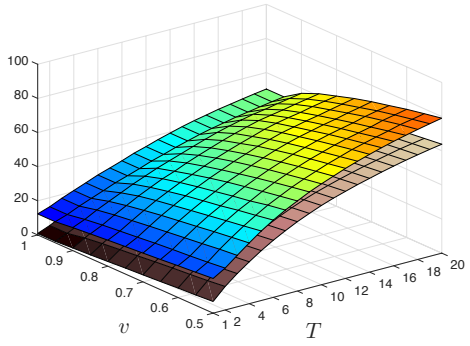
3.6.2 Results

The reported average FDP, $\widehat{\text{FDP}}$, and TPP (all averaged over 955 Monte Carlo replications) in Figures 3.7, 3.8, and 3.9 are estimates of the FDR, $\widehat{\text{FDR}}$, and TPR, respectively. For this reason, the results are discussed in terms of the FDR, $\widehat{\text{FDR}}$, and TPR in the captions of the figures, while the axes labels emphasize that the average FDP, $\widehat{\text{FDP}}$, and TPP are plotted.

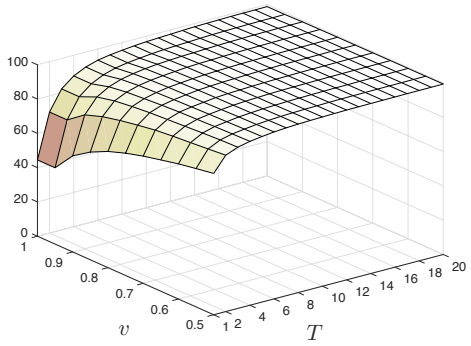
The simulation results confirm that the proposed *T-Rex* selector possesses the FDR control property. Moreover, the simulation results show that the *T-Rex* selector outperforms the benchmark methods while its computation time is multiple orders of magnitude lower than

[13] Additional simulation results that allow for a performance comparison of the proposed *T-Rex* selector to the *BH* method, the *BY* method, and the *fixed-X* knockoff methods in a low-dimensional setting are deferred to Appendix A.8.

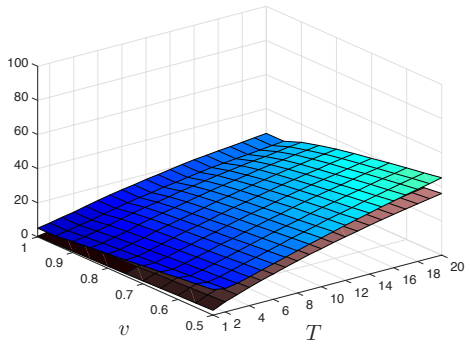
[14] The reason for running 955 Monte Carlo replications is that the simulations were conducted on the Lichtenberg High-Performance Computer of the Technische Universität Darmstadt, which consists of multiple nodes of 96 CPUs each. In order to run computationally efficient simulations, our computation jobs are designed to request 2 nodes and run 5 cycles on each CPU while one CPU acts as the master, i.e., $(2 \cdot 96 - 1) \cdot 5 = 955$.



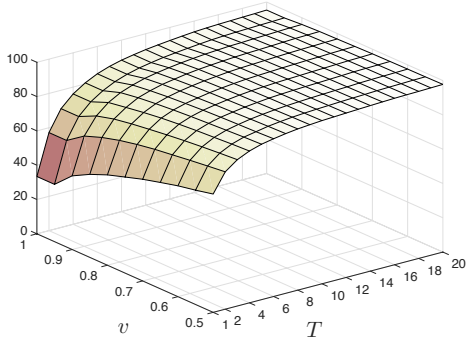
(a) Average FDP and $\widehat{\text{FDP}}$ ($L = p$).



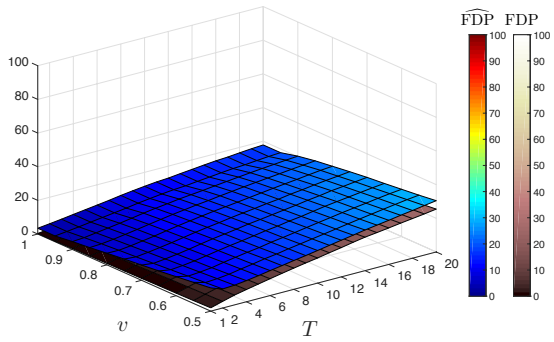
(b) Average TPP ($L = p$).



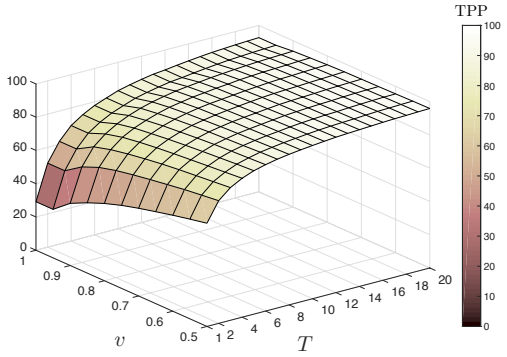
(c) Average FDP and $\widehat{\text{FDP}}$ ($L = 3p$).



(d) Average TPP ($L = 3p$).



(e) Average FDP and $\widehat{\text{FDP}}$ ($L = 5p$).



(f) Average TPP ($L = 5p$).

Figure 3.7: The *T-Rex* selector controls the FDR for all values of v and T while achieving a high power, even at low values of T. Note that the FDR control is tighter for large values of L.This observation led to the development of Algorithm 3. Moreover, we observe that the conditions in Theorem 5 hold on average (i.e., ceteris paribus, $\widehat{\text{FDP}}(v, T, L)$ is monotonically decreasing in v and monotonically increasing in T). Setup: $n = 300$, $p = 1{,}000$, $p_1 = 10$, $K = 20$, SNR $= 1$, $MC = 955$.

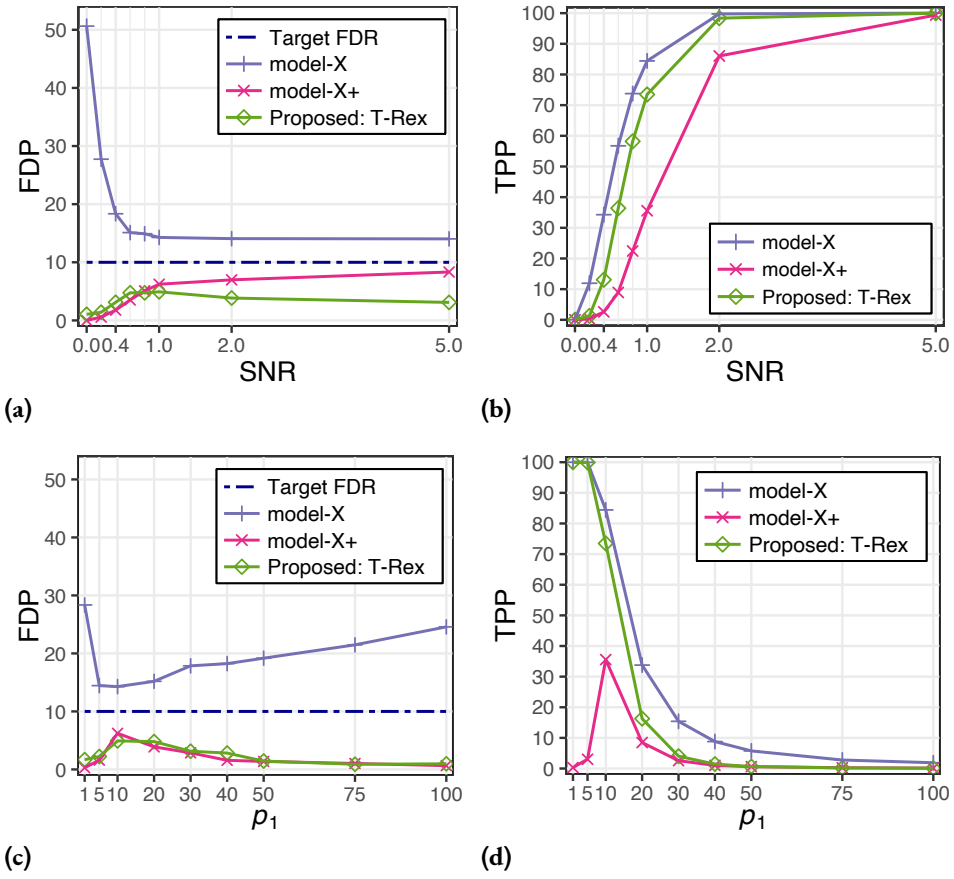


Figure 3.8: General: The *model-X* knockoff method fails to control the FDR. Among the FDR-controlling methods, the *T-Rex* selector outperforms the *model-X* knockoff+ method in terms of power. **Details**: **(a)** The *T-Rex* selector and the *model-X* knockoff+ method control the FDR at a target level of 10% for the whole range of SNR values while the *model-X* knockoff method fails to control the FDR and performs poorly at low SNR values. Setup: $n = 300$, $p = 1{,}000$, $p_1 = 10$, $T_{\max} = \lceil n/2 \rceil$, $L_{\max} = 10p$, $K = 20$, $MC = 955$. **(b)** As expected, the TPR (i.e., power) increases with respect to the SNR. It is remarkable that even though the FDR of the *T-Rex* selector lies below that of the *model-X* knockoff+ method for SNR values larger than 0.6, its power exceeds that of its strongest FDR-controlling competitor. The high power of the *model-X* knockoff method cannot be interpreted as an advantage, because the method does not control the FDR. Setup: Same as in Figure (a). **(c)** As in Figure (a), only the *T-Rex* selector and the *model-X* knockoff+ method control the FDR at a target level of 10%, whereas the *model-X* knockoff method always exceeds the target level. Setup: $n = 300$, $p = 1{,}000$, $T_{\max} = \lceil n/2 \rceil$, $L_{\max} = 10p$, $K = 20$, SNR $= 1$, $MC = 955$. **(d)** Among the FDR-controlling methods, the *T-Rex* selector has by far the highest power for sparse settings. The power of the *model-X* knockoff method exceeds that of the FDR-controlling methods, but this cannot be interpreted as an advantage of the method since it exceeds the target FDR level. Note that for an increasing number of active variables the power drops for all methods since apparently the number of data points $n = 300$ does not suffice in the simulated settings with a low sparsity level, i.e., settings with many active variables. Setup: Same as in Figure (c).

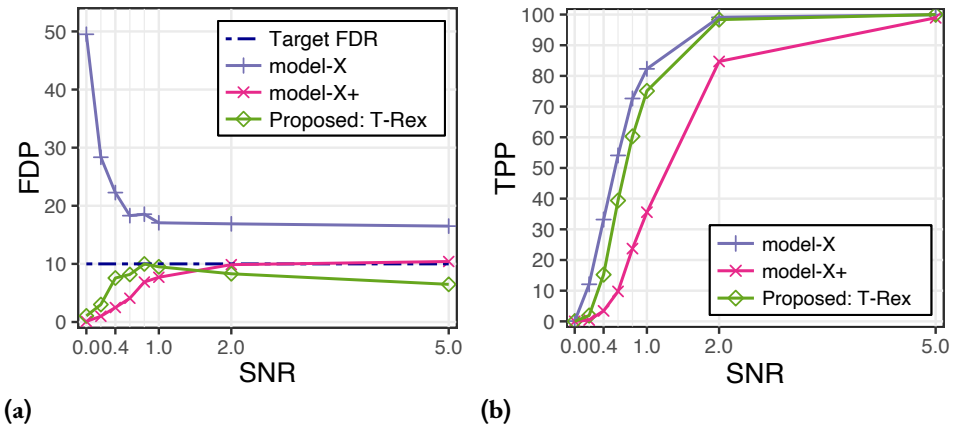


Figure 3.9: Average FDP and TPP in the case of dependent predictors: The *T-Rex* selector controls the FDR, has the highest power among the FDR-controlling methods, and reaches the almost highest possible TPR level at an SNR of 2 while the *model-X* knockoff+ method requires an SNR of 5 to reach the same TPR level. The *model-X* knockoff+ method also controls the FDR except for an SNR of 5, where it slightly exceeds the target FDR, and the *model-X* knockoff method does not control the FDR. The predictors were sampled from an autoregressive model of order one (AR(1)) with Gaussian noise and an autocorrelation coefficient $\rho = 0.5$. Setup: $n = 300, p = 1{,}000, p_1 = 10, T_{\text{max}} = \lceil n/2 \rceil, L_{\text{max}} = 10p, K = 20, MC = 955$.

that of its competitors (see Figure 3.1). The detailed descriptions and discussions of the simulation results are given in the captions of Figures 3.7, 3.8, and 3.9. Furthermore, Appendix A.10 discusses in more detail the robustness of the *T-Rex* selector in the presence of non-Gaussian heavy-tailed noise.

3.7 The Screen-T-Rex Selector

Genomics biobanks are information treasure troves with thousands of phenotypes (e.g., diseases, traits) and millions of single nucleotide polymorphisms (SNPs). Conducting reproducible GWAS for tens of thousands of phenotypes requires fast FDR-controlling variable selection methods such as the proposed *T-Rex* selector. Figure 3.1 shows that the proposed *T-Rex* selector is scalable to millions of variables in a reasonable computation time, while the state-of-the-art benchmark method is practically infeasible in such large-scale high-dimensional settings. Nevertheless, even the comparably low computation time of the *T-Rex* selector for one phenotype might become a burden when conducting GWAS for many phenotypes. Therefore, we propose the *Screen-T-Rex* selector, a fast version of the *T-Rex* selector. The proposed FDR-controlling method is suitable for conducting large-scale GWAS (with up to millions of SNPs) for tens of thousands of phenotypes. It does not ask

the user to set a target FDR level, but provides the user with an estimate of the achieved FDR. In the cases, where the user is not satisfied with the provided FDR estimate, the original *T-Rex* selector should be used with the result of the *Screen-T-Rex* selector and the desired target FDR as inputs. The proposed *Screen-T-Rex* selector has the following major innovations/advantages:

1. It provably controls the FDR at the self-estimated level (see Theorems 6 and 7 in Section 3.7.1).

2. It does not require the choice of any additional parameters (sparsity parameter, target FDR level, etc.).

3. Its computation time is approximately one order of magnitude lower than that of the original *T-Rex* selector and more than three orders of magnitude lower than that of the *model-X* knockoff methods in our simulations (see Table 6.2 in Chapter 6).

4. If, for some phenotypes, the user is not satisfied with the estimated FDR level, then the invested computation time to run the proposed *Screen-T-Rex* selector is not wasted because its computations can be reused by the original *T-Rex* selector to control the FDR at the desired target level.

In the following, the methodology and main theoretical FDR control results are presented and numerically verified and benchmarked against state-of-the-art methods. In Chapter 6, a simulated GWAS and a real-world HIV-1 drug resistance study demonstrate that the performance of the *Screen-T-Rex* selector is superior, and its computation time is multiple orders of magnitude lower compared to current benchmark knockoff methods.

3.7.1 Methodology and Main Results

Two versions of the *Screen-T-Rex* selector are proposed, the corresponding FDR control theorems are presented, and an algorithm for screening genomics biobanks is formulated.

3.7.1.1 Ordinary Screen-T-Rex Selector

While the original *T-Rex* selector determines v, T, and L such that the FDR is controlled at the user-defined target level, the *Screen-T-Rex* selector fixes $(v, T, L) = (0.5, 1, p)$. This is a special case of the original *T-Rex* selector that

1. is harnessed by the proposed *Screen-T-Rex* selector to determine an estimator of the FDR and
2. requires a much lower computation time than the original *T-Rex* selector and other benchmark methods (see Table 6.2 in Chapter 6).

The FDR estimator of the proposed ordinary *Screen-T-Rex* selector is given by

$$\widehat{\alpha} := \frac{1}{R_{1,p}(0.5) \vee 1}, \tag{3.41}$$

i.e., one divided by the number of selected variables. The intuition behind this estimator is as follows: $T = 1$ dummy variable is allowed to enter the solution paths of the random experiments before terminating the forward selection processes. So, in each random experiment one out of p dummies is included. Therefore, we expect, on average, no more than one out of at most p null variables to be included in each candidate set $\mathcal{C}_{k,L}(T)$, and, consequently, no more than one null variable among all selected variables. This idea is formalized in the following FDR control result:

Theorem 6 (FDR control - ordinary *Screen-T-Rex* selector)**.** *Define* $\widehat{\alpha} := 1/(R_{1,p}(0.5) \vee 1)$*, i.e., as in* (3.41)*. Then,*

$$\text{FDR}(0.5, 1, p) = \mathbb{E}[\text{FDP}(0.5, 1, p)] \leq \widehat{\alpha}, \tag{3.42}$$

i.e., the FDR is controlled at the estimated level $\widehat{\alpha}$*.*

Proof. With Definition 7, we obtain

$$\text{FDP}(0.5, 1, p) = \frac{V_{1,p}(0.5)}{R_{1,p}(0.5) \vee 1} = \widehat{\alpha} \cdot V_{1,p}(0.5). \tag{3.43}$$

Let $p = p_1 + p_0$, where p_1 and p_0 are the number of true active and null variables, respectively. Taking the expectation of (3.43) yields

$$\text{FDR}(0.5, 1, p) = \mathbb{E}[\text{FDP}(0.5, 1, p)] = \widehat{\alpha} \cdot \mathbb{E}\big[V_{1,p}(0.5)\big] \leq \widehat{\alpha} \cdot \frac{p_0}{p+1} \leq \widehat{\alpha}, \tag{3.44}$$

where the first inequality follows from $V_{1,p}(0.5)$ being stochastically dominated by the negative hypergeometric distribution $\text{NHG}(p_0 + p, p_0, 1)$, whose expected value is given by $p_0/(p+1)$ (see Appendix A.5). □

3.7.1.2 Confidence-Based Screen-T-Rex Selector

The above proposed ordinary *Screen-T-Rex* selector, as well as the original *T-Rex* selector, only considers the relative occurrences of the candidate variables in the selected active sets $\mathcal{C}_{k,L}(T)$ and disregards the original and dummy coefficient estimates, i.e.,

1. $\hat{\beta}_{j,k}(T, L)$, $j = 1, \ldots, p$, (i.e., coefficient estimate of the jth original variable in the kth random experiment and
2. $\hat{\beta}^{\circ}_{l,k}(T, L)$, $l = 1, \ldots, L$, (i.e., coefficient estimate of the lth dummy variable in the kth random experiment.

However, since the dummy variables act as flagged null variables, the coefficients of the dummies contain information about the distribution of the coefficients of the null variables. Therefore, we propose to harness the coefficient estimates of the dummies to construct a confidence interval

$$C(\gamma) \coloneqq [c_1(\gamma), c_2(\gamma)], \quad \gamma \in [0, 1], \tag{3.45}$$

where $c_1(\gamma)$ and $c_2(\gamma)$ are the lower and upper bound, respectively, and γ is the confidence level. The coefficient estimates of the null variables can also be expected to lie within the same confidence interval. Therefore, instead of selecting variables based on their relative occurrences, we replace $V_{1,p}(0.5)$ and $R_{1,p}(0.5)$ in Definition 7 and Theorem 6 by

$$V^{(C)}_{1,p}(\gamma) \coloneqq \left|\left\{\text{null } j : \overline{\hat{\beta}}_j(1, p) \notin C(\gamma)\right\}\right| \text{ and} \tag{3.46}$$

$$R^{(C)}_{1,p}(\gamma) \coloneqq \left|\widehat{\mathcal{A}}^{(C)}_p(\gamma, 1)\right| \coloneqq \left|\left\{j : \overline{\hat{\beta}}_j(1, p) \notin C(\gamma)\right\}\right|, \tag{3.47}$$

respectively, where $\overline{\hat{\beta}}_j(1, p) \coloneqq \frac{1}{K}\sum_{k=1}^{K} \hat{\beta}_{j,k}(1, p)$. That is, only candidate variables whose averaged (over K random experiments) coefficient estimates are not inside the confidence interval $C(\gamma)$ are selected.

We propose to construct the confidence interval in (3.45) using the non-parametric bootstrap with 1,000 resamples of the vector containing the $K = 20$ non-zero dummy coefficient estimates. Since, in all our simulations, the distribution of the bootstrapped standard errors of the averaged non-zero dummy coefficient estimates followed the standard normal distribution, we construct a normal bootstrap confidence interval (for details, see [ET94; DH97; ZI04]). In the following theorem, we state how the most liberal confidence level γ can be

determined such that the FDR is controlled at the estimated level by the confidence-based *Screen-T-Rex* selector:

Theorem 7 (FDR control - confidence-based *Screen-T-Rex* selector)**.** *Define* $\gamma := \inf\left\{\gamma' \in [0,1] \,:\, R_{1,p}^{(C)}(\gamma') \leq R_{1,p}(0.5)\right\}$ *and* $\widehat{\alpha}_C := 1/(R_{1,p}^{(C)}(\gamma) \vee 1)$*. Suppose that* $V_{1,p}^{(C)}(\gamma) \overset{d}{\leq} V_{1,p}(0.5)$*, where* $\overset{d}{\leq}$ *denotes stochastic dominance. Then,*

$$\mathrm{FDR}(\gamma, T=1, L=p) := \mathbb{E}[\mathrm{FDP}(\gamma, T=1, L=p)] := \mathbb{E}\left[\frac{V_{1,p}^{(C)}(\gamma)}{R_{1,p}^{(C)}(\gamma) \vee 1}\right] \leq \widehat{\alpha}_C. \tag{3.48}$$

Proof. With Definition 7 and Equations (3.46) and (3.47), we obtain

$$\mathrm{FDR}(\gamma, T=1, L=p) = \mathbb{E}[\mathrm{FDP}(\gamma, T=1, L=p)] = \mathbb{E}\left[\frac{V_{1,p}^{(C)}(\gamma)}{R_{1,p}^{(C)}(\gamma) \vee 1}\right] \tag{3.49}$$

$$= \widehat{\alpha}_C \cdot \mathbb{E}\Big[V_{1,p}^{(C)}(\gamma)\Big] \leq \widehat{\alpha}_C \cdot \mathbb{E}\big[V_{1,p}(0.5)\big] \leq \widehat{\alpha}_C, \tag{3.50}$$

where the first inequality follows from $V_{1,p}^{(C)}(\gamma) \overset{d}{\leq} V_{1,p}(0.5)$ and the second inequality is the same as in the proof of Theorem 6. □

3.7.1.3 Screening Genomics Biobanks

The *Screen-T-Rex* selector is intended to be used for screening thousands of phenotypes in large biobanks, while only using the original *T-Rex* selector in the cases where the estimated FDR is not acceptable to the user. Here, the user sets the target FDR for the original *T-Rex* selector and a lower and upper bound α_l and α_u, respectively, for the estimated FDRs by both versions of the *Screen-T-Rex* selector. The lower bound is required to avoid solutions at very low estimated FDRs, since these would yield a low power (i.e., TPR). Algorithm 4 summarizes the proposed work flow.

3.7.2 Numerical Experiments

We simulate a high-dimensional data setting according to the linear model in (2.1) with $n = 300$ samples and $p = 1{,}000$ predictors (i.e, candidate variables), and $p_1 = 10$ true active variables. The noise variance σ^2 is chosen such that the signal-to-noise-ratio

Algorithm 4 Screening Genomics Biobanks.

1. **Input**: α, α_l, and α_u.
2. For each considered phenotype in the biobank do:
   2.1. **Run** the *Screen-T-Rex* selector and obtain the estimated FDR levels $\widehat{\alpha}$ and $\widehat{\alpha}_C$.
   2.2. **Determine** the final set of selected variables $\widehat{\mathcal{A}}$ as follows:

$$\widehat{\mathcal{A}} \coloneqq \begin{cases} \widehat{\mathcal{A}}_p^{(C)}(\gamma, 1), & \alpha_l \leq \widehat{\alpha}_C \leq \alpha_u \,\& \\ & \max\{\widehat{\alpha}_C, \widehat{\alpha} \cdot I(\widehat{\alpha} \leq \alpha_u)\} = \widehat{\alpha}_C \\ \widehat{\mathcal{A}}_p(0.5, 1), & \alpha_l \leq \widehat{\alpha} \leq \alpha_u \,\& \\ & \max\{\widehat{\alpha}_C \cdot I(\widehat{\alpha}_C \leq \alpha_u), \widehat{\alpha}\} = \widehat{\alpha} \\ \varnothing, & \text{otherwise} \end{cases}, \tag{3.51}$$

   where $I(a \leq b)$, $a, b \in \mathbb{R}$, is the indicator function that has the value one if $a \leq b$ and zero otherwise. $\varnothing$ denotes the empty set. Convention: If $\widehat{\alpha} = \widehat{\alpha}_C$ and all conditions in the first two cases are satisfied, then $\widehat{\mathcal{A}} \coloneqq \widehat{\mathcal{A}}_p^{(C)}(\gamma, 1)$.
   2.3. If $\widehat{\mathcal{A}} = \varnothing$, **run** the *T-Rex* selector with target FDR α and **determine**

$$\widehat{\mathcal{A}} \coloneqq \widehat{\mathcal{A}}_L(v^*, T^*). \tag{3.52}$$

3. **Output**: Selected active set $\widehat{\mathcal{A}}$.

SNR $\coloneqq \text{Var}(\boldsymbol{X}\boldsymbol{\beta})/\text{Var}(\boldsymbol{\epsilon})$ takes on the values on the x-axes in Figure 3.10. Note that the FDP and TPP in Figure 3.10 are averaged over 955 Monte Carlo replications, respectively, and, therefore, are estimates of the FDR and TPR, respectively. A discussion of the results is provided within the caption of Figure 3.10.

3.8 Summary

The *T-Rex* selector, a new fast FDR-controlling variable selection framework for high-dimensional settings, has been proposed and benchmarked against existing methods in numerical simulations. The *T-Rex* selector is, to the best of our knowledge, the first multivariate high-dimensional FDR-controlling method that scales to millions of variables in a reasonable amount of computation time. Since the *T-Rex* random experiments can be computed in parallel, multicore computers allow for additional substantial savings in computation time. These properties make the *T-Rex* selector a suitable method especially for

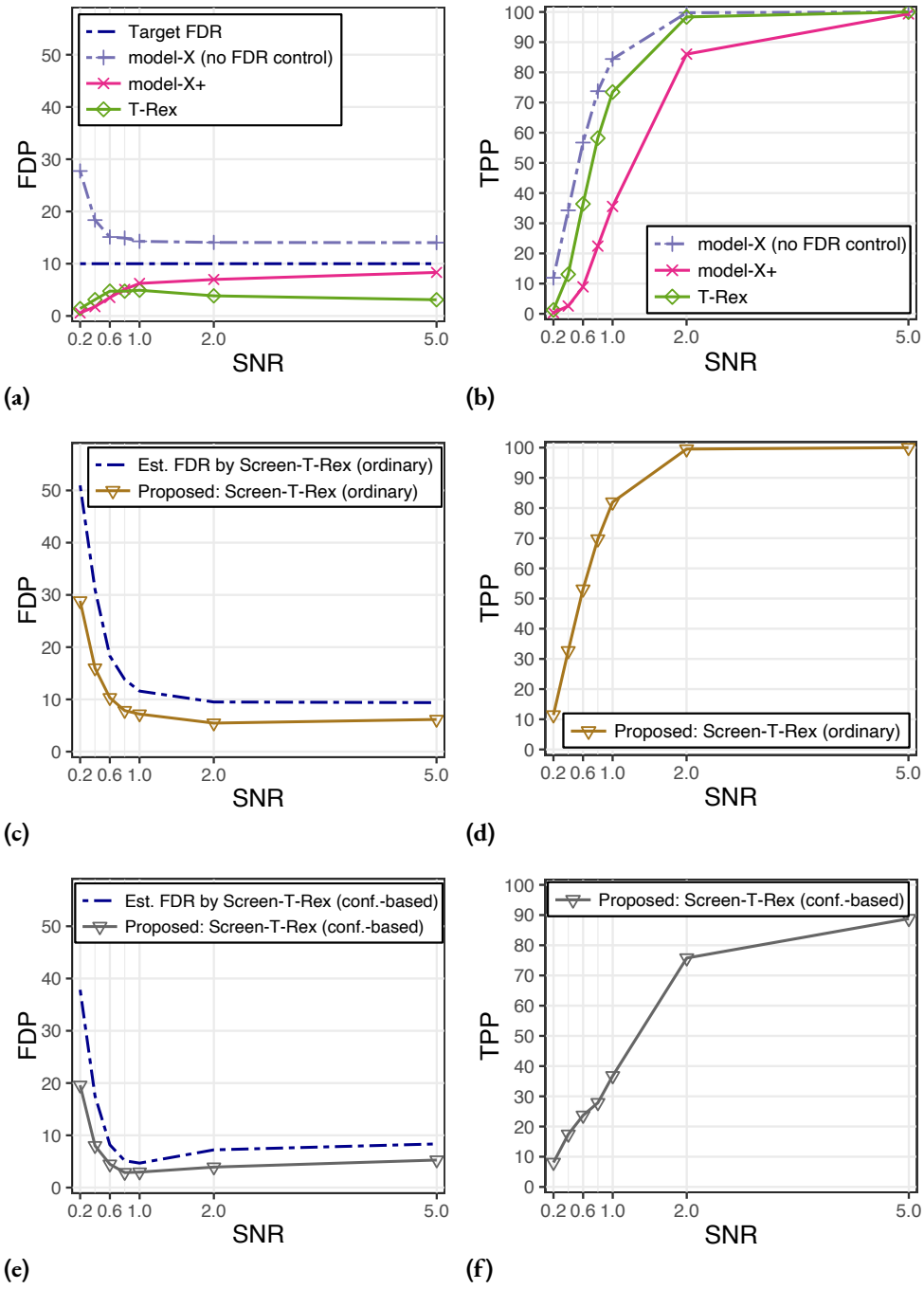


Figure 3.10: Figures (a) and (b) display the results of the benchmark methods. We observe that the *T-Rex* selector and the *model-X* knockoff+ method control the FDR at the target level of 10%, while the *model-X* knockoff method fails to control the FDR. Figures (c) - (f) show that the proposed ordinary and the confidence-based *Screen-T-Rex* selector both control the FDR at the self-estimated levels while achieving a reasonably high TPR. The confidence-based version is capable of controlling the FDR at lower levels than the ordinary version but, in turn, achieves a lower TPR.

large-scale GWAS.

The *Screen-T-Rex* selector, a fast FDR-controlling variable selection method for screening through large-scale genomics biobanks, has been proposed. Using the *Screen-T-Rex* selector in combination with the original *T-Rex* selector, an efficient algorithm for conducting thousands of large-scale GWAS has been proposed.

In Chapter 6, the proposed *T-Rex* and *Screen-T-Rex* selectors have been used to solve real-world data applications in biomedical engineering.

After climbing a great hill, one only finds that there are many more hills to climb.

Nelson Mandela

4

Dependency-Aware High-Dimensional False Discovery Rate Control

In this chapter, the proposed *T-Rex* framework in Chapter 3 is extended to account for groups of highly dependent candidate variables. The developed dependency-aware *T-Rex* (*T-Rex+DA*) selector allows to perform FDR-controlled high-dimensional variable selection in the presence of various types of dependencies among the candidate variables. In these highly correlated settings, which are characteristic for, e.g., genomics and stock returns data, the state-of-the-art benchmark methods loose the FDR control property while the proposed *T-Rex+DA* selector maintains FDR control.

Section 4.1 motivates the proposed *T-Rex+DA* selector and provides an overview of the major contributions. In Section 4.2 the methodology, the main theoretical results, and the calibration algorithm of the proposed *T-Rex+DA* selector are presented. In Section 4.3, the theoretical results are numerically verified and the performance of the *T-Rex+DA* selector is compared against state-of-the-art benchmark methods via numerical experiments. In Section 4.4, the *T-Rex+DA* selector is adapted for advanced dependency models that are especially relevant in financial index tracking. Section 4.5 summarizes this chapter.

In Chapters 6 and 7, the proposed methods have been used to solve challenging real-world

data problems in biomedical and financial engineering. Technical proofs and additional simulations are deferred to Appendix B.

The content of this chapter is based on the publications [MMPewb], [MPMew], and [MMP24]. The implementations of the developed methods are included in the open source R software packages (with C++ backend) TRexSelector [Mac+24c] and tlars [Mac+24b] on CRAN.

4.1 The T-Rex +DA Selector: Motivation and Major Contributions

This section motivates the proposed *T-Rex+DA* selector through a gene expression survival analysis use-case and provides an overview of the major contributions.

4.1.1 Motivation

An important use-case of this work consists in detecting the few genes that are truly associated with the survival time of patients diagnosed with a certain type of cancer [TCW15; ŐLG21; ABG08]. The expression levels of the detected genes are then classified into low- and high-expressing genes, which allows cancer researchers to make statements such as: "The median survival time of breast cancer patients with a high expression of gene A and a low expression of gene B is 10 years higher than the survival time of patients with a low expression of gene A and a high expression of gene B." Such information is invaluable for the development of new therapies and personalized medicine [KA17]. However, the development and clinical trial of new drugs is costly and resources are limited. Therefore, it is crucial to select as many as possible of the few reproducible genes that are truly associated with the survival time of cancer patients while keeping the number of false discoveries (i.e., irrelevant genes) low. This aim is in line with false discovery rate (FDR) controlling methods.

Unfortunately, however, both the *model-X* knockoff methods and the *T-Rex* methods fail to control the FDR reliably in the presence of groups of highly dependent variables, which are characteristic for, e.g., gene expression [SDC03], genomics [Bal06], and stock returns data [MPMew].

In order to reduce the dependencies among the candidate variables, pruning approaches have been used [Can+18; SSC19; MMPewa]. In general, pruning methods cluster highly depen-

dent variables into groups, select a representative variable for each group, and run the FDR-controlling method on the set of representatives. This approach is suitable for genome-wide association studies (GWAS) based on large-scale high-dimensional genomics data from large biobanks [Bun+19; Sud+15], where the goal is to detect the groups of highly correlated single nucleotide polymorphisms (SNPs) that are associated with a disease of interest and not the specific SNPs. However, pruning methods are not applicable in gene expression analysis and other applications where it is crucial to detect specific genes or other variables.

4.1.2 Major Contributions

To address use-cases like the one discussed in Section 4.1.1, we propose a new FDR-controlling and dependency-aware *T-Rex* (*T-Rex+DA*) framework that provably controls the FDR at the user-specified target level. This is achieved and verified through the following theoretical contributions, numerical validations, and real world experiments:

1. A dependency-capturing graphical model is incorporated into the *T-Rex* framework and is used to capture and leverage the dependency structure among variables to develop a variable penalization mechanism that allows for provable FDR control.

2. Using martingale theory (see Section 2.3), we prove that the proposed approach controls the FDR (Theorem 9).

3. We begin with hierarchical graphical models (i.e., binary trees) and then extend the proposed framework by stating and proving a comprehensible condition that must be satisfied for the design of graphical and non-graphical dependency-capturing models to be eligible for being incorporated into the *T-Rex* framework (Theorem 10).

4. We develop a fully integrated optimal calibration algorithm that simultaneously determines the parameters of the incorporated graphical model and of the *T-Rex* framework, such that the FDR is controlled while maximizing the number of selected variables (Theorem 11).

5. Numerical experiments and a real-world breast cancer survival analysis (see Chapter 6) verify the theoretical results and demonstrate the practical usefulness of the proposed framework.

4.2 Methodology and Main Theoretical Results

In this section, the proposed FDR-controlling *T-Rex+DA* framework for general dependency structures is introduced. First, a dependency-capturing graph model is incorporated into the *T-Rex+DA* framework. Second, we prove that the considered group design yields FDR control. Third, we formulate a sufficient group design condition for graphical as well as non-graphical models that can be used as a guiding principle for other application-specific group designs. Finally, the optimal dependency-aware calibration algorithm for the *T-Rex+DA* selector is presented.

4.2.1 Preliminaries

Before the proposed *T-Rex+DA* selector is presented and in order to understand why the ordinary *T-Rex* selector might loose the FDR control property in the presence of highly correlated variables, we establish an interesting relationship between the pairwise relative occurrences of two candidate variables and the correlation coefficient between them. For example, let the *Lasso* in Section 2.1.1 be used to perform the forward variable selection in each random experiment. Within the *T-Rex* framework and for the kth random experiment, the *Lasso* estimator is defined by

$$\hat{\boldsymbol{\beta}}_k(\lambda_k(T,L)) = \underset{\boldsymbol{\beta}_k}{\arg\min}\, \frac{1}{2}\big\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}_k\boldsymbol{\beta}_k\big\|_2^2 + \lambda_k(T,L)\cdot\|\boldsymbol{\beta}_k\|_1, \tag{4.1}$$

where $\lambda_k(T,L) > 0$ is the sparsity parameter that corresponds to the change point in the kth random experiment after T dummies have been included. With these definitions in place, we can formulate the following theorem:

Theorem 8 (Absolute difference of relative occurrences)**.** *Let* $\rho_{j,j'} := \boldsymbol{x}_j^\top\boldsymbol{x}_{j'}$, $j,j' \in \{1,\dots,p\}$*, be the sample correlation coefficient of the standardized variables* j *and* j'*. Suppose that* $\hat{\beta}_{j,k}, \hat{\beta}_{j',k} \neq 0$*. Then, for all tuples* $(T,L) \in \{1,\dots,L\}\times\mathbb{N}_+$ *it holds that*

$$\big|\Phi_{T,L}(j) - \Phi_{T,L}(j')\big| \leq \overline{\Lambda}\|\boldsymbol{y}\|_2\cdot\sqrt{2(1-\rho_{j,j'})}, \tag{4.2}$$

where $\overline{\Lambda} := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\lambda_k(T,L)}$.

Proof. The proof is deferred to Appendix B.1.2. □

4.2.2 The Dependency-Aware T-Rex Selector

From Theorem 8, we know that the pairwise absolute differences between the relative occurrences are bounded and the differences are zero when the corresponding variables are perfectly correlated. That is, even if only one of the variables from the pair of highly correlated variables is a true active variable, both variables might be selected. This is the harmful behavior that leads to the loss of the FDR control property in the presence of highly correlated variables. Loosely speaking, if a candidate variable is highly correlated with another candidate variable and has a similar relative occurrence, then even high relative occurrences are no evidence for that variable being a true active one. Therefore, for such types of data, we propose to replace the ordinary relative occurrences of the *T-Rex* selector $\Phi_{T,L}(j)$ by the dependency-aware relative occurrences $\Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}})$, $j = 1, \ldots, p$, which are defined as follows:

Definition 12 (Dependency-aware relative occurrences)**.** *The dependency-aware relative occurrence of variable $j \in \{1, \ldots, p\}$ is defined by*

$$\Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}}) \coloneqq \Psi_{T,L}(j, \rho_{\text{thr}}) \cdot \Phi_{T,L}(j), \tag{4.3}$$

where

$$\Psi_{T,L}(j, \rho_{\text{thr}}) \coloneqq \begin{cases} \dfrac{1}{2 \;-\; \min\limits_{j' \in \text{Gr}(j, \rho_{\text{thr}})} \left\{ \left| \Phi_{T,L}(j) - \Phi_{T,L}(j') \right| \right\}}, & \text{Gr}(j, \rho_{\text{thr}}) \neq \varnothing \\ 1/2, & \text{Gr}(j, \rho_{\text{thr}}) = \varnothing \end{cases}, \tag{4.4}$$

with $\Psi_{T,L}(j, \rho_{\text{thr}}) \in [0.5, 1]$ being a penalty factor,

$$\text{Gr}(j, \rho_{\text{thr}}) \subseteq \{1, \ldots, p\} \backslash \{j\} \tag{4.5}$$

denoting the generic definition of the group of variables that are associated with variable j, and $\rho_{\text{thr}} \in [0, 1]$ denoting a parameter that determines the size of the variable groups.

In words, the dependency-aware relative occurrence of variable j is designed to penalize the ordinary relative occurrence of variable j according to its resemblance with the relative occurrences of its associated group of variables $\text{Gr}(j, \rho_{\text{thr}})$.

From Definition 12, we can infer that the selected active set of the proposed *T-Rex+DA* selector is a subset of the selected active set of the ordinary *T-Rex* selector in (3.3):

Corollary 1. *Let $\widehat{\mathcal{A}}_L(v, T) \coloneqq \{j : \Phi_{T,L}(j) > v\}$ and $\widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}, T) \coloneqq \{j :$*

$\Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}}) > v\}$ *be the selected active sets of the ordinary T-Rex selector and the T-Rex+DA selector, respectively. Then, it holds that*

$$\widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}, T) \subseteq \widehat{\mathcal{A}}_L(v, T). \tag{4.6}$$

Proof. Using the definition of $\Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}})$ in (4.3), we obtain

$$\begin{aligned} \widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}, T) &= \{j : \Psi_{T,L}(j, \rho_{\text{thr}}) \cdot \Phi_{T,L}(j) > v\} && (4.7)\\ &\subseteq \{j : \Phi_{T,L}(j) > v\} && (4.8)\\ &= \widehat{\mathcal{A}}_L(v, T), && (4.9) \end{aligned}$$

where the second line follows from $\Psi_{T,L}(j, \rho_{\text{thr}}) \leq 1$. □

Loosely speaking, Corollary 1 indicates that the effect of replacing $\Phi_{T,L}(j)$ by $\Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}})$ is that highly correlated variables, for which there is not sufficient evidence to decide if they are active, are removed from the selected active set.

In order to particularize the *T-Rex+DA* selector for different dependency structures among the candidate variables, only the generic definition of the variable groups $\text{Gr}(j, \rho_{\text{thr}})$ in (4.5) has to be specified. Therefore, in the following, we present a rigorous methodology for the design of $\text{Gr}(j, \rho_{\text{thr}})$ such that the FDR is provably controlled at the user-defined target level α while maximizing the number of selected variables and, thus, implicitly maximizing the TPR.

4.2.3 Clustering Variables via Hierarchical Graphical Models

In the following, we specify $\text{Gr}(j, \rho_{\text{thr}})$, $j = 1, \ldots, p$, using a hierarchical graphical model. That is, the variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are clustered in a recursive fashion according to some measure of distance. The resulting binary tree or dendrogram is a structured graph that allows for different distance cutoff values that partition the set of variables. Figure 4.1 depicts such a dendrogram for $p = 6$ variables, where the height of the "⊓"-shaped connector of any two clusters represents the distance of the two connected clusters. At the bottom of the dendrogram, all variables are considered as one-element clusters. Then, starting at the bottom, in each iteration the two clusters with the smallest distance are connected until all variables are clustered into a single cluster at the top. The obtained dendrogram can be evaluated at different distances (i.e., values on the y-axis), resulting in different variable clusters. The p discrete distances between two consecutive cutoff levels that invoke a change in the clusters, are de-

noted by $\Delta\rho_{\text{thr},u}$, $u = 1, \ldots, p$. For example, cutting off the dendrogram in Figure 4.1 at a distance of

$$1 - \rho_{\text{thr}}(u_c = 2) \coloneqq 1 - \sum_{u=1}^{u_c=2} \Delta\rho_{\text{thr},u} \tag{4.10}$$

$$= 1 - (0.1 + 0.3) = 0.6, \tag{4.11}$$

where $u_c \in \{1, \ldots, p\}$ is the discrete cutoff level, yields three disjoint variable clusters: $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, $\{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, and $\{\boldsymbol{x}_6\}$.

With this generic description of hierarchical graphical models in place, we can specify the generic definition of the variable groups in (4.5) in a recursive fashion:

Definition 13 (Hierarchical group design)**.** *The jth variable group following a hierarchical graphical model (i.e., binary tree/dendrogram) is defined by*

$$\text{Gr}(j, \rho_{\text{thr}}(u_c)) \coloneqq \big\{ j' \in \{1, \ldots, p\} \backslash \{j\} : \tag{4.12}$$

$$\text{dist}_{u_c - 1}(j, j') \in [1 - \rho_{\text{thr}}(u_c), 1 - \rho_{\text{thr}}(u_c - 1)] \big\}, \tag{4.13}$$

where $\text{dist}_{u_c-1}(j, j')$ *is a still to be specified measure of distance between the groups* $\text{Gr}(j, \rho_{\text{thr}}(u_c - 1))$ *and* $\text{Gr}(j', \rho_{\text{thr}}(u_c - 1))$.

Note that in this recursive definition of the variable groups, we consider $\rho_{\text{thr}}(u_c)$ to be a variable that can be optimized and, therefore, include it in $\text{Gr}(j, \rho_{\text{thr}}(u_c))$ as a second argument.

Remark 8. *The following three distance measures are frequently used in hierarchical graphical models [MC12]:*

1. *Single linkage:*

$$\text{dist}_{u_c}(g, h) \coloneqq \min_{\substack{g' \in \text{Gr}(g, \rho_{\text{thr}}(u_c)) \\ h' \in \text{Gr}(h, \rho_{\text{thr}}(u_c))}} 1 - |\rho_{g',h'}|, \tag{4.14}$$

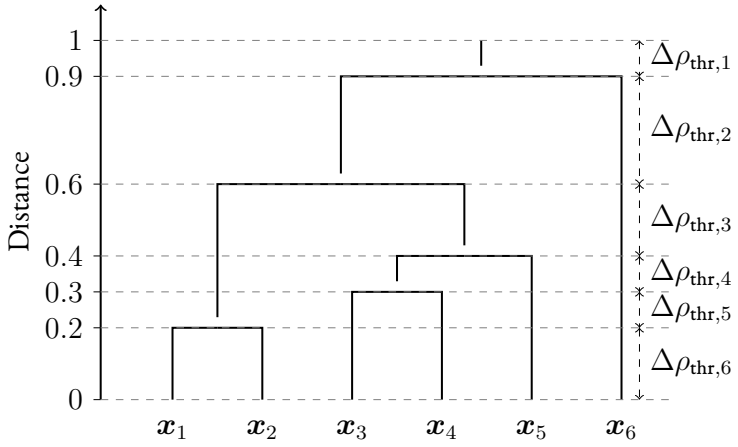


Figure 4.1: Hierarchical graphical models: The dendrogram.

2. *Complete linkage:*

$$\text{dist}_{u_c}(g,h) := \max_{\substack{g' \in \text{Gr}(g,\rho_{\text{thr}}(u_c)) \\ h' \in \text{Gr}(h,\rho_{\text{thr}}(u_c))}} 1 - |\rho_{g',h'}|, \tag{4.15}$$

3. *Average linkage:*

$$\text{dist}_{u_c}(g,h) := \frac{\sum_{g' \in \text{Gr}(g,\rho_{\text{thr}}(u_c))} \sum_{h' \in \text{Gr}(h,\rho_{\text{thr}}(u_c))} \left(1 - |\rho_{g',h'}|\right)}{|\,\text{Gr}(g,\rho_{\text{thr}}(u_c))| \cdot |\,\text{Gr}(h,\rho_{\text{thr}}(u_c))|}. \tag{4.16}$$

Remark 9. *Note that, for all $u_c \in \{1,\dots,p\}$, it holds that*

$$\text{Gr}(j_1,\rho_{\text{thr}}(u_c)) \cap \text{Gr}(j_2,\rho_{\text{thr}}(u_c)) = \tag{4.17}$$

$$\begin{cases} \varnothing, & \nexists\, j \in \{1,\dots,p\} : j_1, j_2 \in \text{Gr}(j,\rho_{\text{thr}}(u_c)) \\ \text{Gr}(j_1,\rho_{\text{thr}}(u_c)) \cup \text{Gr}(j_2,\rho_{\text{thr}}(u_c)), & \textit{otherwise} \end{cases}, \tag{4.18}$$

i.e., any arbitrary pair of groups among the obtained p groups $\text{Gr}(j,\rho_{\text{thr}}(u_c))$, $j = 1,\dots,p$, are disjoint if and only if there exist no two variables j_1 and j_2 that belong to the same group and

identical otherwise. Loosely speaking, due to the binary tree structure of hierarchical graphical models, there exist no "overlapping" variable groups.

4.2.4 Preliminaries for the FDR Control Theorem

Based on the recursive definition of the variable groups in (4.12), we can formulate the FDR, TPR, and the conservative FDP estimator $\widehat{\text{FDP}}$. For this purpose, let

$$V_{T,L}(v, \rho_{\text{thr}}(u_c)) \coloneqq \big|\widehat{\mathcal{A}}^0(v, \rho_{\text{thr}}(u_c))\big| \tag{4.19}$$
$$\coloneqq \big|\{\text{null } j : \Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}}(u_c)) > v\}\big|, \tag{4.20}$$

$$S_{T,L}(v, \rho_{\text{thr}}(u_c)) \coloneqq \big|\widehat{\mathcal{A}}^1(v, \rho_{\text{thr}}(u_c))\big| \tag{4.21}$$
$$\coloneqq \big|\{\text{active } j : \Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}}(u_c)) > v\}\big|, \tag{4.22}$$

$$R_{T,L}(v, \rho_{\text{thr}}(u_c)) \coloneqq \big|\widehat{\mathcal{A}}(v, \rho_{\text{thr}}(u_c))\big| \tag{4.23}$$
$$\coloneqq \big|\{j : \Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}}(u_c)) > v\}\big|, \tag{4.24}$$

be the number of selected null variables, the number of selected active variables, and the total number of selected variables, respectively. Note that the expressions $\widehat{\mathcal{A}}^0(v, \rho_{\text{thr}}(u_c))$, $\widehat{\mathcal{A}}^1(v, \rho_{\text{thr}}(u_c))$, and $\widehat{\mathcal{A}}(v, \rho_{\text{thr}}(u_c))$ are shortcuts (i.e., L and T are dropped) of the expressions $\widehat{\mathcal{A}}^0_L(v, \rho_{\text{thr}}(u_c), T)$, $\widehat{\mathcal{A}}^1_L(v, \rho_{\text{thr}}(u_c), T)$, and $\widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}(u_c), T)$, respectively.

Definition 14 (Dependency-aware FDP and FDR)**.** *The dependency-aware FDR is defined as the expectation of the dependency-aware FDP, i.e.,*

$$\text{FDR}(v, \rho_{\text{thr}}(u_c), T, L) \coloneqq \mathbb{E}\big[\,\text{FDP}(v, \rho_{\text{thr}}(u_c), T, L)\big] \tag{4.25}$$
$$\coloneqq \mathbb{E}\left[\frac{V_{T,L}(v, \rho_{\text{thr}}(u_c))}{R_{T,L}(v, \rho_{\text{thr}}(u_c)) \vee 1}\right]. \tag{4.26}$$

Definition 15 (Dependency-aware TPP and TPR)**.** *The dependency-aware TPR is defined as the expectation of the dependency-aware TPP, i.e.,*

$$\text{TPR}(v, \rho_{\text{thr}}(u_c), T, L) \coloneqq \mathbb{E}\big[\,\text{TPP}(v, \rho_{\text{thr}}(u_c), T, L)\big] \tag{4.27}$$
$$\coloneqq \mathbb{E}\left[\frac{S_{T,L}(v, \rho_{\text{thr}}(u_c))}{p_1 \vee 1}\right]. \tag{4.28}$$

From Definition 14, we know that in order to design a dependency-aware and conservative FDP estimator, we only need to design a dependency-aware estimator of the number of selected null variables $V_{T,L}(v, \rho_{\text{thr}}(u_c))$, since the total number of selected variables $R_{T,L}(v, \rho_{\text{thr}}(u_c))$ is observable. For this purpose, we plug the dependency-aware relative occurrences from Definition 12 and the group design from Definition 13 into the ordinary *T-Rex* estimator of $V_{T,L}(v)$ in Definition 10, which yields the dependency-aware estimator of the number of selected null variables

$$\widehat{V}_{T,L}(v, \rho_{\text{thr}}(u_c)) \coloneqq \sum_{j \in \widehat{\mathcal{A}}(v, \rho_{\text{thr}}(u_c))} \left(1 - \Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}}(u_c))\right) + \widehat{V}'_{T,L}(v, \rho_{\text{thr}}(u_c)), \tag{4.29}$$

where

$$\widehat{V}'_{T,L}(v, \rho_{\text{thr}}(u_c)) \coloneqq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi^{\text{DA}}_{t,L}(q, \rho_{\text{thr}}(u_c))}{L - (t-1)} \cdot \frac{\sum_{j \in \widehat{\mathcal{A}}(v, \rho_{\text{thr}}(u_c))} \Delta\Phi^{\text{DA}}_{t,L}(j, \rho_{\text{thr}}(u_c))}{\sum_{j \in \widehat{\mathcal{A}}(0.5, \rho_{\text{thr}}(u_c))} \Delta\Phi^{\text{DA}}_{t,L}(j, \rho_{\text{thr}}(u_c))} \tag{4.30}$$

and $\Delta\Phi^{\text{DA}}_{t,L}(j, \rho_{\text{thr}}) \coloneqq \Phi^{\text{DA}}_{t,L}(j, \rho_{\text{thr}}) - \Phi^{\text{DA}}_{t-1,L}(j, \rho_{\text{thr}})$ is the increase in the dependency-aware relative occurrence from step $t-1$ to t. The expressions in (4.29) and (4.30) are derived along the lines of the ordinary estimator of $V_{T,L}(v)$ in (3.18) within Definition 10 except that the ordinary relative occurrences in Definition 5 have been replaced by the proposed dependency-aware relative occurrences in Definition 12.

Finally, analogous to Definition 10 that introduces the FDP estimator for the *T-Rex* selector, the conservative dependency-aware estimator of the FDP is defined as follows:

Definition 16 (Dependency-aware FDP estimator)**.** *The dependency-aware FDP estimator is defined by*

$$\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \coloneqq \frac{\widehat{V}_{T,L}(v, \rho_{\text{thr}}(u_c))}{R_{T,L}(v, \rho_{\text{thr}}(u_c)) \vee 1}. \tag{4.31}$$

With all preliminary definitions in place, the overarching goal of this work, i.e., maximizing the number of selected variables while controlling the FDR at the target level α, is formulated as follows:

$$\underset{v, \rho_{\text{thr}}(u_c), T}{\text{maximize}}\, R_{T,L}(v, \rho_{\text{thr}}(u_c)) \qquad \text{subject to} \quad \widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \leq \alpha. \tag{4.32}$$

In Section 4.2.5, we prove that satisfying the condition of the optimization problem in (4.32) yields FDR control and in Section 4.2.7, we propose an efficient algorithm to solve (4.32) and prove that it yields an optimal solution.

4.2.5 Dependency-Aware FDR Control

In this section, using martingale theory (see Section 2.3), we state and prove that controlling $\widehat{\mathrm{FDP}}(v, \rho_{\mathrm{thr}}(u_c), T, L)$ at the target level α (i.e., the condition in the optimization problem in (4.32)) guarantees FDR control.

Theorem 9 (Dependency-aware FDR control)**.** *For all quadruples* $(T, L, \rho_{\mathrm{thr}}(u_c), v) \in \{1, \dots, L\} \times \mathbb{N}_+ \times [0, 1] \times [0.5, 1)$ *that satisfy the equation*

$$v = \inf\big\{\nu \in [0.5, 1) : \widehat{\mathrm{FDP}}(\nu, \rho_{\mathrm{thr}}(u_c), T, L) \leq \alpha\big\}, \tag{4.33}$$

and as $K \to \infty$*, the T-Rex+DA selector with* $\mathrm{Gr}(j, \rho_{\mathrm{thr}}(u_c))$ *from Definition 13 controls the FDR at any fixed target level* $\alpha \in [0, 1]$*, i.e.,*

$$\mathrm{FDR}(v, \rho_{\mathrm{thr}}(u_c), T, L) \leq \alpha. \tag{4.34}$$

Proof. The proof is deferred to Appendix B.1.3. □

4.2.6 General Group Design Principle

The *T-Rex+DA* selector is not restricted to the considered binary tree graphs or dendrograms but also applies for various other dependency models. In fact, a closer look at Lemmas 8 and 9 in Appendix B.1.7, which are essential for the proof of Theorem 9 (dependency-aware FDR control), reveals that the following general design principle for the groups $\mathrm{Gr}(j, \rho_{\mathrm{thr}})$ can be derived from Lemmas 8 and 9:

Theorem 10 (Group design principle)**.** *Consider the generic definition of the variable groups in Definition 12, i.e.,* $\mathrm{Gr}(j, \rho_{\mathrm{thr}}) \in \{1, \dots, p\} \backslash \{j\}$*,* $j = 1, \dots, p$*,* $\rho_{\mathrm{thr}} \in [0, 1]$*. If any* ρ_1*,* $\rho_2 \in [0, 1]$*,* $\rho_2 > \rho_1$*, satisfy*

$$\mathrm{Gr}(j, \rho_2) \subseteq \mathrm{Gr}(j, \rho_1), \quad j = 1, \dots, p, \tag{4.35}$$

then the T-Rex+DA selector controls the FDR at the target level $\alpha \in [0, 1]$*.*

Proof. The proof is deferred to Appendix B.1.4. □

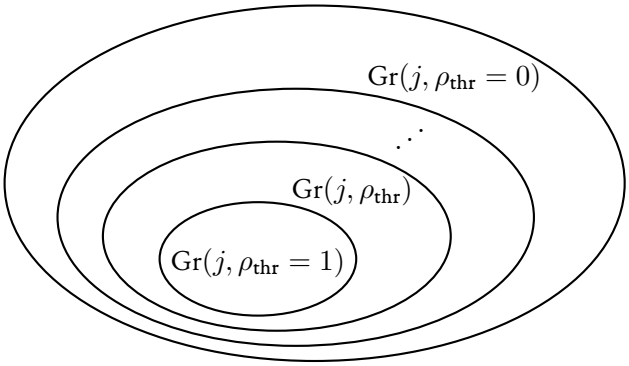


Figure 4.2: Illustration of the group design principle for dependency-aware FDR control in Theorem 10.

Loosely speaking, Theorem 10 states that the cardinalities of any variable group j must be monotonically decreasing in ρ_{thr} and follow the subset structure illustrated in Figure 4.2. Thus, any dependency model (e.g., graph models, time series models, equicorrelated models, etc.) that follows the design principle in Theorem 10 can be incorporated into the *T-Rex+DA* selector. This property makes the *T-Rex+DA* selector a versatile FDR-controlling method that can cope with various dependency models. In conclusion, it is not necessary to redo the FDR control proof for new group designs. Instead, one simply has to check whether that group design follows the group design principle in Theorem 10.

4.2.7 Optimal Dependency-Aware T-Rex Calibration Algorithm

In this section, we propose an efficient calibration algorithm and prove that it yields an optimal solution of (4.32). That is, it optimally calibrates the parameters v, $\rho_{\text{thr}}(u_c)$, and T of the proposed *T-Rex+DA* selector, such that the FDR is controlled at the target level while maximizing the number of selected variables.The pseudocode of the proposed *T-Rex+DA* calibration algorithm is given in Algorithm 5. An open source implementation of the proposed calibration algorithm for the *T-Rex+DA* selector is available within the R package 'TRexSelector' on CRAN [Mac+24c].

First, some hyperparameters, which are relevant for managing the tradeoff between achieving a high TPR, the memory consumption, and computation time but have no influence on the FDR control property of the proposed method, are set. Throughout this dissertation, the hyperparameters are chosen as suggested and discussed in Section 3.5.4 and Appendix A.3 to be $\tilde{v} = 0.75$, $\tilde{\rho}_{\text{thr}} = \rho_{\text{thr}}(\lfloor 0.75 \cdot p \rceil)$, $L_{\max} = 10p$, $T_{\max} = \lceil n/2 \rceil$, where $\lfloor \cdot \rceil$ and $\lceil \cdot \rceil$ denote rounding towards the nearest integer and the nearest higher integer, respectively. Moreover,

Algorithm 5 Extended *T-Rex+DA* Calibration.

1. **Input**: $\alpha \in [0,1]$, $\boldsymbol{X}$, $\boldsymbol{y}$, K, $\tilde{v}$, $\tilde{\rho}_{\text{thr}}$, $L_{\max}$, $T_{\max}$.
2. **Set** $L = p, T = 1$.
3. **While** $\widehat{\text{FDP}}(v = \tilde{v}, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T, L) > \alpha$ and $L \leq L_{\max}$ **do**:
   Set $L \leftarrow L + p$.
4. **Set** $\Delta v = \dfrac{1}{K}$, $\widehat{\text{FDP}}(v = 1 - \Delta v, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T, L) = 0$.
5. **While** $\widehat{\text{FDP}}(v = 1 - \Delta v, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T, L) \leq \alpha$ and $T \leq T_{\max}$ **do**:
   5.1. **For** $v = 0.5, 0.5 + \Delta v, 0.5 + 2 \cdot \Delta v, \ldots, 1 - \Delta v$ **do**:
      5.1.1. **For** $u_c = 1, \ldots, p$ **do**:
         i. **Compute** $\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L)$ as in Def. 16.
         ii. **If** $\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \leq \alpha$
            Compute $\widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}(u_c), T)$ as in (4.36).
         Else
            Set $\widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}(u_c), T) = \varnothing$.
   5.2. **Set** $T \leftarrow T + 1$.
6. **Solve**

$$\begin{aligned} \max_{v', \rho_{\text{thr}}(u_c'), T'} & \left| \widehat{\mathcal{A}}_L(v', \rho_{\text{thr}}(u_c'), T') \right| \\ \text{s.t.} \quad & T' \in \{1, \ldots, T-1\} \\ & u_c' \in \{1, \ldots, p\} \\ & v' \in \{0.5, 0.5 + \Delta v, 0.5 + 2\Delta v, \ldots, 1 - \Delta v\} \end{aligned}$$

   and let $(v^*, \rho_{\text{thr}}(u_c^*), T^*)$ be a solution.
7. **Output**: $(v^*, \rho_{\text{thr}}(u_c^*), T^*, L)$ and $\widehat{\mathcal{A}}_L(v^*, \rho_{\text{thr}}(u_c^*), T^*)$.

it was shown for various applications and in extensive simulations that there are no improvements for $K > 20$ random experiments [MMPewa; MMP22; MMP23a] and, therefore, we set $K = 20$.

The algorithm proceeds as follows: It takes the user-defined target FDR, the original predictor matrix $\boldsymbol{X}$, and the response vector $\boldsymbol{y}$ as inputs. Then, it determines L via a loop that

adds p dummies in each iteration until $\widehat{\text{FDP}}(v = \tilde{v}, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T = 1, L)$ falls below the target FDR level α at a reference point $(v, \rho_{\text{thr}}(u_c), T) = (\tilde{v}, \tilde{\rho}_{\text{thr}}, 1)$ or L reaches the maximum allowed value L_{max}. This guarantees that the FDR is controlled as tightly as possible at the target FDR level α while ensuring that the TPR is as high as possible. Supposing that there exists no $\tilde{\rho}'_{\text{thr}} \neq \tilde{\rho}_{\text{thr}}$ that satisfies $\widehat{\text{FDP}}(v = 1 - \Delta v, \rho_{\text{thr}}(u_c) = \tilde{\rho}'_{\text{thr}}, T, L) < \widehat{\text{FDP}}(v = 1 - \Delta v, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T, L)$, the algorithm proceeds by increasing the number of included dummies T at a reference point $(v, \rho_{\text{thr}}(u_c)) = (1 - \Delta v, \tilde{\rho}_{\text{thr}})$, where $\Delta v = 1/K$, until $\widehat{\text{FDP}}(v = 1 - \Delta v, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T, L)$ exceeds the target level α or reaches the maximum allowed number of included dummies T_{max}. Fixing the optimized parameters L, T and the corresponding FDR-controlled variable sets, the algorithm then determines the optimal values of v and $\rho_{\text{thr}}(u_c)$ that maximize the number of selected variables by solving the optimization problem in (6). Finally, the obtained solution $(v^*, \rho_{\text{thr}}(u_c^*), T^*, L)$ yields the FDR-controlled set of selected variables

$$\widehat{\mathcal{A}}_L(v^*, \rho_{\text{thr}}(u_c^*), T^*) = \left\{ j : \Phi^{\text{DA}}_{T^*,L}(j, \rho_{\text{thr}}(u_c^*)) > v^* \right\}. \tag{4.36}$$

In the following, we state and prove that Algorithm 5 yields an optimal solution of (4.32).

Theorem 11 (Optimal Dependency-Aware Calibration)**.** *Suppose that L, as obtained by Algorithm 5, is fixed and that, ceteris paribus, $\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L)$ is monotonically decreasing in v and monotonically increasing in T. Then, any triple $(v^*, \rho_{\text{thr}}(u_c^*), T^*)$ of a quadruple $(v^*, \rho_{\text{thr}}(u_c^*), T^*, L)$, as obtained by Algorithm 5, is an optimal solution of* (4.32).

Proof. The proof is deferred to Appendix B.1.5. □

4.3 Numerical Experiments

In this section, we verify the FDR control property of the proposed *T-Rex+DA* selector via numerical experiments and compare its performance against three state-of-the-art methods for high-dimensional data, i.e., *model-X* knockoff [Can+18], *model-X* knockoff+ [Can+18], and the proposed *T-Rex* selector from Chapter 3.

4.3.1 Setup

We consider a high-dimensional setting with $p = 500$ variables and $n = 150$ samples and generate the predictor matrix $\boldsymbol{X}$ from a zero mean multivariate Gaussian distribution with

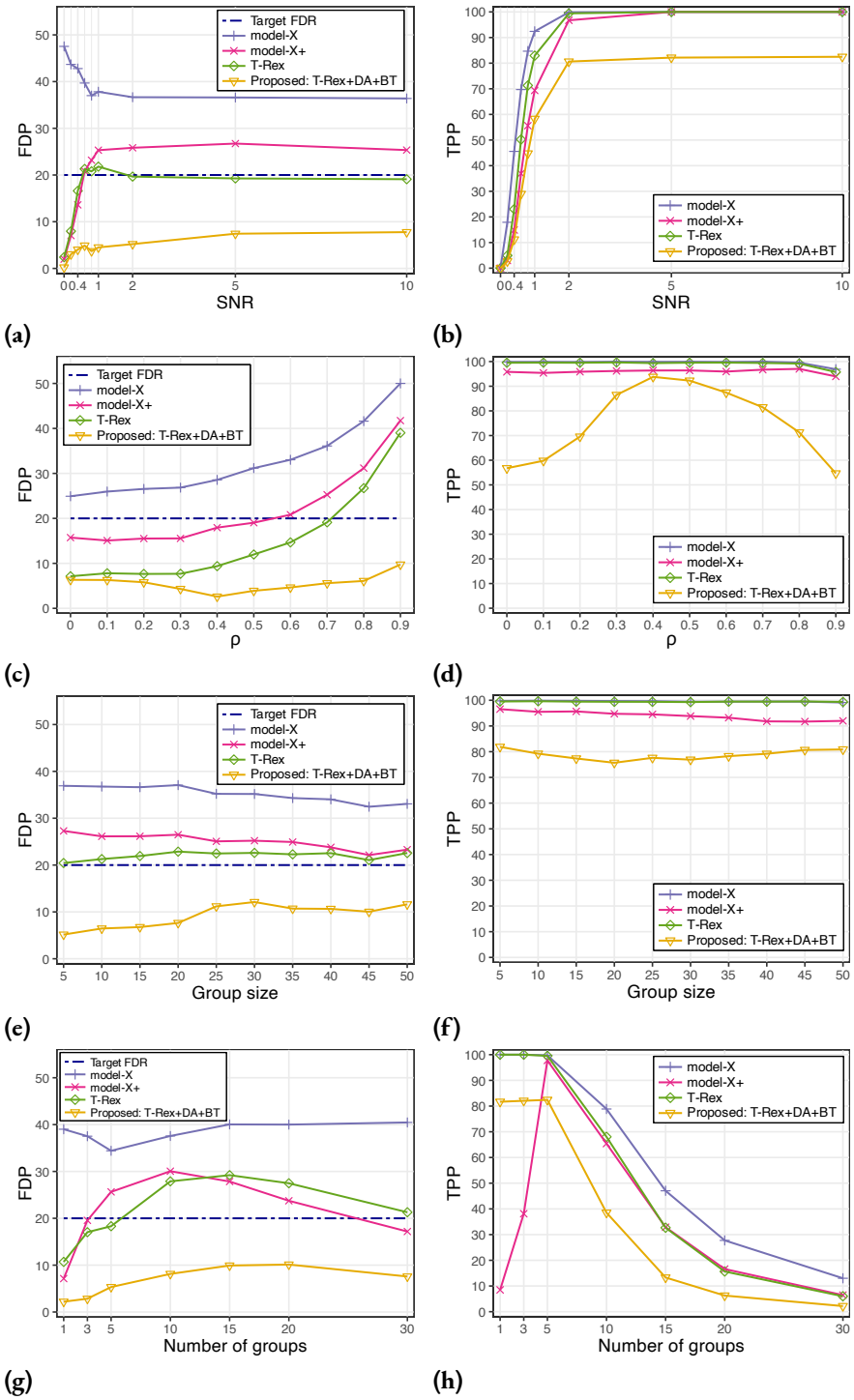
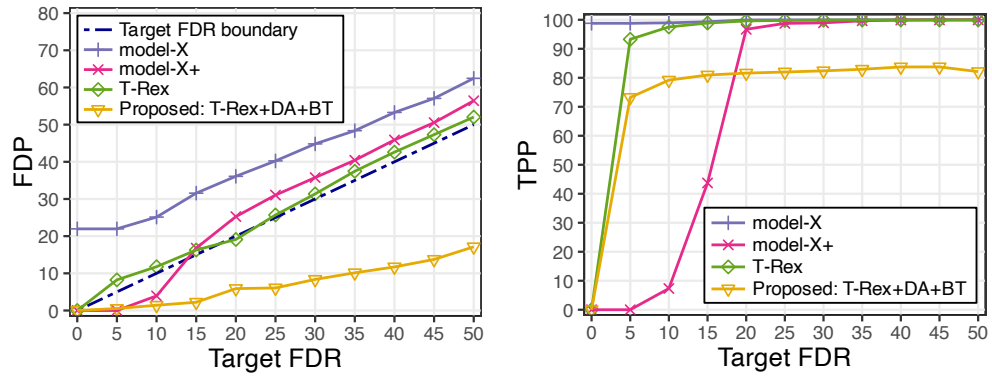
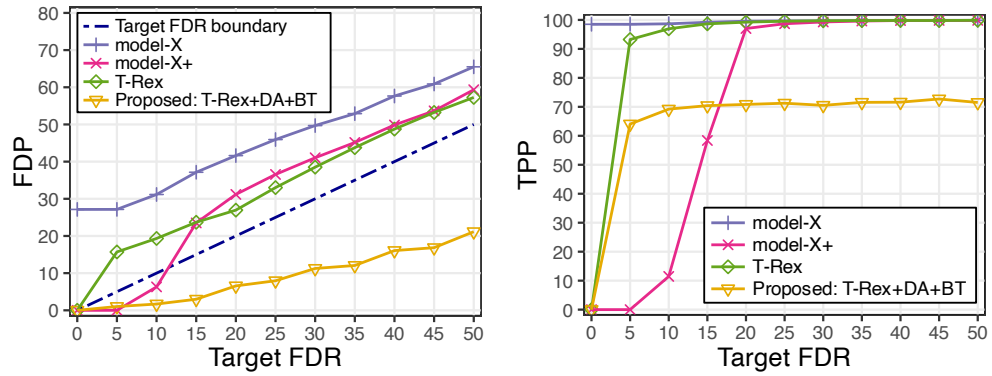


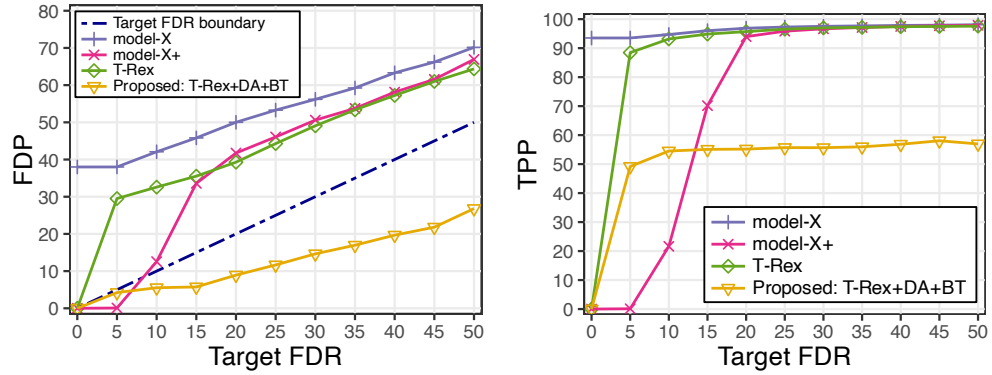
Figure 4.3: Only the proposed *T-Rex+DA* selector with a binary tree group model (*T-Rex+DA+BT*) reliably controls the FDR in all settings while achieving a reasonably high TPR. In Figure (c), we see that with increasing correlations among the variables in a group, the benchmark methods exhibit an alarming increase in FDR.



(a) $\rho = 0.7$ **(b)** $\rho = 0.7$



(c) $\rho = 0.8$ **(d)** $\rho = 0.8$



(e) $\rho = 0.9$ **(f)** $\rho = 0.9$

Figure 4.4: The proposed *T-Rex+DA* selector reliably controls the FDR in all settings while achieving a reasonably high TPR in harsh high correlation settings. We observe that with increasing correlations among the variables in a group, the benchmark methods do not control the FDR for almost any choice of target FDR.

an M block diagonal correlation matrix, where each block is a $Q \times Q$ toeplitz correlation matrix, i.e.,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & & \vdots \\ \vdots & & \boldsymbol{\Sigma}_M & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{Q-1} \\ \rho & 1 & \rho & \cdots & \rho^{Q-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{Q-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{Q-1} & \rho^{Q-2} & \rho^{Q-3} & \dots & 1 \end{bmatrix}. \tag{4.37}$$

That is, each block mimics a dependency structure that is often present in biomedical data (e.g., gene expression [SDC03] and genomics data [Bal06]) and may lead to the breakdown of the FDR control property of existing methods. The response vector $\boldsymbol{y}$ is generated from the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = [\beta_1 \cdots \beta_p]^\top \in \mathbb{R}^p$ is the sparse true coefficient vector and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is an additive noise vector with variance σ^2 and identity matrix $\boldsymbol{I}$. The variance σ^2 is set such that the signal-to-noise ratio $\text{SNR} = \text{Var}[\boldsymbol{X}\boldsymbol{\beta}]/\sigma^2$ has the desired value. In the base setting, we set the parameters as follows: $\text{SNR} = 2$, $\rho = 0.7$, $Q = 5$, $M = 5$, $\alpha = 0.2$. The coefficient vector $\boldsymbol{\beta}$ is generated such that the mth block consists of one true active variable with coefficient value one, while the remaining variables are nulls with coefficient value zero. In the numerical experiments, all parameters except for one parameter of the base setting are varied. That is, ceteris paribus, SNR, ρ, group size Q, number of groups M, and target FDR α are varied.

4.3.2 Results

The results in Figures 4.3 and 4.4 are averaged over 955 Monte Carlo replications.[1] For the performance comparison, we consider the averaged FDP and TPP (in %) which are estimates of the FDR and TPR. We observe that only the proposed *T-Rex+DA* selector with a binary tree group model (*T-Rex+DA+BT*) reliably controls the FDR over all values of SNR, ρ, Q, M, and α, while the benchmarks lose the FDR control property, especially in the practically important case where groups of highly correlated variables are present in the data. It is remarkable that the frequently used *model-X* knockoff method exceeds the target FDR level in all scenarios by far. Note that achieving a higher TPR without controlling the FDR is undesirable, since it leads to reporting false discoveries, which need to be avoided in order to alleviate

[1] The uneven number of Monte Carlo replications was chosen to run the simulations efficiently and in parallel on the Lichtenberg High-Performance Computer of the Technische Universität Darmstadt.

the unfortunately still ongoing reproducibility crisis in many scientific fields [Bak16b].

The results of two additional simulation setups, where

1. the p-dimensional samples of the predictor matrix (i.e., rows of $\boldsymbol{X}$) are sampled from a zero-mean multivariate heavy-tailed Student-t distribution with covariance matrix $\boldsymbol{\Sigma}$ and 3 degrees of freedom,
2. the noise vector $\boldsymbol{\epsilon}$ is sampled from a heavy-tailed Student-t distribution with 3 degrees of freedom,

are deferred to Appendix B.2. These additional simulations verify the theoretical results and show that only the proposed *T-Rex+DA* selector reliably controls the FDR in these heavy-tailed settings.

4.4 The T-Rex+DA Selector for Sparse Financial Index Tracking

In Section 4.2.3, the *T-Rex+DA* framework has been adapted for non-overlapping groups of highly correlated candidate variables. However, there exist applications in areas such as financial engineering that exhibit variable dependency structures that cannot be modeled as non-overlapping groups of variables. In particular, in financial index tracking overlapping groups of highly correlated stocks need to be taken into account to determine an FDR-controlled sparse tracking portfolio. To address this issue, we have expanded the *T-Rex+DA* framework to accommodate overlapping groups of highly correlated variables.

More specifically, we propose a new FDR-controlling index tracking method by

1. extending the *T-Rex+DA* framework to account for strongly overlapping groups of highly correlated variables using a nearest neighbors penalization mechanism,
2. proving that it controls the FDR at the investor-specified target level, and
3. demonstrating its unique capability of accurately tracking the S&P 500 index over the last 20 years based on a sparse, FDR-controlled, and quarterly updated portfolio (see Chapter 7).

Moreover, we propose another extension of the *T-Rex+DA* framework that accounts for autoregressive dependencies among candidate variables. In Chapter 7, its capability of accurately

tracking the S&P 500 index at a very low target FDR level is demonstrated.

While this work highlights an application in finance, the method more generally applies to any setting where the variables exhibit an overlapping groups dependency structure.

4.4.1 Adaptation for Overlapping Groups of Highly Correlated Variables

The binary tree design of the groups $\text{Gr}(j, \rho_{\text{thr}})$ for the *T-Rex+DA+BT* selector that has been introduced in Definition 13 does not allow for arbitrary overlapping groups of highly correlated variables, which are characteristic for stock returns data. Therefore, in this section, we propose a new design for the variable groups $\text{Gr}(j, \rho_{\text{thr}})$, $j = 1, \dots p$, which incorporates a nearest neighbors (NN) penalization mechanism into the dependency-aware *T-Rex* framework. We then prove that the proposed approach controls the FDR at the user-specified target level.

4.4.1.1 Group Design

For the proposed NN group design, each variable j is assigned a variable group that contains variables whose correlations with variable j exceed a threshold $\rho_{\text{thr}} \in [0, 1]$, i.e.,

$$\text{Gr}(j, \rho_{\text{thr}}) \coloneqq \big\{ j' \in \{1, \dots, p\} \backslash \{j\} : \tag{4.38}$$

$$|\operatorname{corr}(\boldsymbol{x}_j, \boldsymbol{x}_{j'})| \geq \rho_{\text{thr}} \big\}. \tag{4.39}$$

Clearly, in contrast to the group design of the *T-Rex+DA+BT* selector in Definition 13, this group design allows for overlapping groups of highly correlated variables. The challenge now is to jointly determine ρ_{thr} in (4.39) and the other parameters of the dependency-aware *T-Rex* selector (i.e., v, T, and L) such that the proposed NN group design yields FDR-controlled solutions

$$\widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}, T) \coloneqq \{j : \Phi^{\text{NN}}_{T,L}(j, \rho_{\text{thr}}) > v\}, \tag{4.40}$$

where

$$\Phi^{\text{NN}}_{T,L}(j, \rho_{\text{thr}}) \coloneqq \Psi^{\text{NN}}_{T,L}(j, \rho_{\text{thr}}) \cdot \Phi_{T,L}(j) \tag{4.41}$$

is the dependency-aware relative occurrence of the jth variable using the proposed NN group design in (4.39).

4.4.1.2 FDR Control

The distinction of the penalty functions $\Psi_{T,L}$ and $\Psi_{T,L}^{\text{NN}}$ in (4.4) and (4.41) stems from their underlying group designs. The underlying group design of $\Psi_{T,L}$ only allows for disjoint variable groups. In contrast, the proposed NN group design in (4.39) extends the scope of the *T-Rex+DA* selector by allowing for overlapping groups of highly correlated variables, which are characteristic for applications such as the considered financial index tracking.

In order to prove that the proposed NN group design controls the FDR, we first define the conservative FDP estimator[2]

$$\widehat{\text{FDP}}(v, \rho_{\text{thr}}, T, L) \coloneqq \frac{\widehat{V}_{T,L}(v, \rho_{\text{thr}})}{\max\left\{1, R_{T,L}(v, \rho_{\text{thr}})\right\}}, \tag{4.42}$$

where $R_{T,L}(v, \rho_{\text{thr}}) \coloneqq |\widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}, T)|$ is the number of selected variables. Second, let $\widehat{\mathcal{A}}(v, \rho_{\text{thr}}) \coloneqq \widehat{\mathcal{A}}_L(v, \rho_{\text{thr}}, T)$ and $\Delta\Phi_{t,L}^{\text{NN}}(j, \rho_{\text{thr}}) \coloneqq \Phi_{t,L}^{\text{NN}}(j, \rho_{\text{thr}}) - \Phi_{t-1,L}^{\text{NN}}(j, \rho_{\text{thr}})$. Then, the proposed estimator of the unknown number of selected null variables $V_{T,L}(v, \rho_{\text{thr}})$ in (4.42) is defined by

$$\widehat{V}_{T,L}(v, \rho_{\text{thr}}) \coloneqq \sum_{j \in \widehat{\mathcal{A}}(v, \rho_{\text{thr}})} \left(1 - \Phi_{T,L}^{\text{NN}}(j, \rho_{\text{thr}})\right) \tag{4.43}$$

$$+ \underbrace{\sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi_{t,L}^{\text{NN}}(q, \rho_{\text{thr}})}{L - (t-1)} \cdot \frac{\sum_{j \in \widehat{\mathcal{A}}(v, \rho_{\text{thr}})} \Delta\Phi_{t,L}^{\text{NN}}(j, \rho_{\text{thr}})}{\sum_{j \in \widehat{\mathcal{A}}(0.5, \rho_{\text{thr}})} \Delta\Phi_{t,L}^{\text{NN}}(j, \rho_{\text{thr}})}}_{=: \widehat{V}'_{T,L}(v, \rho_{\text{thr}})}, \tag{4.44}$$

which is similar to the estimator that was defined in (4.29) with the innovation that the ordinary relative occurrences are replaced by the proposed relative occurrences in (4.41).

With all necessary definitions in place, we now state and prove that any quadruple $(v, \rho_{\text{thr}}, T, L) \in [0.5, 1) \times [0, 1] \times \{1, \ldots, L\} \times \mathbb{N}_+$, for which the conservative FDP estimator in (4.42) does not exceed the user-specified target level $\alpha \in [0, 1]$, yields FDR control. To simultaneously maximize the number of selected variables (to obtain highest possible TPR while controlling FDR), we choose the lowest possible voting threshold v that

[2] Conservative is meant in the sense that $\mathbb{E}[\text{FDP}] \leq \mathbb{E}[\widehat{\text{FDP}}]$.

maintains FDR control, i.e.,

$$v \coloneqq \inf\{\nu \in [0.5, 1) : \widehat{\mathrm{FDP}}(\nu, \rho_{\text{thr}}, T, L) \leq \alpha\}. \tag{4.45}$$

Theorem 12 (FDR control)**.** *Let* $\mathrm{Gr}(j, \rho_{\text{thr}})$ *be as defined in* (4.39) *and* $K \to \infty$*. Suppose that* $\widehat{V}'_{T,L}(v, \rho_{\text{thr}}) > 0$*. Then, for any quadruple* $(v, \rho_{\text{thr}}, T, L) \in [0.5, 1) \times [0, 1] \times \{1, \ldots, L\} \times \mathbb{N}_+$ *that satisfies Equation* (4.45)*, it holds that* $\mathrm{FDR}(v, \rho_{\text{thr}}, T, L) \leq \alpha$*, i.e., the FDR is controlled at the user-specified target level* $\alpha \in [0, 1]$*.*

Proof. The proof is deferred to Appendix B.1.6. □

4.4.2 Adaptation for Autoregressive Variable Dependency Models

The proposed *T-Rex+DA+NN* selector from Section 4.4.1 is able to control the FDR even in the presence of overlapping groups of highly correlated variables. However, opting for group dependency models often demands investors to accept a relatively high target FDR level (e.g., 30%) for a sufficiently diversified tracking portfolio, while lower target levels (e.g., 1%) typically result in an empty portfolio. Tolerating high target FDR levels leads to more irrelevant stocks in the tracking portfolio and consequently to higher transaction costs.

To reinstate the FDR control property amid highly dependent stocks while achieving high selection power at low target FDR levels, this section presents the following methodological and experimental contributions:

1. Proposed: Integrating autoregressive dependency models into the *T-Rex+DA* framework to account for dependencies among stocks.
2. Theoretical analysis and numerical experiments showing that the proposed methodology controls the FDR at the user-specified target level, whereas the ordinary *T-Rex* selector loses the FDR control property in the presence of strong autoregressive dependencies among variables.
3. A real-world S&P 500 index tracking use-case demonstrating that the simple and yet accurate autoregressive dependency models allow to effectively leverage the dependencies among stocks, which yields well diversified tracking portfolios at low target FDR levels (see Chapter 7).

4.4.2.1 Group Design

In order to control the FDR in the presence of autoregressive dependency structures among candidate variables, we mathematically model and account for such dependencies by adapting the formulation of the variable groups $\mathrm{Gr}(j, \rho\mathrm{thr})$, $j = 1, \ldots, p$. We focus on first-order autoregressive (AR(1)) dependencies, since this simple first order model suffices to properly model the dependencies among stocks for the purpose of FDR-controlled index tracking (see Section 7.1.3.2).

First, let $X_1, \ldots, X_p$ be standardized (i.e., $\mathbb{E}[X_j] = 0$ and $\mathrm{Var}[X_j] = 1$, $j = 1, \ldots, p$) random variables associated with the predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. Then, a weak-sense stationary AR(1) process is characterized by

$$X_j = \rho X_{j-1} + E_j, \quad |\rho| < 1, \tag{4.46}$$

where the white noise process E_j is characterized by $\mathbb{E}[E_j] = 0$ and $\mathrm{Var}[E_j] = \sigma_E^2$. The autocorrelation function of the AR(1) process in (4.46) is given by

$$r(\kappa) = \rho^{\kappa}, \tag{4.47}$$

where $\kappa \in \{-p + 1, \ldots, p - 1\}$ is the lag of the autocorrelation function. That is, the correlation between variables is exponentially decaying in κ.

Leveraging this property, we introduce a sliding window function that maps each candidate variable j to a group of neighboring variables, such that the pairwise correlations between variable j and its associated group of variables do not fall below a threshold $\rho_{\text{thr}} \in [0, 1)$. That is, $|r(\kappa)| \leq \rho_{\text{thr}}$, which yields $\kappa \geq \log(\rho_{\text{thr}})/\log(|\rho|)$.

Definition 17 (Sliding window)**.** *Let $\hat{\rho}$ be the maximum likelihood estimator of ρ in* (4.46) *and* (4.47) *averaged over all n samples and $\rho_{\text{thr}} \in [0, 1)$ the autocorrelation threshold. Define $\widehat{|\kappa|} \coloneqq \lceil \log(\rho_{\text{thr}})/\log(|\hat{\rho}|) \rceil$. Then, the jth variable group $\mathrm{Gr}(j, \rho_{\text{thr}})$ is specified by*

$$\mathrm{Gr}(j, \rho_{\text{thr}}) \coloneqq \mathrm{SW}(j, \rho_{\text{thr}}), \tag{4.48}$$

where the sliding window function is defined by

$$\mathrm{SW}(j, \rho_{\text{thr}}) \coloneqq \left\{ \max\left\{1, j - \widehat{|\kappa|}\right\}, \ldots, \min\left\{p, j + \widehat{|\kappa|}\right\}\right\} \backslash \{j\}. \tag{4.49}$$

With this definition in place, the original relative occurrences of the *T-Rex* selector $\Phi_{T,L}(j)$ are replaced by the dependency-aware relative occurrences for the autoregressive model $\Phi^{\text{AR}}_{T,L}(j, \rho_{\text{thr}})$, which yields the proposed dependency-aware *T-Rex* selector for autoregressive dependency models (*T-Rex+DA+AR*1).

Definition 18 (Autoregressive dependency-aware relative occurrences)**.** *Let* $\text{Gr}(j, \rho_{\text{thr}}) \coloneqq \text{SW}(j, \rho_{\text{thr}})$ *and* $\Phi_{T,L}(j)$ *be the ordinary relative occurrence in* (3.4)*. Then, the* j*th dependency-aware relative occurrence for the autoregressive dependency model is given by*

$$\Phi^{\text{AR}}_{T,L}(j, \rho_{\text{thr}}) \coloneqq \Psi^{\text{AR}}_{T,L}(j, \rho_{\text{thr}}) \cdot \Phi_{T,L}(j), \tag{4.50}$$

where the penalty factor is defined by

$$\Psi^{\text{AR}}_{T,L}(j, \rho_{\text{thr}}) \coloneqq \frac{1}{2 \ - \min\limits_{j' \in \text{SW}(j, \rho_{\text{thr}})} \left\{ \left| \Phi_{T,L}(j) - \Phi_{T,L}(j') \right| \right\}}. \tag{4.51}$$

4.4.2.2 Theoretical Analysis

From Definition 18, it is clear that the jth dependency-aware relative occurrence arises from the penalization of the corresponding ordinary relative occurrence according to its similarity with all other ordinary relative occurrences within its associated variable group $\text{SW}(j, \rho_{\text{thr}})$.

In the following, we state and prove for the *Lasso* forward selector [Tib96; Efr+04] within the *T-Rex+DA* framework that the dependency-aware relative occurrence of any variable j is related to its largest sample correlation coefficient with respect to the proposed sliding window group design in Definition 17. For this purpose, let

$$\hat{\boldsymbol{\beta}}_k \coloneqq [\hat{\beta}_{1,k} \ \cdots \ \hat{\beta}_{p,k}]^\top \tag{4.52}$$

$$\coloneqq \arg\min_{\boldsymbol{\beta}_k} \frac{1}{2} \left\| \boldsymbol{y} - \widetilde{\boldsymbol{X}}_k \boldsymbol{\beta}_k \right\|_2^2 + \lambda_k(T, L) \cdot \|\boldsymbol{\beta}_k\|_1 \tag{4.53}$$

be the *Lasso* estimator in the kth random experiment and let $\lambda_k(T, L) > 0$ be the sparsity parameter, as in (4.1).

Theorem 13. *Define* $\overline{\Lambda} \coloneqq \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\lambda_k(T,L)}$ *and* $\rho_{j,j'} \coloneqq \boldsymbol{x}_j^\top \boldsymbol{x}_{j'}$*. Then, for any* $j, j' \in \{1, \ldots, p\}$*, that satisfy* $\hat{\beta}_{j,k}, \hat{\beta}_{j',k} \neq 0$*, it holds that*

$$\Phi^{\text{AR}}_{T,L}(j, \rho_{\text{thr}}) \leq \left[2 - \overline{\Lambda} \|\boldsymbol{y}\|_2 \sqrt{2\Big(1 - \max_{j' \in \text{SW}(j, \rho_{\text{thr}})} \{\rho_{j,j'}\}\Big)} \right]^{-1} \Phi_{T,L}(j). \tag{4.54}$$

Proof. Considering the "absolute difference of relative occurrences" theorem (i.e., Theorem 8), which states that

$$\left|\Phi_{T,L}(j) - \Phi_{T,L}(j')\right| \leq \overline{\Lambda}\|\boldsymbol{y}\|_2\sqrt{2(1-\rho_{j,j'})}, \tag{4.55}$$

we obtain

$$\begin{aligned}
\Phi^{\text{AR}}_{T,L}(j,\rho_{\text{thr}}) &= \left[2 - \min_{j'\in \text{SW}(j,\rho_{\text{thr}})}\left\{\left|\Phi_{T,L}(j) - \Phi_{T,L}(j')\right|\right\}\right]^{-1}\Phi_{T,L}(j) && (4.56)\\
&\leq \left[2 - \overline{\Lambda}\|\boldsymbol{y}\|_2 \min_{j'\in \text{SW}(j,\rho_{\text{thr}})}\left\{\sqrt{2(1-\rho_{j,j'})}\right\}\right]^{-1}\Phi_{T,L}(j) && (4.57)\\
&= \left[2 - \overline{\Lambda}\|\boldsymbol{y}\|_2\sqrt{2\Big(1 - \max_{j'\in \text{SW}(j,\rho_{\text{thr}})}\{\rho_{j,j'}\}\Big)}\right]^{-1}\Phi_{T,L}(j), && (4.58)
\end{aligned}$$

where the inequality in the second line follows from (4.55). □

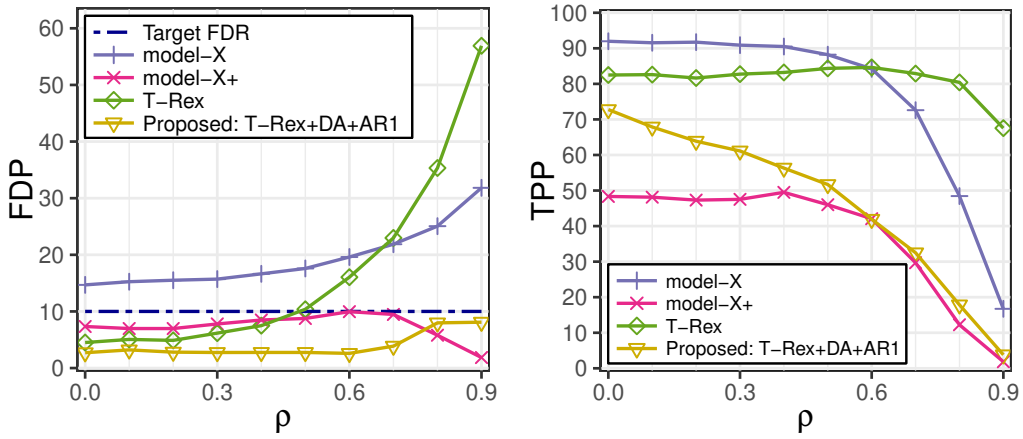
Remark 10. *Note that $\rho_{j,j'}$ considers the sample correlations. The expected value of $\rho_{j,j'}$ is equal to ρ from (4.46) for all j, j'. That is, in the idealized case we have*

$$\Phi^{\text{AR}}_{T,L}(j,\rho_{\text{thr}}) \leq \left[2 - \overline{\Lambda}\|\boldsymbol{y}\|_2\cdot\sqrt{2(1-\rho)}\right]^{-1}\cdot\Phi_{T,L}(j), \tag{4.59}$$

which reveals that in the perfectly correlated case $\rho \to 1$, where candidate variables cannot be distinguished into true active and null variables anymore, $\Phi^{\text{AR}}_{T,L}(j,\rho_{\text{thr}}) = 0.5\,\Phi_{T,L}(j,\rho_{\text{thr}}) \leq 0.5$, $j = 1,\ldots,p$. Thus, in accordance with (3.3), no variable is selected. In Section 4.4.2.3 (see Figure 4.5b), this is verified numerically for reasonably high values of ρ.

4.4.2.3 Numerical Experiments

In the following, the proposed *T-Rex+DA+AR*1 selector for the autoregressive dependency model is compared against the original *T-Rex* selector from Chapter 3, the *model-X* knockoff method [Can+18], and the *model-X* knockoff+ method [Can+18]. The sparse and high-dimensional simulation setting consists of $p = 500$ candidate variables and $n = 300$ samples. The predictors are sampled according to the AR(1) model in (4.46) with Gaussian white noise with $\sigma_E = 1$ and the response is generated according to the the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Here, $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$ is the predictor matrix, $\boldsymbol{y} \in \mathbb{R}^n$ the response, $\boldsymbol{\beta} \in \mathbb{R}^p$ the coefficient vector and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ the Gaussian noise vector with standard deviation σ and identity matrix $\boldsymbol{I}$. The standard deviation σ is chosen such that the signal-to-noise ratio



(a) Average FDP (in %). **(b)** Average TPP (in %).

Figure 4.5: The numerical experiments verify that the proposed *T-Rex+DA+AR*1 selector for the autoregressive dependency model controls the FDR for all autocorrelation values ρ, while the original *T-Rex* selector loses its FDR control property for $\rho > 0.5$. The *model-X* knockoff method does not control the FDR. Among the two methods that control the FDR over the whole range of ρ, the proposed *T-Rex+DA+AR*1 selector has the highest power despite the conservative value of $\rho_{\text{thr}} = 0.02$.

(SNR) is one, i.e., $\text{SNR} = \text{Var}(\boldsymbol{X}\boldsymbol{\beta})/\,\text{Var}(\boldsymbol{\epsilon}) = 1$. Only $p_1 = 10$ randomly selected entries of the coefficient vector $\boldsymbol{\beta}$ are set to one and the remaining ones are set to zero, i.e., 10 out of 500 variables are true active variables. The target FDR is set to 10%. The FDP and TPP in Figure 4.5 are averaged over 955 Monte Carlo replications. The numerical experiments confirm empirically that the proposed *T-Rex+DA+AR*1 selector controls the FDR. Furthermore, it has a higher TPR compared to the *model-X+* knockoff method, which also controls the FDR in this example.

4.5 Summary

The dependency-aware *T-Rex* (*T-Rex+DA*) selector has been proposed. In contrast to existing methods, it reliably controls the FDR in the presence of groups of highly correlated variables in the data. The *T-Rex+DA* selector was specified for dependency structures that can be modeled as groups of non-overlapping variables (i.e., *T-Rex+DA+BT* selector). A real world TCGA breast cancer survival analysis is provided in Chapter 6. It shows that the proposed method selects genes that have been previously identified to be related to breast cancer. Thus, the *T-Rex+DA* selector is a promising tool for making reproducible discoveries in biomedical applications. Moreover, the derived group design principle allows to easily adapt the method

to various application-specific dependency-structures, which opens the door to many other fields that require large-scale high-dimensional variable selection with FDR-control guarantees.

The *T-Rex+DA+NN* selector has been proposed. It is an extension of the *T-Rex+DA* framework that accounts for overlapping groups of highly correlated variables. The proposed method provably controls the FDR and has been successfully applied in tracking the S&P 500 index over the last 20 years with higher accuracy and fewer stocks compared to state-of-the-art methods (see Chapter 7). Therefore, it is a promising approach for index tracking as well as other applications where overlapping groups of highly correlated variables exist.

The *T-Rex+DA+AR*1 selector has been proposed. It is another extension of the *T-Rex+DA* selector that accounts for autoregressive dependency models. It ensures reproducibility of discoveries through FDR-controlled high-dimensional variable selection in the presence of autoregressive dependency structures among the candidate variables. In FDR-controlled sparse financial index tracking, the *T-Rex+DA+AR*1 selector has outperformed the *T-Rex+DA+NN* selector at low target FDR levels. That is, the *T-Rex+DA+AR*1 selector has been capable of selecting the relevant stocks, while the *T-Rex+DA+NN* selector has suffered from missed detections at low target FDR levels.

A great idea solves multiple problems at the same time.

Shigeru Miyamoto

5

Joint Grouped Variable Selection and False Discovery Rate Control

This chapter tackles the challenge posed by the presence of groups of highly correlated candidate variables in the data from a different perspective and with a different aim than the proposed *T-Rex+DA* selector in Chapter 4. Wile the approach of the more conservative *T-Rex+DA* selector is to detect the few individual true active variables among highly dependent candidate variables, the aim of this chapter is to select entire groups of highly correlated variables that contain at least one true active variable. This is especially useful in biomedical applications like GWAS, where it is not necessary to select individual SNPs but rather entire groups of highly correlated SNPs that point to a few locations on the genome.

In Section 5.1, the proposed *T-Rex* selector for grouped variable selection (*T-Rex+GVS*) selector is presented. Section 5.2 presents the proposed *informed elastic net (IEN)* base selector, which has been developed to replace the *elastic net (EN)* base selector within the *T-Rex+GVS* selector. The advantage of the *IEN* over the *EN* base selector is that it reduces the overall computation time of the *T-Rex+GVS* selector while maintaining its grouped variable selection (GVS) property. In Section 5.3, the *T-Rex+GVS* selector is leveraged to perform FDR-controlled sparse principal component analysis (PCA). The proposed *T-Rex* PCA method yields FDR-controlled and, therefore, interpretable loading vectors and principal components

(PCs).

In Chapters 6 and 7, the proposed methods have been used to solve challenging real-world data problems in biomedical and financial engineering. The technical proofs of the theoretical results are deferred to Appendix C.

The main content of this chapter is based on the publications [MMP22], [MMP23b] and [Mac+24a]. The implementations of the developed methods are included in the open source R software packages (with C++ backend) TRexSelector [Mac+24c] and tlars [Mac+24b] on CRAN.

5.1 The T-Rex+GVS Selector

This section motivates the proposed *T-Rex+GVS* selector, provides an overview of the major contributions, and presents the grouped variable selection property and the dummy generation algorithm of the *T-Rex+GVS* selector.

5.1.1 Motivation

High-dimensional variable selection is a challenging task, especially when groups of highly correlated variables are present in the data. However, in applications such as genome-wide association studies (GWAS), it is not necessary to detect the few individual single nucleotide polymorphisms (SNPs) that are associated with a phenotype of interest among highly correlated candidate SNPs. Instead, it is sufficient to detect entire groups of highly correlated SNPs that point to a few associated locations on the genome. The reason for this is that groups of nearby and, therefore, highly correlated SNPs occur throughout the genome due to a phenomenon called linkage disequilibrium [Rei+01]. This less stringent requirement allows to develop more liberal FDR-controlling methods that have a higher TPR at the cost of not determining the individual true active variables but rather the true active groups of variables, i.e., groups that contain at least one true active variable.

5.1.2 Major Contributions

Real-world problems, such as the one described in Section 5.1.1, are addressed by adapting the forward variable selection procedure and the dummy generation process within the *T-Rex* framework for grouped variable selection. The proposed *T-Rex+GVS* selector has four major innovations that distinguish it from the original *T-Rex* selector:

1. It replaces the originally used variable selection method (*LARS* or *Lasso*) by the *elastic net* (see Section 2.1), which renders it suitable for performing grouped variable selection.

2. This replacement requires a major adjustment of the dummy generation process. A new dummy generation process that mimics the group correlation structure of $\boldsymbol{X}$ is proposed. This necessary adjustment fosters the generation of groups of highly correlated dummies that allow for a fair competition of original variables and dummies to be included along the forward variable selection process within each random experiment.

3. We prove that the *T-Rex+GVS* selector possesses the desirable grouped variable selection property (Theorem 14).

4. Through a simulated high-dimensional genome-wide association study (GWAS), it is demonstrated that the proposed *T-Rex+GVS* selector significantly increases the TPR, while controlling the FDR at the target level (see Chapter 6).

5.1.3 Grouped Variable Selection

The *elastic net* in (2.20) requires choosing the tuning parameters λ_1 and λ_2, where $\lambda_1, \lambda_2 > 0$ are the weights of the sparsity inducing ℓ_1-norm penalty (*Lasso*) and the grouped selection fostering ℓ_2-norm penalty (ridge regression), respectively. Since we are interested in performing grouped variable selection, we require a sufficiently large value of λ_2, such that the grouping effect is sufficiently strong. However, since the strength of the grouping effect is not very sensitive to the choice of λ_2, we choose λ_2 by performing 10-fold cross validated ridge regression (see Section 2.1.2) and fix the obtained λ_2-value. With a fixed λ_2, the *elastic net* optimization problem can be reformulated as a *Lasso optimization problem* (see Section 2.1.4) and, therefore, it can be solved by the *LARS* algorithm. Thus, the *elastic net* with fixed λ_2 can be integrated into the original *T-Rex* framework, which removes the necessity of choosing λ_1 because the random experiments are automatically terminated by the *T-Rex* selector, as discussed in Section 3.2.

In the following, we will prove for the special case, where the variables within a group are perfectly correlated, that the desirable grouped variable selection property of the *elastic net* carries over to the proposed *T-Rex+GVS* selector. Considering this idealized case is common in theory, since it reveals whether a method is generally capable of performing grouped variable selection [ZH05]. First, for each standardized variable $m \in \{1, \ldots, p\}$, we define a group of perfectly correlated variables $\mathcal{G}_m$ that contains variable m. Then, we show that if any variable

contained in $\mathcal{G}_m$ is selected (not selected) by the *T-Rex+GVS* selector, then the entire group is selected (not selected).

Theorem 14 (Grouped variable selection)**.** *Define* $\rho_{g,m} \coloneqq \boldsymbol{x}_g^\top \boldsymbol{x}_m$ *and*

$$\mathcal{G}_m \coloneqq \big\{g \in \{1,\ldots,p\} : |\rho_{g,m}| = 1\big\}, m = 1,\ldots,p. \tag{5.1}$$

The following two statements hold for all triples $(v, T, L) \in [0.5, 1) \times \{1,\ldots,L\} \times \mathbb{N}_+$*:*

(i) Suppose that $j \in \mathcal{G}_m$ *and* $j \in \widehat{\mathcal{A}}_L(v,T)$*. Then, it holds that* $\mathcal{G}_m \subseteq \widehat{\mathcal{A}}_L(v,T)$*.*

(ii) Suppose that $j \in \mathcal{G}_m$ *and* $j \notin \widehat{\mathcal{A}}_L(v,T)$*. Then, it holds that* $\mathcal{G}_m \cap \widehat{\mathcal{A}}_L(v,T) = \varnothing$*, where* $\varnothing$ *denotes the empty set.*

Proof. The proof is deferred to the Appendix C.1. □

5.1.4 Dummy Generation

The goal of the proposed dummy generation algorithm is to generate dummy matrices $\mathring{\boldsymbol{X}}_k$, $k = 1,\ldots,K$, that mimic the group correlation structure that is present within $\boldsymbol{X}$.

First, in order to cluster the variables into groups of highly correlated variables with low correlations between variables from different clusters, we apply single-linkage hierarchical clustering [MC12] to the predictors in $\boldsymbol{X}$, where the sample correlation is used as the similarity measure. Then, the obtained dendrogram is cut at the lowest level where the sample correlations of any two predictors from different clusters are not higher than the threshold value $\rho_{\text{thr}} = 1/3$. The value of ρ_{thr} is determined empirically, such that the resulting clusters capture the characteristic group correlation structure of SNPs. Such a clustering approach was proposed to be used as an SNP clustering method in, e.g., the supplementary material of [SSC19]. As specified in the extended calibration algorithm (i.e., Algorithm 3) that determines the value of L (i.e., the number of dummies), L is a multiple of the number of predictors p. Thus, L/p sub-dummy matrices that mimic the group correlation structure of $\boldsymbol{X}$ are generated and appended together to obtain the final dummy matrices $\mathring{\boldsymbol{X}}_k$, $k = 1,\ldots,K$.

The annotated pseudocode of the proposed *T-Rex+GVS* dummy generation process for the generation of the kth dummy matrix $\mathring{\boldsymbol{X}}_k$ is given in Algorithm 6.

Algorithm 6 *T-Rex+GVS* dummies

1. **Input**: $\boldsymbol{X}$, ρ_{thr}, L.
2. **Apply** single-linkage hierarchical clustering [MC12] to the predictors in $\boldsymbol{X}$ and **cut** the resulting dendrogram at the lowest level where the sample correlations of any two predictors from different clusters are not higher than ρ_{thr}.
   Result: Z clusters with associated disjoint variable index sets $\mathcal{J}_1, \ldots, \mathcal{J}_Z \subseteq \{1, \ldots, p\}$, where $\bigcup_{z=1}^{Z} \mathcal{J}_z = \{1, \ldots, p\}$.
3. **For** $w = 1, \ldots, w_{\max}$, where $w_{\max} := \dfrac{L}{p}$, **do**:
   3.1. **For** $z = 1, \ldots, Z$ **do**:
      i. **Compute** the sub-cluster covariance matrix
      $$\boldsymbol{\Sigma}_z = \frac{1}{n-1} \boldsymbol{X}_{\mathcal{J}_z}^\top \boldsymbol{X}_{\mathcal{J}_z}, \tag{5.2}$$
      where $\boldsymbol{X}_{\mathcal{J}_z}$ is the sub-matrix of $\boldsymbol{X}$ that contains the predictors corresponding to $\mathcal{J}_z$.
      ii. **Compute** the sub-dummy matrix
      $$\mathring{\boldsymbol{X}}_{z,w} = \begin{bmatrix} \mathring{\boldsymbol{x}}'^\top_{z,w,1} \\ \vdots \\ \mathring{\boldsymbol{x}}'^\top_{z,w,n} \end{bmatrix}, \quad \mathring{\boldsymbol{x}}'_{z,w,i} \sim \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{\Sigma}_z\big), \tag{5.3}$$
      where $\mathring{\boldsymbol{x}}'^\top_{z,w,i}$ is the ith row of $\mathring{\boldsymbol{X}}_{z,w}$.
4. **Output**: kth dummy matrix
$$\mathring{\boldsymbol{X}}_k = \Big[\mathring{\boldsymbol{X}}_{1,1} \cdots \mathring{\boldsymbol{X}}_{Z,1} \quad \cdots \quad \mathring{\boldsymbol{X}}_{1,w_{\max}} \cdots \mathring{\boldsymbol{X}}_{Z,w_{\max}}\Big]. \tag{5.4}$$

5.2 The Informed Elastic Net for Grouped Variable Selection and FDR Control

This section motivates the proposed *informed elastic net (IEN)* and its incorporation into the *T-Rex+GVS* selector as a fast base selector, provides an overview of the major contributions, presents the methodology and theoretical analysis of the proposed method, and assesses the performance via numerical simulations. A simulated GWAS using the proposed method is

presented in Chapter 6.

5.2.1 Motivation

The proposed *T-Rex+GVS* selector in Section 5.1 uses the *elastic net (EN)* as a base selector to perform grouped variable selection. Although it significantly increases the TPR in simulated GWAS compared to the original *T-Rex* selector (see Chapter 6), its comparably high computational cost limits scalability.

The reason for the high computational cost of the *EN* base selector compared to the ordinary *LARS* algorithm in Section 2.1.3 is that solving it in a forward selection manner using the *LARS* algorithm requires to augment the input data to the *LARS* algorithm as detailed in (2.22). More specifically, p rows of dimension p have to be added to the predictor matrix $\boldsymbol{X}$ and the response vector has to be augmented accordingly by adding zeros. That is, the $(n \times p)$-dimensional predictor matrix $\boldsymbol{X}$ and the n-dimensional response vector $\boldsymbol{y}$ are replaced by the $((n + p) \times p)$-dimensional augmented predictor matrix $\boldsymbol{X}'$ and the $(n + p)$-dimensional augmented response vector $\boldsymbol{y}'$. Furthermore, when using the *EN* as a base selector within the *T-Rex* framework, we also need to account for L dummies within the enlarged predictor matrices in (3.2). That is, p becomes $p + L$.

In conclusion, the required augmentation of the data leads to a high computational cost that does not allow the *EN* base selector to scale to large-scale high-dimensional problems such as GWAS. Therefore, we propose the *IEN*, which has a reduced computational cost compared to the *EN* while maintaining a sufficiently strong and provable grouped variable selection property. This is achieved by formulating a new penalty term that incorporates information about the group correlation structure among the candidate variables to reduce the size of the augmented predictor matrix and the response vector.

5.2.2 The T-Rex+GVS Selector with the Informed Elastic Net: Overview of Major Contributions

To alleviate the drawbacks of the *EN* base selector, we propose the *IEN*, a new base selector for the *T-Rex+GVS* method that significantly reduces computation time while retaining the grouped variable selection property. We prove that the proposed *IEN*

1. can be formulated as a *Lasso*-type optimization problem (Theorem 15) and, therefore, can be solved efficiently in a forward-selection manner, as required by the *T-Rex* frame-

work, using the *LARS* algorithm (Algorithm 1) and

2. exhibits a grouping effect (Theorem 16) that is similar to that of the *elastic net*.

Additionally, we validate empirically that the proposed *IEN*, as suggested by Theorems 15 and 16,

1. produces solution paths that are similar to the ones of the *EN* (see Figure 5.1),

2. significantly reduces the computation time compared to the *EN* when incorporated into the *T-Rex* framework as the base selector (see Figure 5.2), and

3. has the same TPR as the *T-Rex+GVS* selector using the *EN* while achieving a much lower FDR in a simulated GWAS (see Chapter 6).

5.2.3 Methodology and Theoretical Analysis

While the *EN* achieves its grouping effect by penalizing $\|\boldsymbol{\beta}\|_2^2$, this section presents a new *GVS* method that incorporates the information of how the variables are grouped into its penalty term. We show that the proposed *IEN* can be formulated as a *Lasso*-type optimization problem (Theorem 15), so that it can be incorporated into the *T-Rex* framework. We also analyze the grouping effect (Theorem 16) and show that the *IEN* boils down to the *EN* when every variable is considered to be its own group (Corollary 2).

In particular, the proposed *IEN* uses single-linkage hierarchical clustering [MC12] with the pairwise correlations of the original variables as a distance measure to cluster variables into groups of highly correlated variables, which are present in genomics data due to a phenomenon called linkage disequilibrium [Rei+01]. The obtained dendrogram from the hierarchical clustering can be cut at different levels to obtain M disjoint groups of variables $\mathcal{G}_1, \ldots, \mathcal{G}_M$, where $M \leq p$. To represent the mth, $m = 1, \ldots, M$, group mathematically, we define the binary support vector $\mathbf{1}_m = [1_{m,1} \cdots 1_{m,p}]^\top \in \{0,1\}^p$ that has one entries for variables in the mth group and zero entries otherwise. The corresponding group size is $p_m := \sum_{j=1}^{p} 1_{m,j}$. With these definitions in place, we define the proposed *IEN*.

Definition 19 (*Informed elastic net (IEN)*)**.** *Let* $\lambda_1, \lambda_2 > 0$ *and let* p_m, $m = 1, \ldots, M$, *and* $\mathbf{1}_m \in \{0,1\}^p$ *be the known group size and the binary support vector of the* m*th group,*

respectively. Then, the Lagrangian of the informed elastic net (IEN) is defined by

$$\mathcal{L}_{\text{IEN}}(\boldsymbol{\beta}) \coloneqq \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{m=1}^{M} \frac{(\mathbf{1}_m^\top \boldsymbol{\beta})^2}{p_m} \tag{5.5}$$

and the solution of the IEN is defined by

$$\hat{\boldsymbol{\beta}} \coloneqq \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\text{IEN}}(\boldsymbol{\beta}). \tag{5.6}$$

The following theorem shows that the proposed *IEN* can be cast as a *Lasso*-type optimization problem and, therefore, can be solved efficiently and in a forward selection fashion, as required by the *T-Rex* framework, using the *LARS* algorithm:[1]

Theorem 15 (*Lasso*-type optimization problem)**.** *Let $\boldsymbol{X}$, $\boldsymbol{y}$, and $\lambda_1, \lambda_2 > 0$ be given and let $\mathbf{0}_M$ be the M-dimensional vector of zeros. Define*

$$\boldsymbol{X}' \coloneqq \sqrt{\lambda_2} \cdot \begin{bmatrix} \boldsymbol{X}/\sqrt{\lambda_2} \\ \mathbf{1}_1^\top/\sqrt{p_1} \\ \vdots \\ \mathbf{1}_M^\top/\sqrt{p_M} \end{bmatrix}, \quad \boldsymbol{y}' \coloneqq \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0}_M \end{bmatrix}. \tag{5.7}$$

Then, the IEN can be formulated as a Lasso-type optimization problem, i.e.,

$$\mathcal{L}_{\text{IEN}}(\boldsymbol{\beta}) = \|\boldsymbol{y}' - \boldsymbol{X}'\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1. \tag{5.8}$$

Proof. Deferred to Appendix C.2. □

Remark 11. *Note that, in contrast to the EN, the IEN data augmentation presented in Theorem 15 requires appending only M additional rows to the original predictor matrix $\boldsymbol{X}$, while solving the elastic net in a forward selection manner requires appending p rows to $\boldsymbol{X}$ (for details, see Section 2.1.4). Since the number of variable groups M, especially in genomics data, is much smaller than the number of variables p (i.e., $M \ll p$), the IEN exhibits a significantly reduced computation time when p is very large.*

The next theorem shows that the proposed *informed elastic net* exhibits a grouping effect, i.e.,

[1] Note that the proposed *IEN* is fundamentally different from the *group Lasso* approach in [YL06], since the solution path of the *group Lasso* is not piecewise linear and, therefore, computationally much more expensive.

the difference of the averaged coefficients of any two variable groups is shrinked towards zero according to the maximum correlation between two variables from different groups:

Theorem 16 (*IEN* grouping effect)**.** *Define* $\rho_{j,j'} \coloneqq \boldsymbol{x}_j^\top \boldsymbol{x}_{j'}$*, where* $\boldsymbol{x}_j$ *and* $\boldsymbol{x}_{j'}$ *are standardized predictors. Suppose that* $\hat{\beta}_j \hat{\beta}_{j'} > 0$ *and, without loss of generality,* $j \in \mathcal{G}_1$ *and* $j' \in \mathcal{G}_2$*. Then, it holds that*

$$\frac{1}{\|\boldsymbol{y}\|_2} \left| \frac{1}{p_1} \sum_{g \in \mathcal{G}_1} \hat{\beta}_g - \frac{1}{p_2} \sum_{g \in \mathcal{G}_2} \hat{\beta}_g \right| \tag{5.9}$$

$$= \frac{1}{\|\boldsymbol{y}\|_2} \left| \frac{\mathbf{1}_1^\top \hat{\boldsymbol{\beta}}}{p_1} - \frac{\mathbf{1}_2^\top \hat{\boldsymbol{\beta}}}{p_2} \right| \leq \frac{1}{\lambda_2} \sqrt{2 \left(1 - \max_{j \in \mathcal{G}_1, j' \in \mathcal{G}_2} \{ \rho_{j,j'} \} \right)}. \tag{5.10}$$

Proof. Deferred to Appendix C.3. □

Corollary 2. *The grouping effect of the proposed IEN is identical to that of the EN, when every variable is considered to be a group.*

Proof. When every variable is considered to be a group, we have $M = p, p_1 = \ldots = p_p = 1$, and the third summand in (5.5) boils down to $\lambda_2 \sum_{m=1}^{p} \beta_m^2 = \lambda_2 \|\boldsymbol{\beta}\|_2^2$ and, thus, (2.20) and (5.6) are equivalent. □

Remark 12. *Note that, as desired, the difference of the averaged coefficients of any two variable groups is exactly zero if two variables from different groups are perfectly correlated.*

5.2.4 Numerical Experiments

In this section, we evaluate the grouping effect of the *EN* and the proposed *IEN* and their relative computation times when incorporated into the *T-Rex* framework.

5.2.4.1 Grouping Effect and Solution Path

To compare the solution paths of the *elastic net* and the proposed *informed elastic net*, we first generate a setting with $p = 100$ and 6 active standard normal variables that are split into two independent groups of highly correlated variables $\mathcal{G}_1$ and $\mathcal{G}_2$ (i.e., any pair of the three variables in one group has a correlation of 0.75). The remaining 94 null variables are sampled independently from the standard normal distribution. The coefficient vector $\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_p]^\top$ of the variables is chosen as follows: $\beta_j = 1$ for $j \in \mathcal{G}_1$, $\beta_j = -1$ for $j \in \mathcal{G}_2$ and $\beta_j = 0$ otherwise. The response variable $\boldsymbol{y}$ is generated from the linear regression

model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ is the noise vector. In order to make the grouping effect and the distinction between nulls and active variables visually noticeable, we generated $n = 150$ samples and set the noise variance σ^2 such that the signal-to-noise-ratio $\text{SNR} := \text{Var}(\boldsymbol{X}\boldsymbol{\beta})/\text{Var}(\boldsymbol{\epsilon}) = 3$. To obtain the binary support vectors that are required for the data augmentation presented in Theorem 15, we use single-linkage hierarchical clustering with the pairwise correlations between variables as distance measures and cut the resulting dendrogram at the maximum height that satisfies the conservative condition that there exist no two variables from different clusters with a correlation higher than 0.2.

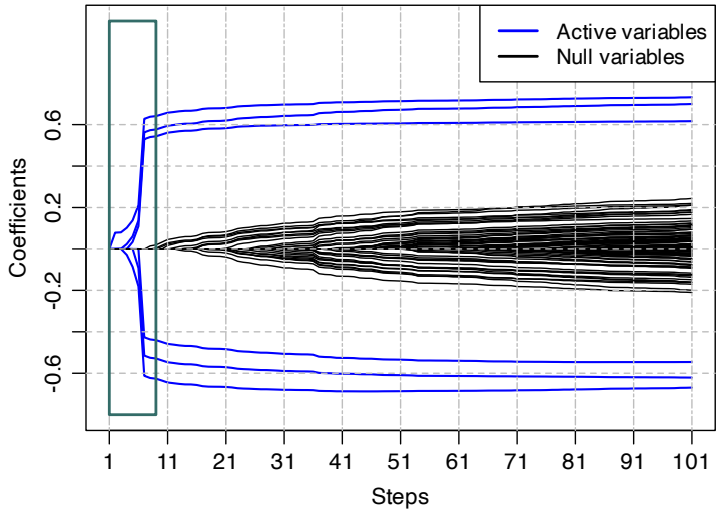
In Figure 5.1, it can be observed that the *EN* and the *IEN* both exhibit the grouping effect in the sense that the coefficients of the two groups of highly correlated active variables are increased in a correlated fashion. We also observe that the grouping effect of the proposed *IEN* is slightly weaker than that of the *EN*. However, since we use the *IEN* as the base selector within the *T-Rex* framework, which terminates the solution paths of all random experiments early, we are primarily interested in the early steps, where a sufficient grouping is observed for both methods, as illustrated in the boxed regions of Figure 5.1.

5.2.4.2 Relative Computation Time

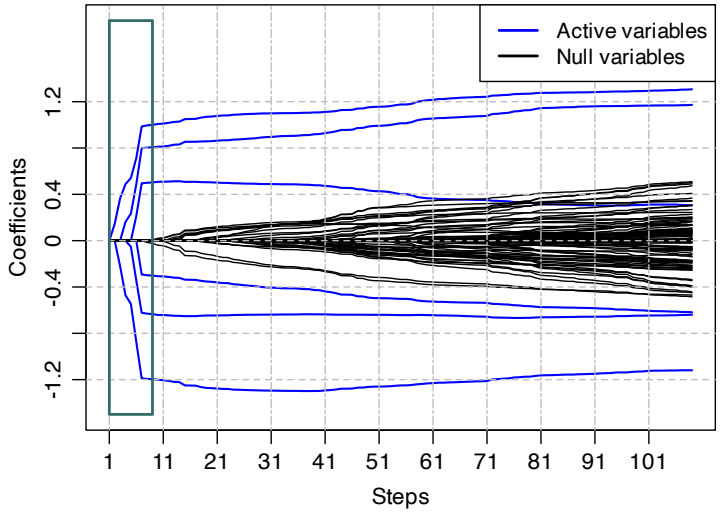
We compare the relative computation times of one random experiment of

1. the original *T-Rex* selector with the *LARS* algorithm as the base selector,
2. the *T-Rex+GVS* selector with the *EN* base selector, and
3. the *T-Rex+GVS* selector with the *IEN* as the base selector

in a setting that is as described in Section 5.2.4.1, except that we fix the number of samples to $n = 50$, set the correlation cutoff of the dendrogram to 0.5 and increase p from 100 to $5{,}000$. The computation times are averaged over 50 Monte Carlo replications. In Figure 5.2, we see that with a growing number of variables p, the relative computation time of the *T-Rex+GVS* selector with the proposed *IEN* as the base selector decreases significantly. In particular, the savings in computation time start to manifest in larger settings with $p \geq 500$ variables, where the *EN* always needs to augment $\boldsymbol{X}$ with p (i.e., number of variables) rows, while the *IEN* only requires augmenting $\boldsymbol{X}$ with M (i.e., number of groups) rows, where $M \ll p$.



(a) *Elastic net (EN)* solution path.



(b) Proposed: *Informed elastic net (IEN)* solution path.

Figure 5.1: Solution paths of the (a) *EN* [ZH05] and (b) the proposed *IEN*.

5.3 Sparse PCA with FDR-Controlled Variable Selection

This section introduces the proposed FDR-controlled PCA methods. First, the general concept of sparse PCA is introduced, the proposed paradigm shift towards FDR-controlled sparse PCA is motivated, and an overview of the major contributions is provided. Then, the methodology of the proposed *T-Rex* PCA methods is presented and verified through numerical exper-

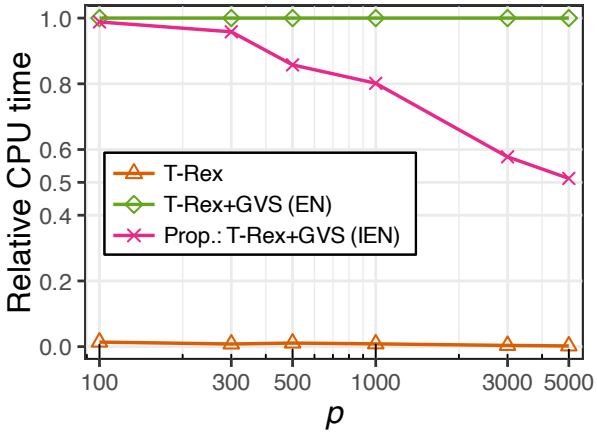


Figure 5.2: Relative computation times of one random experiment with $L = p$ and $T = 1$ of the *T-Rex*, *T-Rex+GVS (EN)*, and the proposed *T-Rex+GVS (IEN)*.

iments. A real-world data factor analysis of the major stocks of the S&P 500 index is presented in Chapter 7.

5.3.1 Introduction to Sparse PCA

Principal component analysis (PCA) aims at mapping large dimensional data to a linear subspace of lower dimension. For this purpose, let us consider consider n samples of p-dimensional observations stored (row-wise) in the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Additionally, let the ordered singular value decomposition (SVD) of $\boldsymbol{X}$ be given by

$$\boldsymbol{X} \overset{\text{SVD}}{=} \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{5.11}$$

where $\mathbf{V} = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_p] \in \mathbb{R}^{p \times p}$ contains the p loading vectors. Then, the rank-M ($M < p$) ordinary principal component analysis (PCA) is commonly used to reduce the data dimension by projecting the data on its M leading principal components (PCs)

$$\mathbf{Z}_M \coloneqq [\boldsymbol{z}_1 \ \cdots \ \boldsymbol{z}_M] \coloneqq \boldsymbol{X}\mathbf{V}_M \coloneqq \boldsymbol{X}[\mathbf{v}_1 \ \cdots \ \mathbf{v}_M]. \tag{5.12}$$

The column vector $\boldsymbol{z}_m = \boldsymbol{X}\mathbf{v}_m$, $m \in \{1, \ldots, M\}$, is called the mth PC, while the associated vector $\mathbf{v}_m$ is referred to as the mth loading vector [Jol03; JC16]. Note that the PCs are thus created from weighted linear combinations of all variables in $\boldsymbol{X}$, which can be problematic in terms of interpretation.

Sparse PCA aims at alleviating the aforementioned drawback of ordinary PCA by imposing some level of sparsity on the loading vectors, i.e., incorporating variable selection in the process of linear dimension reduction [ZHT06; ZX18; US08; Bre+23; Ben+16]. This is generally achieved by casting and solving a trade-off optimization problem of the form

$$\begin{aligned} \underset{\mathbf{V}_M \in \mathbb{R}^{p \times M}}{\text{minimize}} \quad & f(\boldsymbol{X}, \mathbf{V}_M) + \lambda h(\mathbf{V}_M) \\ \text{subject to} \quad & \mathbf{V}_M^\top \mathbf{V}_M = \mathbf{I}_M, \end{aligned} \tag{5.13}$$

where $f(\boldsymbol{X}, \mathbf{V}_M)$ is a data fitting term, $h(\mathbf{V}_M)$ is a sparsity promoting penalty, and $\lambda \in \mathbb{R}^+$ is the corresponding regularization parameter. Such a generic formulation has motivated numerous developments in terms of problem design and optimization methods (see, e.g., [Hu+15; Bre+21; WT08] and references therein). A seminal formulation of sparse PCA ties the problem of penalized maximization of the explained variance (with relaxed orthogonality constraint) to a series of M *elastic net* variable selection problems [ZHT06].

5.3.2 Motivation for FDR-Controlled Sparse PCA

By imposing loading vectors to be sparse, sparse PCA performs the double duty of dimension reduction and variable selection. Nevertheless, sparse PCA methods are often assessed and compared using performance metrics such as the explained variance (EV) of a PC, which is a measure of the variation in the data that is captured by that PC. However, a high EV is not necessarily synonymous with relevant information and these methods are prone to select irrelevant variables. More specifically, sparse PCA algorithms as formulated in (5.13) trade-off the explained variance and the sparsity level and, therefore, suffer from two major issues:

1. Maximizing the explained variance does not inherently yield the most meaningful projection for exploratory data analysis: highly noisy variables will tend to be selected, although not being necessarily informative.

2. Lowering the sparsity to achieve a higher explained variance does not guarantee that, in turn, more meaningful variables have been selected.

These observations motivate controlling sparse PCA variable selection processes with a criterion that ensures that the number false discoveries (i.e., irrelevant variables) used to create a sparse PC is low. Thus, to overcome these issues, we propose an alternative formulation of sparse PCA driven by the FDR. We then leverage the *T-Rex+GVS* selector to automatically

determine an FDR-controlled support of the loading vectors. A major advantage of the resulting *T-Rex* PCA is that no sparsity parameter tuning is required.

5.3.3 The T-Rex PCA Method: Overview of Major Contributions

This work proposes an alternative approach for sparse PCA, where the selection of variables for the loading vectors is driven by the FDR. Although there exist many FDR-controlling methods (e.g., [BH95; BY01; STS04; BC15; Can+18]), only the *T-Rex* framework provides the possibility of solving the *elastic net* based sparse PCA optimization problem in [ZHT06] in an FDR-controlled manner. Thus, our proposed *T-Rex* PCA approach

1. harnesses the *elastic net* based sparse PCA formulation of [ZHT06]
2. and solves it by leveraging the *T-Rex+GVS* selector, which yields
3. FDR-controlled solutions while maximizing the number of selected (informative) variables and implicitly maximizing the explained (non-null) variance.

5.3.4 Methodology

In the following, the proposed *T-Rex* PCA approach is explained and a comprehensive definition of the percentage of explained variance (PEV) for sparse PCA methods is presented.

5.3.4.1 T-Rex PCA Algorithm

We propose to leverage the *T-Rex* selector to obtain FDR-controlled solutions of the formulation of sparse PCA as a collection of the M *elastic net* problems [ZHT06], i.e.,

$$\left\{ \underset{\boldsymbol{\beta}_j \in \mathbb{R}^p}{\text{minimize}} \left\| \boldsymbol{z}_m - \boldsymbol{X}\boldsymbol{\beta}_j \right\|_2^2 + \lambda_1 \left\| \boldsymbol{\beta}_j \right\|_1 + \lambda_2 \left\| \boldsymbol{\beta}_j \right\|_2^2 \right\}_{m=1}^{M} \tag{5.14}$$

where $\lambda_1, \lambda_2 > 0$ are tuning parameters and $\boldsymbol{z}_m$ is the plug-in estimate of the mth PC (i.e., the ordinary PC $\boldsymbol{z}_m = \boldsymbol{X}\mathbf{v}_m$). The parameter λ_1 controls the sparsity level, while the ridge parameter λ_2 determines the strength of the variable grouping effect [ZH05].

Our goal is to obtain FDR-controlled solutions of (5.14) (i.e., $\{\hat{\boldsymbol{\beta}}_m\}_{m=1}^{M}$) that provide a basis of (sparse) loading vectors for the dimension reduction. For this purpose, the ordinary PCs $\boldsymbol{z}_m, m = 1, \ldots, M$, serve as supervising response vectors within the *T-Rex* selector (i.e., $\boldsymbol{y} =$

$\boldsymbol{z}_m$ in Figure 3.2) and we incorporate the *elastic net* as the forward variable selector into the *T-Rex* framework. This is achieved by reformulating the *elastic net* as a *Lasso*-type problem and solving it using the Terminating-*LARS* (*T-LARS*) forward selection algorithm (for details, see Section 5.1 and [Mac+24b]). This approach yields the sparse and FDR-controlled supports of the M loading vectors, i.e.,

$$\widehat{\mathcal{A}}_{L_m^*}(v_m^*, T_m^*), \quad m = 1, \dots, M. \tag{5.15}$$

To convert the supports into loading vectors, we leverage the fact that the loading vectors can be linked to the ridge regression estimator [ZHT06]. That, in combination with the selected active set $\widehat{\mathcal{A}}_m := \widehat{\mathcal{A}}_{L_m^*}(v_m^*, T_m^*)$ as obtained by the *T-Rex* selector, yields

$$\hat{\mathbf{v}}_m = \frac{\hat{\boldsymbol{\beta}}_{m,\text{Ridge}}}{\|\hat{\boldsymbol{\beta}}_{m,\text{Ridge}}\|_2}, \quad m = 1, \dots, M, \tag{5.16}$$

Algorithm 7 *T-Rex* PCA.

1. **Input**: α, K, M, $\boldsymbol{X}$, $\boldsymbol{y}$.
2. **Compute** the SVD of $\boldsymbol{X}$, i.e., $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\mathbf{V}^\top$ and **determine** the ordinary PC matrix $\mathbf{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_M] = \boldsymbol{U}\boldsymbol{D}$ that contains the first $M \leq \min\{n, p\}$ ordinary PCs.
3. **For** $m = 1, \dots, M$ **do**:
   - 3.1. **Run** the *T-Rex* selector with
     - a. the target FDR level α,
     - b. the extended predictor matrices $\widetilde{\boldsymbol{X}}_{m,k} := \left[\boldsymbol{X} \ \mathring{\boldsymbol{X}}_{m,k}\right]$, $k = 1, \dots, K$, and
     - c. the mth PC $\boldsymbol{z}_m$ as the common response for all $\widetilde{\boldsymbol{X}}_{m,k}$, $k = 1, \dots, K$.
   - 3.2. **Obtain** the FDR-controlled support of the mth loading vector $\widehat{\mathcal{A}}_{L_m^*}(v_m^*, T_m^*)$.
   - 3.3. **Compute** the mth loading vector $\hat{\mathbf{v}}_m = \hat{\boldsymbol{\beta}}_{m,\text{Ridge}} / \|\hat{\boldsymbol{\beta}}_{m,\text{Ridge}}\|_2$.
   - 3.4. **Compute** the mth PC $\hat{\boldsymbol{z}}_m = \boldsymbol{X}_{\widehat{\mathcal{A}}_m} \hat{\mathbf{v}}_m$.
4. **Output**:
   - 4.1. *T-Rex* supports $\widehat{\mathcal{A}}_{L_m^*}(v_m^*, T_m^*)$, $m = 1, \dots, M$, and
   - 4.2. *T-Rex* PC matrix $\widehat{\mathbf{Z}} = [\widehat{\boldsymbol{z}}_1 \ \cdots \ \widehat{\boldsymbol{z}}_M]$.

where

$$\hat{\boldsymbol{\beta}}_{m,\text{Ridge}} := \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{z}_m - \boldsymbol{X}_{\widehat{\mathcal{A}}_m}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \tag{5.17}$$

and $\boldsymbol{X}_{\widehat{\mathcal{A}}_m}$ contains only the predictors corresponding to $\widehat{\mathcal{A}}_m$. The *T-Rex* PCs are then given by $\hat{\boldsymbol{z}}_m = \boldsymbol{X}_{\widehat{\mathcal{A}}_m}\hat{\mathbf{v}}_m$, $m = 1, \ldots, M$.

Note that $\hat{\mathbf{v}}_m$ is independent of λ_2 because of the scaling with the ℓ_2-norm of $\hat{\boldsymbol{\beta}}_{m,\text{Ridge}}$ [ZHT06] and, therefore, we simply set $\lambda_2 = 10^{-6}$. A major advantage of the *T-Rex* selector framework is that when incorporating the *elastic net* into it, the choice of λ_1 becomes obsolete, since the random experiments are terminated after T^* dummies have entered the solution paths such that the FDR is controlled at the user-defined target level α, which corresponds to choosing λ_1 for each random experiment such that an FDR-controlled selected active set $\widehat{\mathcal{A}}_{L_m^*}(v_m^*, T_m^*)$ is obtained. This is especially important, since sparse PCA relies on hand-selecting the sparsity parameter specifically for every problem [ZHT06]. The pseudocode of the proposed *T-Rex* PCA is given in Algorithm 7.

The obtained FDR-controlled selected active sets can also be used to threshold the loading vectors of the ordinary PCA. Thus, in addition to the *T-Rex* PCA, we also propose the *T-Rex* Thresholded PCA, which is obtained by thresholding each loading vector $\mathbf{v}_m$ such that only the $|\widehat{\mathcal{A}}_{L_m^*}(v_m^*, T_m^*)|$ strongest loadings remain active (i.e., non-zero). The thresholded loading vector is then rescaled by its ℓ_2-norm to ensure that $\|\hat{\mathbf{v}}_m\|_2 = 1$.

5.3.4.2 Percentage of Explained Variance

The explained variance (EV) in ordinary PCA is defined by $\text{tr}(\widehat{\mathbf{Z}}^\top\widehat{\mathbf{Z}})$, where $\text{tr}(\cdot)$ is the trace-operator. Since we are interested in the variance that corresponds to signal and not pure null (i.e., non-signal) components, we define the percentage of explained variance (PEV) as follows:

Definition 20. *Let $\widehat{\mathbf{V}} = \widehat{\mathbf{V}}_{\mathcal{A}} + \widehat{\mathbf{V}}_{\mathcal{A}^C} \in \mathbb{R}^{p\times M}$, where $\widehat{\mathbf{V}}_{\mathcal{A}}$ is the estimated loading matrix whose entries are set to zero except for the positions containing true active loadings and $\widehat{\mathbf{V}}_{\mathcal{A}^C}$ is the estimated loading matrix whose true active loadings are set to zero. Then, $\widehat{\mathbf{Z}} = \boldsymbol{X}\widehat{\mathbf{V}} = \boldsymbol{X}\widehat{\mathbf{V}}_{\mathcal{A}} + \boldsymbol{X}\widehat{\mathbf{V}}_{\mathcal{A}^C} =: \widehat{\mathbf{Z}}_{\mathcal{A}} + \widehat{\mathbf{Z}}_{\mathcal{A}^C}$ and the signal EV, mixed EV, and null EV are defined by*

$$\text{EV} := \text{tr}(\widehat{\mathbf{Z}}^\top\widehat{\mathbf{Z}}) = \underbrace{\text{tr}(\widehat{\mathbf{Z}}_{\mathcal{A}}^\top\widehat{\mathbf{Z}}_{\mathcal{A}})}_{\textit{Signal EV}} + \underbrace{2\,\text{tr}(\widehat{\mathbf{Z}}_{\mathcal{A}}^\top\widehat{\mathbf{Z}}_{\mathcal{A}^C})}_{\textit{Mixed EV}} + \underbrace{\text{tr}(\widehat{\mathbf{Z}}_{\mathcal{A}^C}^\top\widehat{\mathbf{Z}}_{\mathcal{A}^C})}_{\textit{Null EV}}, \tag{5.18}$$

and the PEV is defined by

$$\text{PEV} := \frac{\text{EV}}{\textit{Signal EV} + \textit{Mixed EV}}. \tag{5.19}$$

Our goal is to explain the signal and mixed EV with few PCs and sparse loadings to allow for interpretability of the obtained PCs. Non-sparse PCA methods or methods that do not provide accurate estimates of $\widehat{\mathbf{V}}$ are prone to have a high null EV and, therefore, capture variance that merely corresponds to null (i.e., non-active) variables/loadings. In that case, the PEV in Definition 20 exceeds 100%, which indicates an inferior performance of the respective method. Moreover, since the orthogonality constraint in (5.13) is dropped for sparse PCA methods, we replace the EV in Definition 20 by the adjusted EV that accounts for the lack of orthogonality of the loading vectors as suggested in [ZHT06]. The adjusted EV is defined by

$$\text{EV}_{\text{adj}} := \sum_{m=1}^{M} r_{m,m}^2, \tag{5.20}$$

where $r_{m,m}$ is the mth diagonal element of the upper triangular matrix $\mathbf{R}$ from the QR-decomposition of $\widehat{\mathbf{Z}}$ (i.e., $\widehat{\mathbf{Z}} = \mathbf{QR}$).

5.3.5 Numerical Experiments

We consider a high-dimensional data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with $n = 50$ samples, $p = 100$ variables, and centered columns that follows the sparse M-factor model

$$\boldsymbol{X} = \mathbf{Z}\mathbf{V}^\top + \boldsymbol{E} = [\boldsymbol{z}_1 \cdots \boldsymbol{z}_M][\mathbf{v}_1 \cdots \mathbf{v}_M]^T + [\boldsymbol{\epsilon}_1 \cdots \boldsymbol{\epsilon}_p] \tag{5.21}$$

$$= \begin{bmatrix} z_{1,1} & \cdots & z_{1,M} \\ z_{2,1} & \cdots & z_{2,M} \\ \vdots & & \vdots \\ z_{n,1} & \cdots & z_{n,M} \end{bmatrix} \begin{bmatrix} v_{1,1} & \cdots & v_{p,1} \\ v_{1,2} & \cdots & v_{p,2} \\ \vdots & & \vdots \\ v_{1,M} & \cdots & v_{p,M} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,1} & \cdots & \epsilon_{1,p} \\ \epsilon_{2,1} & \cdots & \epsilon_{2,p} \\ \vdots & & \vdots \\ \epsilon_{n,1} & \cdots & \epsilon_{n,p} \end{bmatrix} \tag{5.22}$$

where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$ are Gaussian factors (i.e., $z_{i,m} \sim \mathcal{N}(0, \sigma_m^2)$), $\mathbf{v}_1, \ldots, \mathbf{v}_M$ are the corresponding sparse loading vectors of the factors (i.e., $v_{j,m} \in [0, 1]$), and $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_p$ are Gaussian noise vectors (i.e., $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$). We generate $M = 3$ factors with standard deviations $(\sigma_1, \sigma_2, \sigma_3) = (5, 3, 1)$. For each of the three factors, p_1 true active loadings are randomly

selected among only the first 30 out of $p = 100$ variables to simulate the more challenging case of overlapping loadings among the three factors. The values of p_1 are varied over a range from 1 to 30. The values of the randomly selected loadings are set to 0.9 (i.e., $v_{j,m} = 0.9$). The noise variance σ^2 is chosen such that the signal-to-noise ratio (SNR) is controlled over a range from -10 dB to $+10$ dB. The SNR is defined by

$$\text{SNR} := 10 \log_{10} \left(\frac{\text{Var}\left[\text{vec}\left(\mathbf{Z}\mathbf{V}^\top\right)\right]}{\text{Var}\left[\text{vec}\left(\boldsymbol{E}\right)\right]} \right), \tag{5.23}$$

where $\text{Var}(\boldsymbol{a})$ and $\text{vec}(\boldsymbol{A})$ denote the sample variance of a vector $\boldsymbol{a}$ and the vectorization operator that stacks the columns of a matrix $\boldsymbol{A}$ on top of each other, respectively. Finally, we set all simulation parameters that are not varied as follows: SNR = 0 dB, $\alpha = 10\%$ (target FDR level), $K = 20$ (number of *T-Rex* random experiments), $p_1 = 5$ (number of true active loadings) in Figure 5.3, and $p_1 = 10$ in Figure 5.4. The following three benchmark methods are considered:

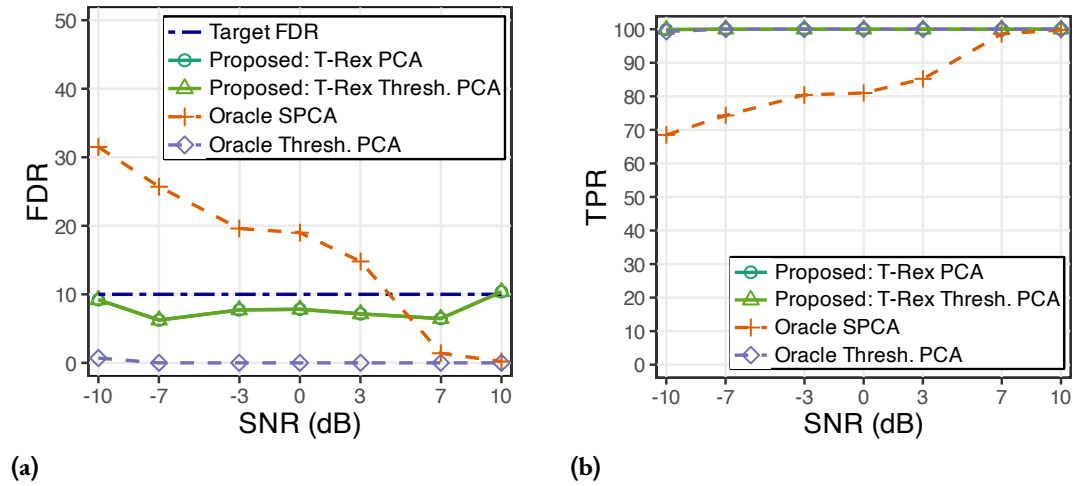
1. Ordinary (non-sparse) PCA.
2. The oracle thresholded PCA solution, which is obtained by thresholding each loading vector $\mathbf{v}_m$ such that only the p_1 strongest loadings remain active (i.e., non-zero). The thresholded loading vector is then rescaled by its ℓ_2-norm to ensure that $\|\hat{\mathbf{v}}_m\|_2 = 1$.
3. The oracle sparse PCA (oracle SPCA) solution of (5.14), which is obtained by choosing the sparsity parameter λ_1 for each plug-in PC $\boldsymbol{z}_m$ such that only p_1 loadings remain active.

Note that we are considering the best-case performances of the benchmark methods. In practice, however, only the proposed *T-Rex* PCA methods are feasible without choosing any sparsity parameter.

The results are averaged over 200 Monte Carlo replications. A discussion is provided in the captions of Figures 5.3 and 5.4.

5.4 Summary

The *T-Rex+GVS* selector for FDR-controlled grouped variable selection in high-dimensional settings has been proposed and its FDR control property has been empirically verified in the



(a) (b)

Figure 5.3: For the first PC, the proposed *T-Rex* PCA methods empirically control the FDR at a level of 10% while achieving an optimal TPR of 100% even at low SNRs. Only the infeasible oracle thresholded PCA achieves the same TPR at an FDR of almost zero. Except for high SNRs, the oracle SPCA is dominated by all other methods.

presence of groups of highly correlated variables. In order to reduce the computation time of the *T-Rex+GVS* selector that uses the *elastic net (EN)* as a base selector, the *informed elastic net (IEN)* has been developed. Incorporating the proposed *IEN* as a base selector into the *T-Rex+GVS* framework yields a significant reduction in computation time. The proposed methods outperformed existing state-of-the-art benchmark methods in terms of the TPR (i.e., power) on a high-dimensional simulated GWAS (see Chapter 6).

The *T-Rex+GVS* selector has been leveraged to develop an FDR-controlling method for sparse PCA. The proposed *T-Rex* PCA and *T-Rex* Thresholded PCA methods perform the double duty of dimension reduction and variable selection while controlling the FDR of the sparse loading vectors. They do not require any tuning of sparsity parameters and are capable of explaining the signal variance in the data with few PCs, which allows for meaningful interpretations of the PCs. The proposed methods have shown a promising performance in simulated data. Moreover, they have proven to be useful for revealing the interdependencies among stocks in a factor analysis of the 500 index (see Chapter 7).

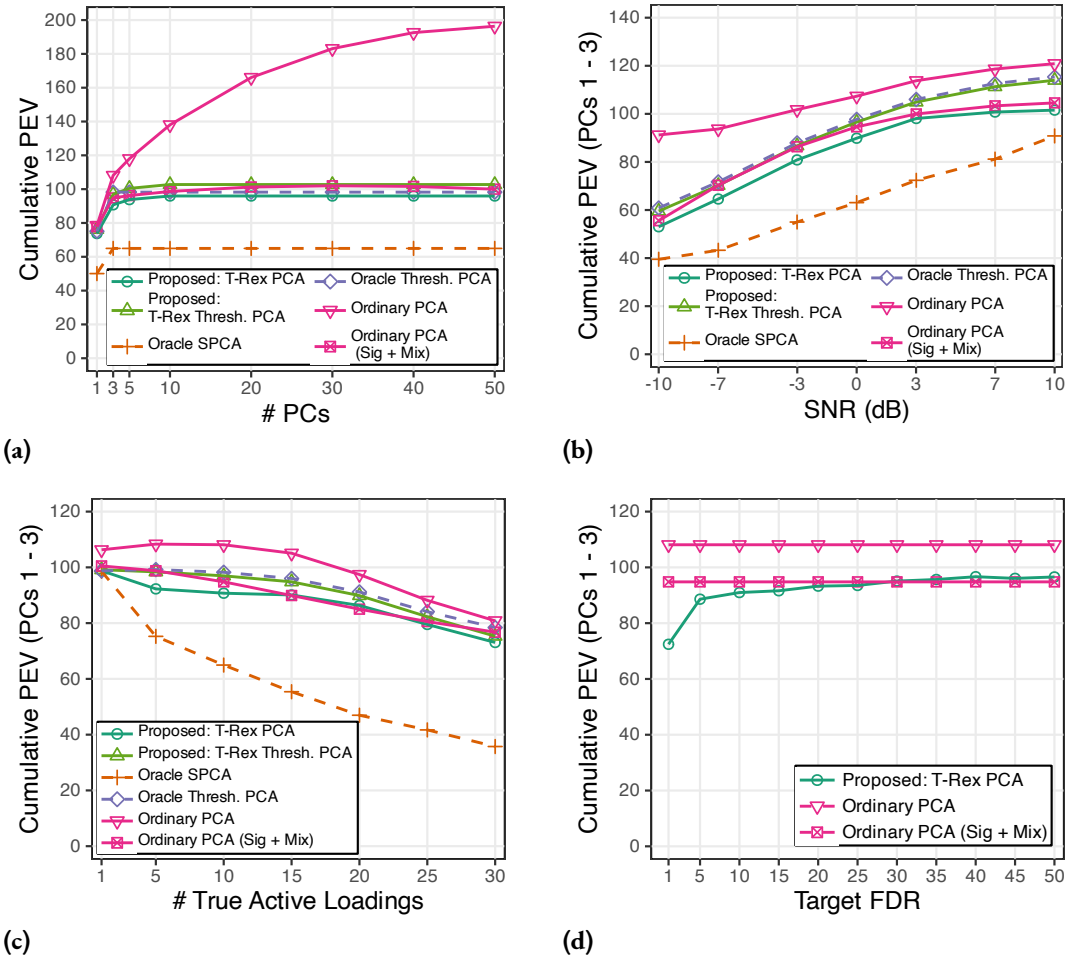


Figure 5.4: Cumulative percentage of explained variance (PEV): (a) - (c) As desired, the proposed *T-Rex* PCA and *T-Rex* Thresholded PCA require only very few PCs to explain the signal and mixed variance while not explaining any additional variance that is purely associated with null loadings. The oracle SPCA is outperformed by all other methods and the ordinary PCA explains all the variance in the data, including the variance that is merely associated with null loadings. (d) The cumulative PEV is not very sensitive with respect to the choice of the target FDR level for the *T-Rex* PCA, which allows the user to set almost any (preferably low) target FDR and still achieve a high cumulative PEV.

He who seeks for methods without having a definite problem in mind seeks in the most part in vain.

David Hilbert

Every scientist dreams of doing something that can help the world.

Tu Youyou

6

Applications in Biomedical Engineering

This chapter presents advanced biomedical applications that have been addressed using the proposed *T-Rex* methods from Chapters 3, 4, and 5. In Section 6.1, the results of a simulated genome-wide association study (GWAS) are presented. In Section 6.2, a real-world data human immunodeficiency virus type-1 (HIV-1) drug resistance analysis is conducted. Section 6.3 presents the results of a real-world data breast cancer survival analysis.

The main content of this chapter is based on the publications [MMPewa], [MMP22], [MMP23b], [MMP23a], and [MMPewb]. The results have been produced using the developed open source R software packages TRexSelector [Mac+24c] and tlars [Mac+24b].

6.1 Simulated Genome-Wide Association Study

The proposed *T-Rex* methods and the state-of-the-art benchmark methods are used to conduct a high-dimensional simulated case-control GWAS. The size of the GWAS was chosen, such that it was still practically feasible to compute the computationally intensive benchmark methods. The general goal is to detect the single nucleotide polymorphisms (SNPs) that are associated with a disease of interest (i.e., active variables), while keeping the number of selected SNPs that are not associated with that disease (i.e., null variables) low.

6.1.1 GWAS in a Nutshell

The general goal of a genome-wide association study (GWAS) is to discover reproducible associations between single nucleotide polymorphisms (SNPs), which are base-pair variations at specific positions on the genome, and a phenotype (e.g., trait, disease) of interest [Uff+21]. An important resource for genomics data is the UK biobank [Sud+15]. It contains genome sequences for roughly $500{,}000$ participants with about $800{,}000$ genotyped SNPs and about $120{,}000{,}000$ imputed SNPs and thousands of phenotypes. A systematic collection of associations that have been discovered through GWAS can be found in the curated GWAS catalog [Bun+19].

The discoveries made in GWAS can be promising candidates for further experiments and the study of potentially functional associations, which require time and cost intensive follow-up investigations [GC18]. Hence, it is of utmost importance to keep the number of false discoveries low while discovering as many true associations as possible. This can be achieved by using FDR-controlling methods. While classical FDR-controlling methods based on marginal p-values have been used, multivariate FDR-controlling methods such as the proposed *T-Rex* methods or the *model-X* knockoff methods usually have a higher TPR and other advantages [Can+18].

In order to conduct an FDR-controlled GWAS, the following data is required:

1. A predictor matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_p]$ that contains p SNPs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ as columns, where $\boldsymbol{x}_j = [x_{1j} \cdots x_{nj}]^\top$ contains n observations of the jth SNP. That is, the ith row of $\boldsymbol{X}$ contains the measurements of all SNPs for the ith subject.
2. A response vector $\boldsymbol{y} = [y_1 \cdots y_n]^\top$ that contains the phenotypes of all n subjects. These can be measurements of the disease progression or, in a simple case-control study, the value "1" for cases and the value "0" for controls.
3. The target FDR level $\alpha \in [0, 1]$.

In the following, we present the results of a relatively small FDR-controlled simulated GWAS to showcase the performance of the proposed and the benchmark methods.

6.1.2 Setup and Problem Statement

The genotypes of 700 cases and 300 controls are simulated based on haplotypes from phase 3 of the International HapMap project [The10] using the software HAPGEN2 [SMD11]. The

software takes into account biological characteristics of genomics principles to simulate realistic genotypes (i.e., predictor matrix $\boldsymbol{X}$) with known ground truth. We simulated 10 randomly selected disease loci on the first 20,000 SNPs of chromosome 15 (contains 42,351 SNPs in total) with randomly selected risk alleles (either 0 or 1 with $\mathbb{P}(\text{“0”}) = \mathbb{P}(\text{“1”}) = 0.5$) and with the heterozygote risks and the homozygote risks being sampled from the uniform distribution on the intervals $[1.5, 2]$ and $[2.5, 3]$, respectively. Since we are conducting a case-control study, the control and case phenotypes are 0 and 1, respectively. Note that the SNPs and the phenotype represent the candidate variables and the response, respectively, while the disease loci represent the indices of the active variables. Thus, we have $p_1 = 10$ active variables and $p_0 = 19{,}990$ null variables. The number of observations is $n = 1{,}000$ (700 cases and 300 controls).

The genotype matrix, i.e., the matrix $\boldsymbol{X}$ containing the SNPs as columns consists of groups of highly correlated SNPs. This is due to a phenomenon called linkage disequilibrium [Rei+01]. Figure 6.1 visualizes the correlation matrix of the first 150 SNPs in $\boldsymbol{X}$. We can observe the dependency structure among the predictors/SNPs that form groups of highly correlated variables.

Although the general goal of a GWAS is to detect associations between SNPs and phenotypes, the overarching goal of such studies is not to find specific SNPs/variables that are associated with a phenotype of interest but rather to find the groups of highly correlated SNPs/variables that point to the broader locations on the genome that are associated with the disease of interest. Therefore, in genomics research, it is a standard procedure to apply a preprocessing method called *SNP pruning* before applying any variable selection method (see, e.g., [SSC19]). The main idea behind *SNP pruning* is to cluster the SNPs into groups of highly correlated SNPs using a dendrogram and to select one representative from each group of highly correlated SNPs. After this procedure has been carried out, we are left with an SNP matrix whose dimension is reduced and that exhibits only weak dependencies among the representative SNPs.

For the simulated GWAS, we generated 100 data sets satisfying the above specifications of n, p, p_1, p_0, etc. According to the authors, HAPGEN2 uses the time of the current day in seconds to set the seed of the random number generator, and, therefore multiple simulations should not be started very close in time to avoid identical results. Therefore we have generated the data sets sequentially and since generating a single data set took roughly six minutes, a sufficient

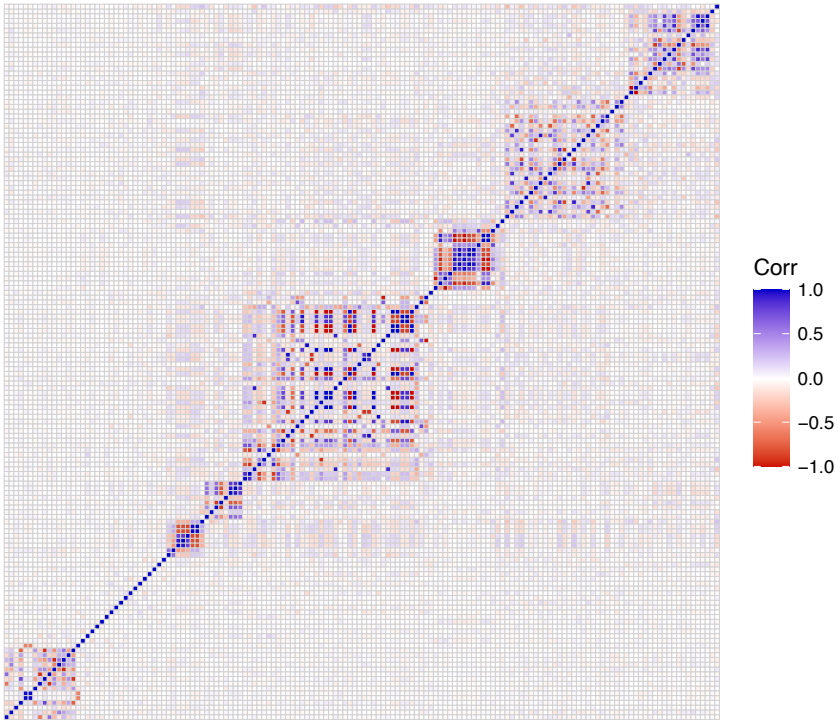


Figure 6.1: The heatmap visualizes the correlation matrix of the first 150 SNPs that were generated using the software HAPGEN2 [SMD11].

time period between the starts of consecutive simulations was allowed.[1] The reported results in Section 6.1.4 have been averaged over these 100 data sets.

6.1.3 Preprocessing

The preprocessing is carried out as suggested in [SSC19] and on the accompanying website.[2] That is, SNPs with a minor allele frequency or call rate lower than 1% and 95%, respectively, are removed. Additionally, SNPs that violate the Hardy-Weinberg disequilibrium with a cut-off of 10^{-6} are removed. Since proximate SNPs are highly correlated, the remaining SNPs are clustered using SNP pruning that ensures that there exist no absolute sample correlations above 0.75 between any two SNPs belonging to different clusters. The resulting average number of clusters is 8211 while the minimum and maximum numbers of clusters are 8120 and

[1] The data sets were generated on a compute node of the Lichtenberg High-Performance Computer of the Technische Universität Darmstadt that consists of two "Intel® Xeon® Platinum 9242 Processors" with 96 cores and 384 GB RAM (DDR4-2933) in total.

[2] URL: https://web.stanford.edu/group/candes/knockoffs/tutorials/gwas_tutorial.html (last access: June 26, 2024).

8326, respectively. Each cluster is represented by the strongest cluster representative which is selected by computing the marginal p-values using the Cochran-Armitage test based on 20% of the data and picking the SNP with the smallest p-value. The marginal p-values that will be plugged into the *BH* method and the *BY* method are also computed using the Cochran-Armitage test but with the full data set.

6.1.4 Results

The reported results in Tables 6.1 and 6.2 have been averaged over 100 data that have been generated as described in Section 6.1.2 and preprocessed as described in Section 6.1.3. Additionally, Figure 6.2 shows how the FDP and TPP vary around the mean using box plots. The *T-Rex* selector demonstrates its applicability to GWAS, as it is the only FDR-controlling method with a positive TPR, and its sequential computation time is about 4 minutes (vs. more than 12 hours for the knockoff methods). A detailed discussion of the results is given in the captions of Table 6.1 and Figure 6.2.

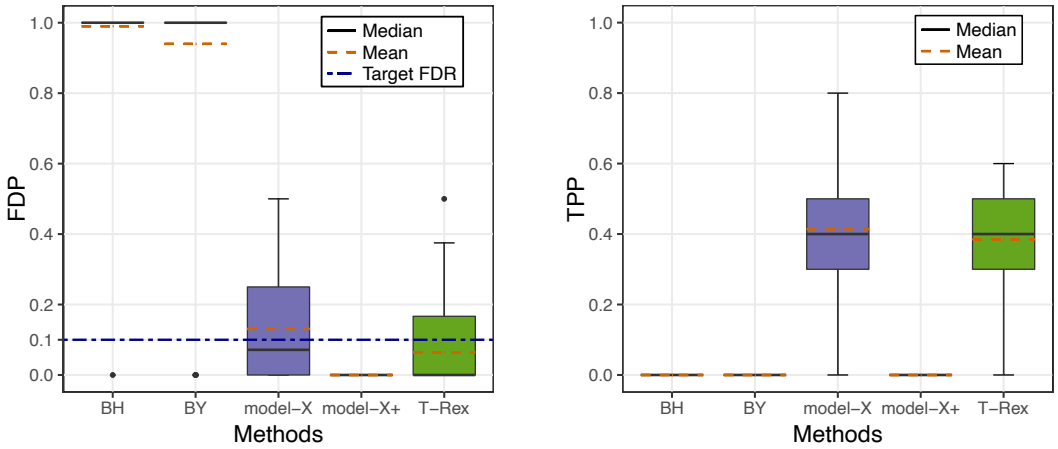
Note that the results of the *Screen-T-Rex* methods are reported in a separate table (Table 6.2). The reason for this is that the computations were conducted several months apart, and due to the energy crisis in Europe, the CPUs of the Lichtenberg High-Performance Computer of the Technische Universität Darmstadt were operating at a reduced clock frequency. The tables are therefore based on two different reference computation times. Thus, only the relative but not the absolute computation times are representative.

Due to the comparatively high computation times of the *T-Rex+GVS* methods, a smaller GWAS is considered for these methods. That is, we generate 100 data sets that each contain 500 cases, 200 controls, the first 1,000 SNPs on Chromosome 15, and 10 true active SNPs. Everything else is as described in Sections 6.1.2 and 6.1.3, except for the SNP pruning. That is, in contrast to the preprocessing in Section 6.1.3, we do not carry out SNP pruning to reduce the dimension of the data but keep all SNPs. Since the ultimate goal of a GWAS is to detect disease positions on the genome and not specific SNPs, it is reasonable to consider groups of highly correlated SNPs as active if they contain a disease SNP (see, e.g., [Can+18; SSC19]). In this regard, a group of highly correlated SNPs is defined as a collection of SNPs of which no SNP has a correlation higher than $\rho_{\text{thr}} = 1/3$ with an SNP from another collection. The choice of ρ_{thr} is based on the same reasoning as in Algorithm 6.

Figure 6.3 presents the box plots of the false discovery proportion (FDP) and the true positive

Table 6.1: The proposed *T-Rex* selector is the only method whose average FDP lies below the target FDR level of 10% while achieving a non-zero power. The only competitor that provably possesses the FDR control property, namely the *model-X* knockoff+ method, has an average FDP of 0% but also an average TPP of 0%, i.e., it has no power. The *model-X* knockoff method exceeds the target FDR level. The computationally cheap procedure of plugging the marginal *p*-values into the *BH* method or the *BY* method, which has been a standard procedure in GWAS, fails in this high-dimensional setting. The sequential computation time of the proposed *T-Rex* selector in combination with the extended calibration algorithm in Algorithm 3 is roughly 4 minutes as compared to more than 12.5 hours for the *model-X* methods. That is, the *T-Rex* selector is 183 times faster than its strongest competitors. Note that this is only a comparison of the sequential computation times. Since the random experiments of the proposed *T-Rex* selector are independent and, therefore, can be run in parallel on multicore computers, an additional substantial speedup can be achieved.

Methods	**FDR control?**	**Average FDP (in %)**	**Average TPP (in %)**	**Average sequential computation time (hh:mm:ss)**	**Average relative sequential computation time**
T-Rex	✓	**6.45**	**38.50**	**00:04:05**	**1**
model-X+	✓	0.00	0.00	12:32:47	183.71
model-X	✗	13.07	41.40	12:32:47	183.71
BY	✗	94.00	0.00	00:00:00	0.00
BH	✗	99.00	0.00	00:00:00	0.00



(a) FDP box plots.

(b) TPP box plots.

Figure 6.2: The proposed *T-Rex* selector is the only method that has an average FDP below the target FDR level and that has a non-zero power. Note that the FDP can be different across the realizations and even for FDR-controlling methods it is not necessarily below the target level for every realization. We use box plots to visualize the distribution of the results and give the reader a sense of how the FDP and TPP (i.e., power) vary around the mean.

Table 6.2: For all methods, the average achieved FDP is lower than the average estimated/target FDR, i.e., all methods control the FDR. The sequential computation time (of both versions) of the proposed *Screen-T-Rex* selector is more than three orders of magnitude lower than that of the *model-X* knockoff+ method. Nearly one order of magnitude is gained compared to the original *T-Rex* selector. Applying Algorithm 4 with $\alpha = 10\%$, $\alpha_l = 5\%$, and $\alpha_u = 20\%$, yields an average FDP and TPP of 15.96% and 47.2%, respectively, without requiring Step 2.3. of Algorithm 4.

Methods	**FDR control?**	**Av. FDP (in %)**	**Av. estimated/target FDR (in %)**	**Av. TPP (in %)**	**Av. sequential comp. time (hh:mm:ss)**	**Av. relative sequential comp. time**
Proposed:						
1. Screen-T-Rex (ordinary)	✓	15.96	18.57	47.2	**00:00:44**	**1**
2. Screen-T-Rex (conf.-based)	✓	10.16	12.5	31.7	00:00:45	1.02
Benchmarks:						
3. T-Rex	✓	6.45	10	38.5	00:06:39	8.88
4. model-X+	✓	0	10	0	20:00:38	1601.39

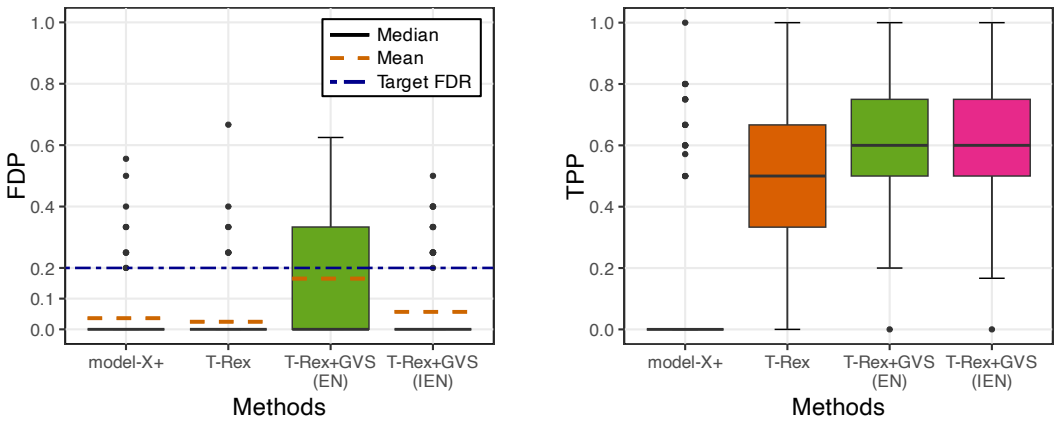


Figure 6.3: The proposed *T-Rex+GVS* methods have the highest TPR values (i.e., average TPP), while their FDR values (i.e., average FDP) stay below the target level of 20%.

proportion (TPP), and the means of the FDP, which are estimates of the FDR, since the FDR is defined as the expectation of the FDP. First, we observe that all methods control the FDR. The median TPP of the benchmark *model-X* knockoff method is zero. The *T-Rex+GVS (EN)* selector shows a significant improvement in TPP compared to the original *T-Rex* selector. As desired, the significant increase in TPP is also achieved by the proposed *T-Rex+GVS (IEN)*

selector, while the FDP is much lower compared to that of the *T-Rex+GVS (EN)* selector. Thus, in this GWAS use-case, the proposed *T-Rex+GVS (IEN)* selector dominates the *T-Rex+GVS (EN)* selector, while exhibiting a much lower computation time, especially in large-scale high-dimensional settings (see Figure 5.2 in Chapter 5).

6.2 HIV-1 Drug Resistance Analysis

In order to also compare the proposed *T-Rex* selector and *Screen-T-Rex* selector from Chapter 3 against FDR-controlling methods for the low-dimensional setting, we consider a low-dimensional benchmark HIV-1 data set that was described and analyzed in [Rhe+05; Rhe+06] and served as a benchmark data set in [BC15]. It can be downloaded from a Stanford University database.[3] As benchmark methods for the low-dimensional setting, we consider the *BH* method and the *BY* method (see Section 2.2.2).

6.2.1 Problem Statement and Setup

Many antiretroviral drugs are used in HIV-1 infection therapies. However, mutations may decrease the susceptibility to some drugs and, thus, lead to an increased drug resistance of the virus. Therefore, it is desired to detect mutations associated with resistance against existing drugs to determine which drugs to use for treating HIV-1 and to develop new drugs to which mutated HIV-1 viruses are highly susceptible.

The same setup as in [BC15] is used. That is, the same preprocessing steps are applied and the same benchmark mutation positions from treatment-selected mutation (TSM) lists in [Rhe+05] are considered. Each drug's response variable contains measurements of drug resistance, while the predictor matrix contains binary data that only distinguishes between the existence or non-existence of the jth mutation, $j = 1, \ldots, p$, in the ith observation, $i = 1, \ldots, n$.

6.2.2 Results

The performance of the proposed *Screen-T-Rex* selector and the benchmark methods in detecting the mutations that are associated with HIV-1 drug resistance for several *protease inhibitor* (PI)-type drugs is assessed. An overview of the dimensions of the data sets correspond-

[3] URL: https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/ (last access: June 26, 2024).

Table 6.3: Results for the HIV-1 PI-type drugs.

Drug	n	p	Target FDR	Est. FDR (ordinary)	Est. FDR (conf.-based)
Amprenavir (**APV**)	767	201	3 %	3.57 %	3.70 %
Atazanavir (**ATV**)	328	147	3 %	4.76 %	0.00 %
Indinavir (**IDV**)	825	206	3 %	3.33 %	3.33 %
Lopinavir (**LPV**)	515	184	3 %	3.85 %	0.00 %
Nelfinavir (**NFV**)	842	207	3 %	3.70 %	0.00 %
Ritonavir (**RTV**)	793	205	3 %	3.33 %	2.86 %
Saquinavir (**SQV**)	824	206	3 %	3.45 %	0.00 %

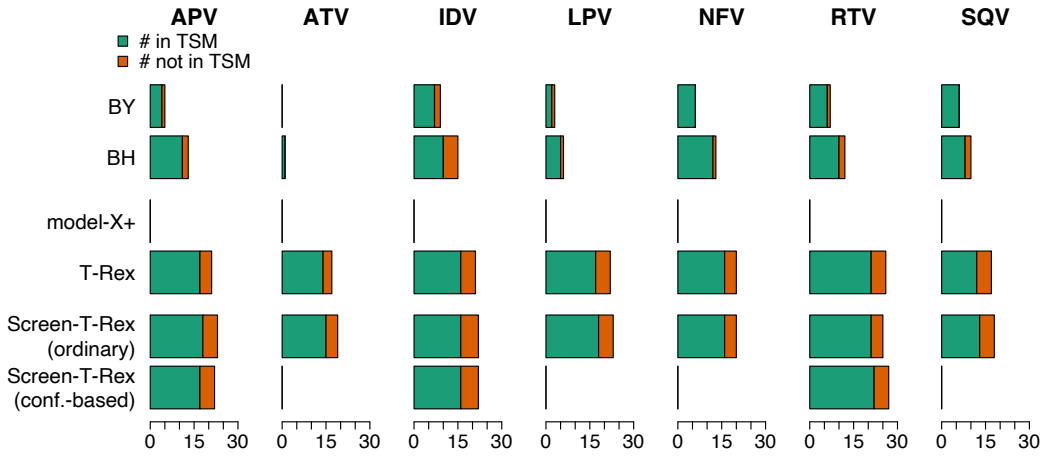


Figure 6.4: Number of selected mutations that are reported (green) and not reported (orange) in TSM lists for HIV-1 PI-type drugs.

ing to the PI-type drugs, the fixed target FDR level for the benchmark methods, and the estimated FDR levels by the proposed ordinary and confidence-based *Screen-T-Rex* selectors are provided in Table 6.3.

The proposed algorithm for screening biobanks (i.e., Algorithm 4 in Section 3.7.1.3) is applied with a fixed target FDR for the *T-Rex* selector of $\alpha = 3\%$, and FDR bounds for the *Screen-T-Rex* selector of $\alpha_l = 2\%$ (lower bound), and $\alpha_u = 4\%$ (upper bound). Thus, whenever the self-estimated FDR of the *Screen-T-Rex* selector does not lie between the user-specified lower and upper bounds, the result of the *T-Rex* selector is favored.

Table 6.3 shows that the result of the *T-Rex* selector is only favored for ATV. For APV and IDV,

the results of the confidence-based *Screen-T-Rex* selector are favored. For the remaining drugs, the results of the ordinary *Screen-T-Rex* selector are favored. Figure 6.4 shows the number of selected mutations that have been previously reported (green) and not reported (orange) in TSM lists for HIV-1 PI-type drugs. The *T-Rex* methods dominate the benchmark methods in terms of the number of selected mutations reported in TSM lists. Moreover, especially the fast ordinary *Screen-T-Rex* selector selects in all cases almost as many mutations as the *T-Rex* selector.

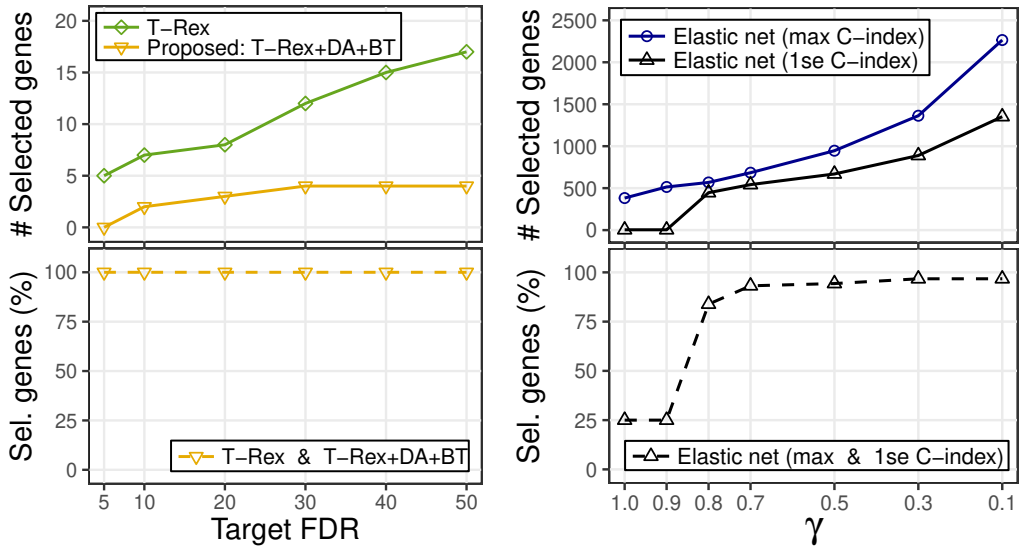
Summarizing, the results indicate that it is reasonable to follow the strategy of the proposed screening biobanks algorithm (i.e., Algorithm 4 in Section 3.7.1.3). That is, it is computationally beneficial to use the fast *Screen-T-Rex* selector to screen through large biobanks and only invoke the *T-Rex* selector whenever the self-estimated FDR of the *Screen-T-Rex* selector is not acceptable.In this example, the results of the *Screen-T-Rex* selector are acceptable for most drugs

6.3 Breast Cancer Survival Analysis

Identifying the few genes that affect the survival time of cancer patients is an important task in the development of new therapies and personalized medicine [KA17]. In this section, we consider gene expression and survival time data from the open source resource The Cancer Genome Atlas (TCGA) [TCW15; Col+16]. In order to detect the genes that are truly associated with the survival time of breast cancer patients, we conduct an FDR-controlled breast cancer survival analysis.

6.3.1 TCGA Breast Cancer Data

The gene expression levels are derived from the RNA-sequencing (RNA-seq) count data. The raw RNA-seq count data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ contains $n = 1{,}095$ samples (i.e., breast cancer patients) and $p = 19{,}962$ protein coding genes. After two standard preprocessing steps, which are removing all genes with extremely low expression levels (i.e., where the sum of the RNA-seq counts is less than 10) and performing a standard variance stabilizing transformation on the count data using the DESeq2 software [LHA14], $p = 19{,}405$ candidate genes are left. The response vector $\boldsymbol{y} \in \mathbb{R}^n$ contains the log-transformed survival times of the patients. After removing missing and uninformative entries (i.e., entries with a survival time of zero days) from $\boldsymbol{y}$, $n = 1{,}072$ samples are left. During the study, the event (i.e., death) occurred



(a) *T-Rex* methods **(b)** Cox *Lasso* and *EN*

Figure 6.5: Number of selected genes in the TCGA breast cancer survival analysis study.

for only 149 patients, while 923 patients were either still alive after the end of the study or dropped out of the study. That is, the survival times of 923 patients are right censored. This is dealt with by treating these entries in $\boldsymbol{y}$ as missing data and imputing them using the well-known Buckley-James estimator [BJ79].

6.3.2 Methods and Results

As motivated in Chapter 4, the most suitable method for this application is the *T-Rex+DA+BT* selector. The considered benchmark methods are the Cox proportional hazards *Lasso* and *elastic net* [Sim+11], which are specifically designed for censored survival data, and the ordinary *T-Rex* selector from Chapter 3. The elastic net Cox model requires the tuning of two parameters, i.e., a sparsity parameter λ and a mixture parameter $\gamma \in [0, 1]$ that balances a convex combination of the ℓ_1- and ℓ_2-norm regularization terms. Here, $\gamma = 1$ sets the ℓ_2 regularization term to zero and yields the *Lasso* solution. As suggested in [Sim+11], we evaluate a range of values for γ and, for each fixed γ, we perform 10-fold cross-validation to choose λ. We consider, as suggested by the authors, the λ-value that achieves the maximum C-index and the λ-value that deviates by one standard error (1se criterion) from the maximum C-index to obtain a sparser solution. Due to the high computational complexity of the *model-X* knockoff method (see Figure 3.1), it is practically infeasible in this large-scale

high-dimensional setting. Therefore, we cannot consider it in this survival analysis.

Figure 6.5 shows the number of selected genes for different target FDR levels (in %) and different values of γ. First, we observe that the ordinary *T-Rex* method selected more genes than the *T-Rex+DA+BT* selector. In accordance with Corollary 1, all genes that were selected by the *T-Rex+DA+BT* selector were also selected by the more liberal ordinary *T-Rex* selector. In contrast, the regularized Cox methods did not provide consistent results for many values of γ because many genes that were selected by the more conservative 1se criterion do not appear in the selected set of the more liberal maximum C-index criterion. Moreover, it seems that many choices of γ lead to a very high number of selected genes, which raises some suspicion with respect to reproducibility because only 149 non-censored data points are usually not sufficient to reliably detect thousands of genes. By contrast, all three genes that were selected by the proposed method at a target FDR level of 20% (i.e., 'ITM2A', 'SCGB2A1', 'RYR2') have been previously identified to be related to breast cancer [Zho+19; Lac06; Xu+21].

Opportunity is missed by most people because it is dressed in overalls and looks like work.

Thomas Edison

7

Applications in Financial Engineering

In this chapter, the proposed FDR-controlling *T-Rex* methods from Chapters 3, 4, and 5 are used to solve two real-world data problems in financial engineering. In Section 7.1, a sparse, FDR-controlled, and quarterly updated portfolio is constructed and used to accurately track the S&P 500 index. In Section 7.2, a factor analysis of the most influential stocks in the S&P 500 index is conducted and used to reveal the interdependencies among stocks based on the idiosyncratic component.

The main content of this chapter is based on the publications [MPMew], [MMP24], and [Mac+24a]. The results have been produced using the developed open source R software packages TRexSelector [Mac+24c] and tlars [Mac+24b].

7.1 FDR-Controlled Sparse Index Tracking

In the following, the sparse index tracking problem is described and the stock returns data model is introduced. Then, the proposed FDR-controlling index tracking algorithm is formulated. Finally, the real-world S&P 500 index tracking setups and results are presented and discussed.

7.1.1 Problem Statement and Stock Returns Data Model

Financial index tracking is a fundamental task for the design of asset portfolios that are used to create exchange traded funds (ETFs) and hedging strategies of mutual funds [Pal24; BFP17; Pri07]. Prevalent index tracking approaches replicate an entire index (e.g., SPDR S&P 500 ETF) by creating and regularly updating a full tracking portfolio. However, this leads to high transaction costs because it requires the regular purchase and disposition of all assets in an index. Therefore, sparse index tracking methods, which use a small fraction of the stocks that constitute an index, have been proposed [JV02; MO07; Sco+13; XLX16; BFP17]. The common disadvantage of existing sparse approaches is that they require the investor to choose the size of the tracking portfolio or the value of a sparsity tuning parameter. Since there exist no optimal strategies for the choice of these parameters, the authors of the aforementioned approaches resort to experimental choices or rules-of-thumb that often lead to sub-optimal tracking portfolios.

In this work, we use the proposed FDR-controlling *T-Rex* methods to automatically determine the size and composition of a sparse index tracking portfolio. In the index tracking context, the FDR is the expected fraction of irrelevant stocks (i.e., stocks that are irrelevant for tracking an index) among all selected stocks. The target FDR between 0 and 100% expresses the level of the investor's willingness to sacrifice a small amount of transaction costs (arising from the inclusion of a few irrelevant stocks into the tracking portfolio) in order to obtain a diversified and yet small tracking portfolio.

As suggested in [BFP17; Pal24], we model the stocks returns as a linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{7.1}$$

where $\boldsymbol{w} = [w_1 \cdots w_p]^\top \in \mathbb{R}^p$ is the asset weight vector and $\boldsymbol{\epsilon} = [\epsilon_1 \cdots \epsilon_n]^\top \in \mathbb{R}^n$ is an additive noise vector. Here, $\boldsymbol{y} = [y_1 \cdots y_n]^\top \in \mathbb{R}^n$ is the daily index returns vector, i.e.,

$$y_i = \frac{\text{index}_i - \text{index}_{i-1}}{\text{index}_{i-1}}, \quad i = 1, \ldots, n, \tag{7.2}$$

where index_i is the closing price of the index on day i and index_0 is the closing price at the first day of the considered period. Analogously, $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_p] \in \mathbb{R}^{n \times p}$ is the matrix

containing the daily returns of the stocks $\boldsymbol{x}_j = [x_{1,j} \cdots x_{n,j}]^\top \in \mathbb{R}^n$, i.e.,

$$x_{i,j} = \frac{\text{price}_{i,j} - \text{price}_{i-1,j}}{\text{price}_{i-1,j}}, \quad i = 1, \ldots, n, \tag{7.3}$$

where $\text{price}_{i,j}$ is the closing price of the jth stock on day i and $\text{price}_{0,j}$ is the closing price on the first day of the period.

The general goal of sparse index tracking is to determine a sparse estimator of the asset weight vector $\boldsymbol{w}$ that tracks the index $\boldsymbol{y}$ sufficiently well using few relevant assets while obeying the following two rules [BFP17]:

1. Shorting stocks is not allowed, i.e., $w_j \geq 0$, $j = 1, \ldots, p$.
2. The available budget has to be invested, i.e., $\|\boldsymbol{w}\|_1 = 1$.

Note that the "no-shorting" constraint is technically a non-negativity constraint that can be sparsity inducing [Mei13]. Therefore, some additional stocks might be dropped in this second step.

7.1.2 Algorithm: FDR-Controlled Index Tracking

The proposed algorithm regularly updates the FDR-controlled tracking portfolio in a rolling-window fashion, as suggested in [BFP17], while satisfying the "no-shorting" and budget constraint in every training period $m \in \{1, \ldots, M\}$. Therefore, we first run the dependency-aware *T-Rex* selector with the target FDR $\alpha \in [0, 1]$, the stock returns matrices of the training periods $\boldsymbol{X}_m$, $m = 1, \ldots, M$, and the corresponding index returns vectors $\boldsymbol{y}_m$, $m = 1, \ldots, M$, as inputs, and with either

1. the nearest neighbors group design in (4.39) or
2. the autoregressive group design in (4.48).

Then, we solve the constrained quadratic problem

$$\underset{\boldsymbol{w}_m}{\text{minimize}} \, \|\boldsymbol{y}_m - \boldsymbol{X}_{m,\widehat{\mathcal{A}}^{(m)}} \cdot \boldsymbol{w}_m\|_2^2 + \lambda_2 \|\boldsymbol{w}_m\|_2^2 \quad \text{subject to} \quad \boldsymbol{w}_m \geq \boldsymbol{0} \tag{7.4}$$

$$\|\boldsymbol{w}_m\|_1 = 1. \tag{7.5}$$

Here, $\boldsymbol{X}_{m,\widehat{\mathcal{A}}^{(m)}}$ contains only the daily returns of the selected subset of stocks in $\widehat{\mathcal{A}}^{(m)}$ and

Algorithm 8 FDR-controlled index tracking.

1. **Input**: $\alpha \in [0, 1]$, $\boldsymbol{X}_m, \boldsymbol{y}_m, m = 1, \ldots, M$.
2. **For** $m = 1, \ldots, M$ **do**:
   - 2.1. **Compute** $\mathrm{Gr}(j, \rho_{\text{thr}})$, $j = 1, \ldots, p$, as defined
     - 2.1.1. for the *T-Rex+DA+NN* selector in (4.39) for all $\rho_{\text{thr}} \in \{0, 0.01, 0.02, \ldots, 1\}$ or
     - 2.1.2. for the *T-Rex+DA+AR*1 selector in (4.48).
   - 2.2. **Run** the *T-Rex+DA* selector in Algorithm 5 with $\mathrm{Gr}(j, \rho_{\text{thr}})$ from Step 2.1 to obtain an optimal solution $(v^*, \rho^*_{\text{thr}}, T^*, L)$ and the corresponding FDR-controlled set of selected stocks $\widehat{\mathcal{A}}^{(m)} := \widehat{\mathcal{A}}^{(m)}_L(v^*, \rho^*_{\text{thr}}, T^*)$.
   - 2.3. **Solve** the quadratic optimization problem in (7.5).
3. **Output**: Portfolio weight vectors for each quarter, i.e., $\widehat{\boldsymbol{w}}_1, \ldots, \widehat{\boldsymbol{w}}_M$.

the ℓ_2-penalty term $\lambda_2 \|\boldsymbol{w}_m\|_2^2$ with $\lambda_2 = 10^{-9}$ is only added to ensure a unique solution when the number of selected stocks exceeds the number of trading days in a training period. Note that for the autoregressive group design in (4.48), the ordering of the stocks matters. Therefore, the stocks in $\boldsymbol{X}_m$ are sorted using single-linkage hierarchical clustering [MC12] with pairwise sample correlations as a similarity measure.[1] This ensures that highly correlated stocks are placed next to each other before fitting an AR(1) model as described in Section 4.4.2. Figure 7.1 shows the returns of the sorted stocks on three different trading days. Algorithm 8 summarizes the proposed index tracking method.

7.1.3 Real-World S&P 500 Index Tracking

First, the proposed *T-Rex+DA+NN* selector from Section 4.4.1 is used to perform sparse FDR-controlled index tracking and its performance is compared against the benchmark methods. Then, the proposed *T-Rex+DA+AR*1 selector from Section 4.4.2 is used to track the index at a much lower target FDR level of 1%.

The *T-Rex+DA+NN* selector and the *T-Rex+DA+AR*1 selector both perform FDR-controlled stock selection for index tracking. The difference between them is that they use different dependency models to capture and leverage the dependency structure among the

[1] Other variants of hierarchical clustering, such as complete linkage and average linkage [MC12], have produced comparable results.

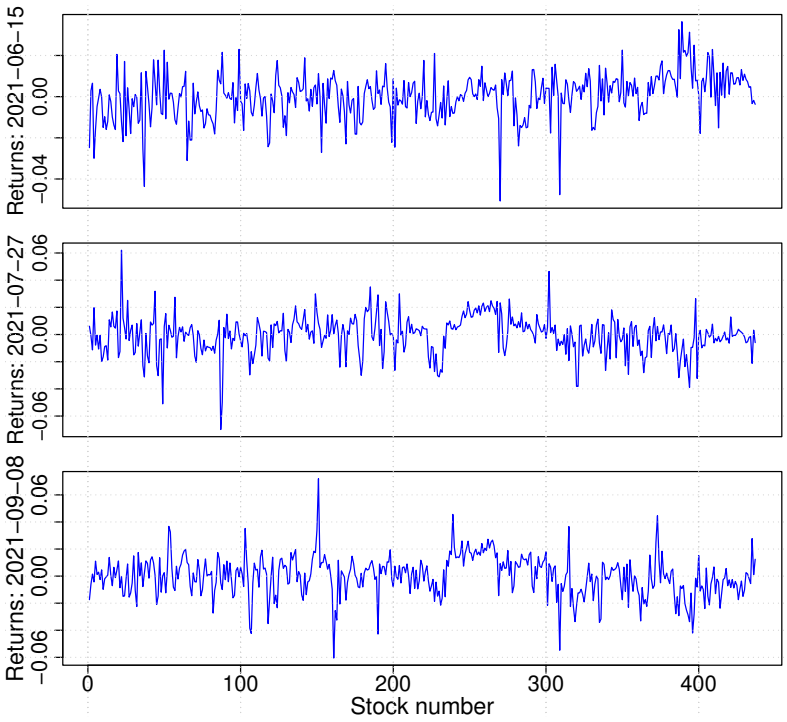


Figure 7.1: Exemplary daily returns of the sorted 438 stocks.

stocks. The *T-Rex+DA+NN* selector uses a nearest neighbors group correlation model, while the *T-Rex+DA+AR*1 selector uses a first order autoregressive stock dependency model. In the following sections, the index tracking performances of both methods are compared and discussed.

7.1.3.1 T-Rex+DA+NN Selector: Setup and Results

We consider the S&P 500 index in the 20 year period from $01/01/2003-29/09/2023$, which gives us $5{,}220$ trading days (i.e., samples) in total. These are divided into $M = 86$ training and 86 testing periods, where the first period can only be used for training and the last period only for testing. Using the same rolling window approach as in [BFP17], the portfolio is updated quarterly, i.e., all training and testing periods consist of $n = 60$ trading days. After removing all stocks that contain missing values, $p = 390$ candidate stocks are left. In summary, the response vector of the mth training period $\boldsymbol{y}_m \in \mathbb{R}^{60\times 1}$, $m = 1, \ldots, 86$, contains the daily index returns, and the corresponding predictor matrix $\boldsymbol{X}_m \in \mathbb{R}^{60\times 390}$, $m = 1, \ldots, 86$, contains the daily stock returns as columns.

As FDR-controlling benchmark methods, we consider the ordinary *T-Rex* selector from Chapter 3 and the *model-X* knockoff+ method [Can+18] for the variable selection step in Algorithm 8. We also consider the non-FDR-controlling state-of-the-art sparse index tracking method *ALAIT - ETE* [BFP17] that solves a *Lasso*-type optimization problem with the same no-shorting and budget constraints using a majorization minimization approach. The target FDR α is set to 30% and the sparsity tuning parameter for *ALAIT - ETE* to $\lambda = 10^{-7}$, as suggested in [BFP17; BP19].

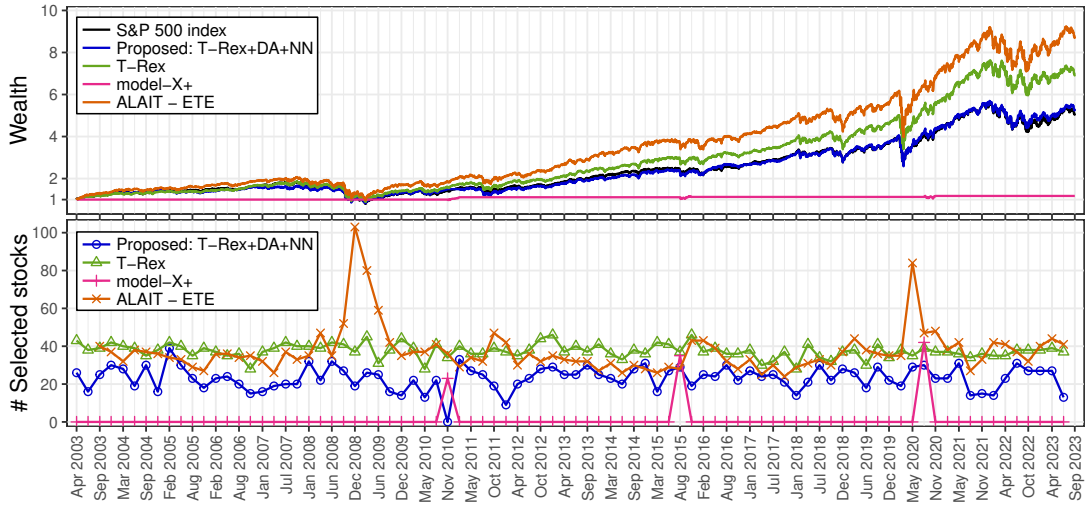


Figure 7.2: The proposed *T-Rex+DA+NN* selector closely follows the S&P 500 index using the fewest number of stocks in almost all quarters.

Table 7.1: The proposed *T-Rex+DA+NN* selector achieves a lower average (over all quarters) mean squared wealth tracking error (MSTE) using a smaller portfolio. Another major advantage is that the proposed method only requires choosing an interpretable target FDR level α (in %), while the benchmark method relies on the uninterpretable tuning parameter λ ($\times 10^{-7}$).

	ALAIT - ETE			*T-Rex+DA+NN*	
Param. λ	# stocks	MSTE	Param. α	# stocks	MSTE
16.67	12.72	0.96	20%	**9.62**	**0.91**
6.67	18.73	0.76	25%	**13.67**	**0.03**
3.33	24.23	1.32	30%	**19.72**	**0.22**

As suggested in [BFP17], the tracking performance is measured by the more comprehensible wealth. That is, we do not consider the absolute value of the index but set the value of the

index at the start of the tracking period to one and the wealth represents how an investors wealth has changed with respect to the reference. Note that index tracking does not aim at achieving the highest possible wealth but at an accurate tracking of the wealth corresponding to the index.

The results in Figure 7.2 and Table 7.1 show that the proposed dependency-aware *T-Rex* selector with the nearest neighbors penalization mechanism (*T-Rex+DA+NN*) has the best index tracking performance, while requiring the smallest number of stocks in almost all quarters.

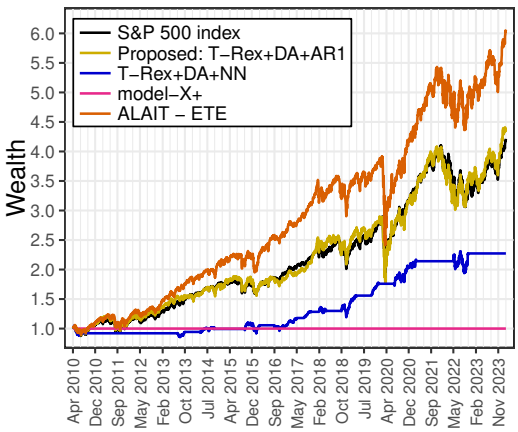
7.1.3.2 T-Rex+DA+AR1 Selector: Setup and Results

We consider the period $01/01/2010-30/01/2024$, which amounts to 3540 trading days (i.e., samples). These are divided into 58 training and 58 testing periods, where the first period can only be used for training and the last period can only be used for testing. As in Section 7.1.3.1, the portfolio is updated quarterly in a rolling window fashion and, thus, each period consists of 60 trading days. The response vector of the mth training period $\boldsymbol{y}_m \in \mathbb{R}^{60\times 1}$, $m = 1,\ldots,58$, contains the daily index returns. After removing all stocks that contain missing values, 438 candidate stocks are left. The predictor matrix $\boldsymbol{X}_m \in \mathbb{R}^{60\times 438}$, $m = 1,\ldots,58$, contains the daily returns of the stocks as column vectors.

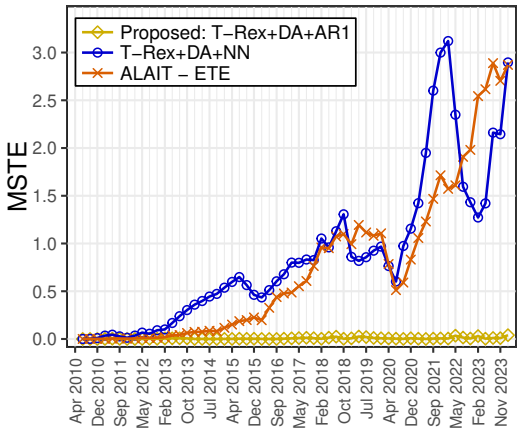
The results in Figure 7.3 and Table 7.2 show that the proposed *T-Rex+DA+AR*1 selector has the best index tracking performance and the lowest mean squared wealth tracking error (MSTE) while requiring only few stocks at a low target FDR of 1%. In comparison, the *T-Rex+DA+NN* selector performs well at higher target FDR levels (see Figure 7.2 and Table 7.1) but at very low target FDR levels, it becomes too conservative and, therefore, selects too few stocks, while the more liberal *T-Rex+DA+AR*1 selector selects sufficiently many stocks to accurately track the S&P 500 index. In conclusion, the *T-Rex+DA+AR*1 selector is favorable in scenarios where the investor can only tolerate a very small fraction of irrelevant stocks in the tracking portfolio.

7.2 Factor Analysis of S&P 500 Stock Returns

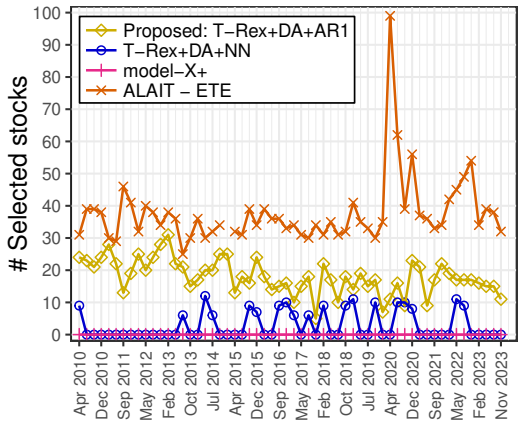
Understanding the interdependencies among stocks in an index such as the S&P 500 index is crucial for the analysis of portfolios. However, computing a simple sample correlation matrix does not allow to assess the fine interdependencies among stocks. The reason is that all stocks in the S&P 500 index are part of the same market and, therefore, are obscured by



(a) Wealth



(b) MSTE

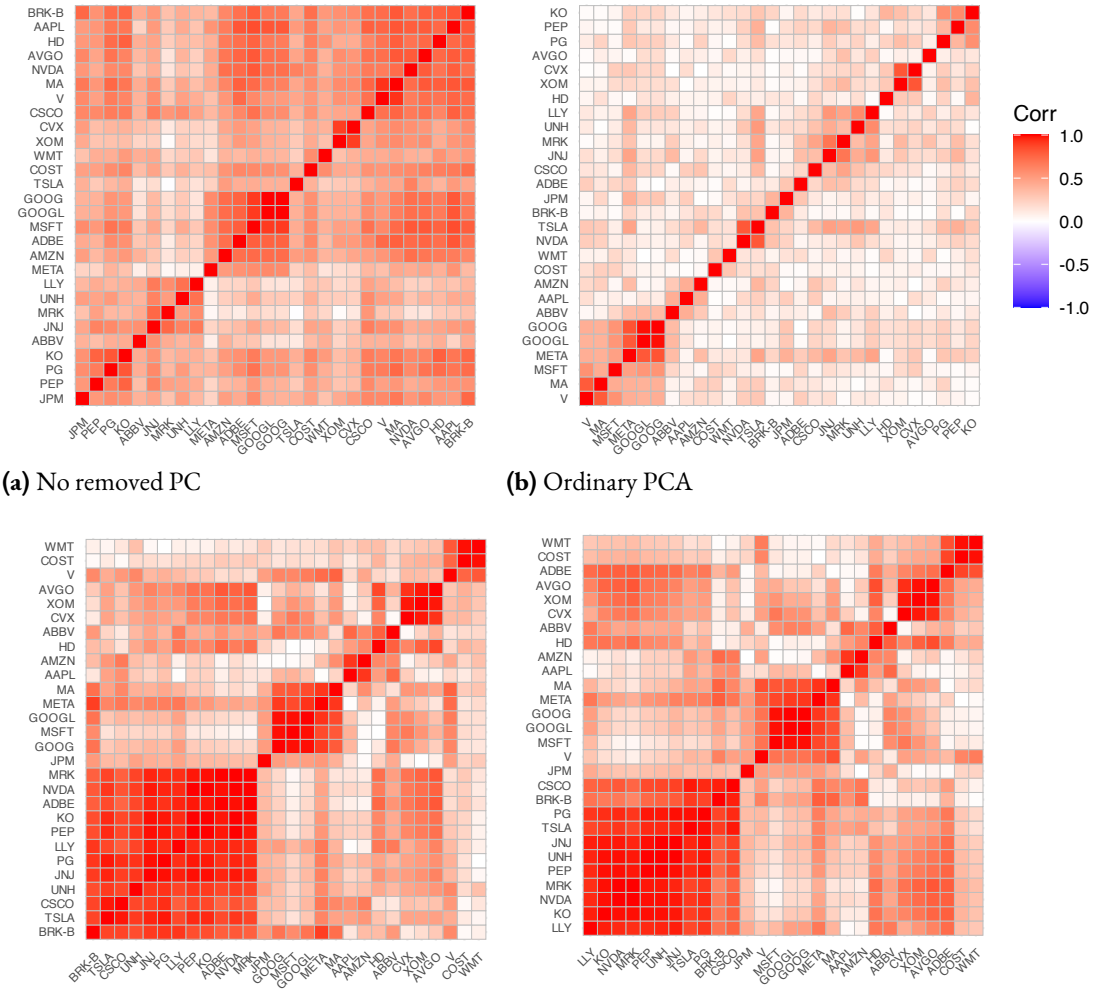


(c) Number of selected stocks

Figure 7.3: The proposed *T-Rex+DA+AR*1 selector shows (a) the best wealth tracking performance and (b) the lowest mean squared wealth tracking error (MSTE) in all quarters while (c) requiring only few stocks at a target FDR of 1% and $\rho_{\text{thr}} = 0.07$. The *model-X* knockoff+ method does not select any stocks and, therefore, does not track the index. *ALAIT-ETE* and *T-Rex+DA+NN* have high MSTEs in most quarters. The vertical grey lines indicate the days of the quarterly updates of the portfolio. For visual clarity, the high MSTE of the *model-X* knockoff+ method and the number of selected stocks by *ALAIT-ETE* in January, 2015 (i.e., 437), are omitted in Figures (b) and (c), respectively.

Table 7.2: MSTE and number of selected stocks averaged over all testing periods.

Methods	**Av. MSTE**	**Av. # selected stocks**
*T-Rex+DA+AR*1	**0.0067**	**17.98**
T-Rex+DA+NN	0.8774	2.88
model-X+	2.2806	0.00
ALAIT-ETE	0.7538	44.41



(a) No removed PC

(b) Ordinary PCA

(c) *T-Rex* PCA

(d) *T-Rex* Thresh. PCA

Figure 7.4: Correlation matrices of the 28 most influential stocks (according to their index weights) in the S&P 500 index.

strong statistical market factors [RM11; AL10]. Therefore, our goal in this application is to use the proposed *T-Rex* PCA and *T-Rex* Thresholded PCA from Section 5.3 to determine the strongest common factors, remove them from the data (which leaves us with the idiosyncratic component), and, thereby, reveal the fine interdependencies among the stocks. That is, we compute $\widehat{\boldsymbol{X}} = \widehat{\mathbf{Z}}'\widehat{\mathbf{V}}'^{\top}$, where $\widehat{\mathbf{Z}}'$ and $\widehat{\mathbf{V}}'$ are copies of the estimated PC matrix $\widehat{\mathbf{Z}}$ and the estimated loading matrix $\widehat{\mathbf{V}}$, respectively, except that the first three columns (i.e., the first three PCs) are removed, which is motivated by the Fama–French three-factor model for stock returns [FF92]. For this purpose, we consider the returns of the stocks that constitute the

S&P 500 index in the three month period from 01/10/2022 to 31/12/2022. The stock returns matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_p] \in \mathbb{R}^{n \times p}$ contains n daily returns of p stocks, as defined in Section 7.1.1.

Figure 7.4 presents the correlation matrices of the 28 most influential stocks (i.e., stocks with index weight larger than 0.6%) in the S&P 500 index. In order to visually distinguish groups of highly associated stocks, the correlation matrices are reordered using complete linkage hierarchical clustering. Even after reordering, the correlation matrix that corresponds to no removed PCs barely reveals any groups of stocks. The ordinary PCA removes too much variance and, therefore, does not allow to distinguish groups of highly correlated stocks. In contrast, after removing the first three PCs, the proposed methods (i.e., *T-Rex* PCA and *T-Rex* Thresholded PCA at a target FDR level of 10%) reveal that there exist meaningful groups of highly correlated stocks that are not explained by the three leading PCs but by the idiosyncratic component. Since the oracle SPCA, which has been considered as a benchmark method in the numerical experiments in Section 5.3.5, is infeasible in this real world example, it is omitted. The results indicate meaningful relationships among stocks from different industries. However, a detailed interpretation of the results from a portfolio design perspective goes beyond the scope of this dissertation.

It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment.

Carl Friedrich Gauss

The important thing is not to stop questioning. Curiosity has its own reason for existing.

Albert Einstein

8 Conclusion

This chapter concludes the dissertation with a summary of the contributions in Section 8.1 and an overview of limitations and open challenges for future research in Section 8.2.

8.1 Summary

The first main contribution of this dissertation (see Chapter 3) is the development and analysis of the *T-Rex* selector, a new fast FDR-controlling variable selection framework for high-dimensional data [MMPewa]. The *T-Rex* selector is, to the best of our knowledge, the first multivariate high-dimensional FDR-controlling method that scales to millions of variables in a reasonable amount of computation time. Since the *T-Rex* random experiments can be computed in parallel, multicore computers allow for additional substantial savings in computation time. These properties make the *T-Rex* selector a suitable method especially for large-scale GWAS. For cases where thousands of GWAS need to be conducted efficiently in order to screen through all the phenotypes of a large-scale biobank such as the UK biobank [Sud+15], the *Screen-T-Rex* selector has been developed [MMP23a]. It is a fast FDR-controlling variable selection method that does not ask the user to set a target FDR level, but provides the user with an estimate of the achieved FDR. In the few cases, where the user might not be satisfied with the provided FDR estimate, the original *T-Rex* selector can be used with the result of the Screen-T-Rex selector as a warm start and the desired target FDR as an input.

The second main contribution of this dissertation (see Chapter 4) is the development and analysis of the dependency-aware *T-Rex* (*T-Rex+DA*) selector [MMPewb]. The *T-Rex+DA* selector is an extension of the *T-Rex* selector that provides FDR control in situations where existing methods including the *T-Rex* selector break down. In particular, it performs high-dimensional FDR-controlled variable selection in the presence of strong dependencies at the cost of a reduced power compared to the *T-Rex* selector. It enables the solution of various problems such as high-dimensional survival analysis [MMPewb] and sparse financial index tracking [MPMew; MMP24], where strong dependencies among the variables (e.g., gene expression levels, stock returns) exist and common SNP pruning or other preprocessing techniques are not applicable.

The third main contribution of this dissertation (see Chapter 5) is the development and analysis of algorithms for FDR-controlled joint grouped variable selection and FDR control. The proposed *T-Rex+GVS* selector uses the *elastic net (EN)* as a base forward selector within the *T-Rex* framework and modifies the dummy generation procedure, which leads to a significant TPR increase in grouped variable selection tasks compared to the ordinary *T-Rex* selector and other benchmark methods [MMP22]. Moreover, the *informed elastic net (IEN)*, a fast grouped variable selection method for high-dimensional settings, has been proposed [MMP23b]. Replacing the *EN* base forward selector with the proposed *IEN* has lead to a significantly reduced computation time in large-scale high-dimensional settings and a better performance in a simulated GWAS compared to the *EN*-based *T-Rex+GVS* selector. Leveraging the formulation of sparse PCA as a series of grouped variable selection tasks, the developed *T-Rex+GVS* selector has been used to develop two new methods: *T-Rex* PCA and *T-Rex* Thresholded PCA [Mac+24a]. These methods perform the double duty of dimension reduction and variable selection while controlling the FDR of the sparse loading vectors. In contrast to existing methods, they do not require the tuning of any sparsity parameters and are capable of explaining the signal variance in the data with few PCs, which allows for meaningful interpretations of the obtained sparse PCs. The proposed FDR-controlled PCA methods have shown a promising performance in simulated data and have proven to be useful for revealing the interdependencies among stocks from the S&P 500 index.

The fourth main contribution of this dissertation (see Chapters 6 and 7) is the usage of the developed FDR-controlling methods to solve challenging real-world data problems in biomedical and financial engineering. The biomedical engineering applications include a high-dimensional GWAS, an HIV-1 drug resistance association study, and a sparse breast

cancer survival analysis study. The financial engineering applications include the design of a quarterly updated and rebalanced FDR-controlled tracking portfolio that allows for accurately tracking the S&P 500 index over decades and a factor analysis of S&P 500 stock returns. In all of these applications, the developed *T-Rex* methods have shown a remarkably accurate variable selection and FDR control performance compared to the state-of-the-art benchmark methods. Moreover, for large-scale data applications like the GWAS, the computation time has been multiple orders of magnitude lower compared to the benchmark methods.

The fifth main contribution of this dissertation is the development of the actively maintained open-source R software packages *tlars* [Mac+24b] and *TRexSelector* [Mac+24c] that have been published on CRAN. These software packages ensure the reproducibility of the presented results and facilitate the resource-friendly usability of the proposed *T-Rex* methods through an efficient C++ backend. Much work has gone into the creation, testing, improvement, and maintenance of these software packages. While none of this is documented in the dissertation, a testimony to the impact of this work is the already high number of downloads. Figure 8.1 shows the accumulated monthly downloads until the date of submission.

Beyond the presented contributions, the proposed *T-Rex* framework and its extensions have

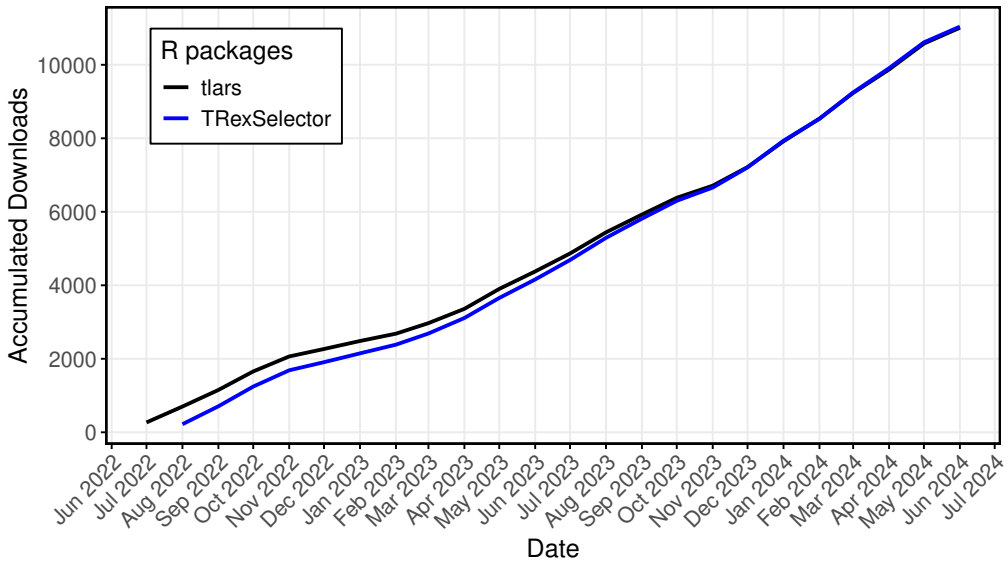


Figure 8.1: The plot shows the accumulated monthly downloads of the developed open source R software packages TRexSelector [Mac+24c] and tlars [Mac+24b]. The tlars package was released about one month before the TRexSelector package. Both packages have been downloaded more than 11,000 times since their initial release.

already proven to be useful and versatile tools for the efficient computation in big data applications [SMM23] and the estimation of Gaussian graphical models [KMM24].

8.2 Limitations and Open Challenges for Future Research

Since the development of the *T-Rex* selector in Chapter 3, several limitations have been addressed by extending the framework to handle various dependency structures among candidate variables in Chapter 4, grouped variable selection in Chapter 5, and Gaussian graphical models [KMM24]. However, beyond these advancements, many extensions are still possible, and some limitations remain. The following non-exhaustive list presents open challenges for future research, highlights limitations of the developed *T-Rex* methods, and proposes potential remedies:

1. **FDR-controlled GWAS catalog of reproducible discoveries**: The aim of this future research challenge is to conduct FDR-controlled multivariate reproducibility studies with high TPR using the *T-Rex* selector on large-scale genotype and phenotype data from the UK Biobank [Sud+15] in order to reproduce the reported results in the GWAS catalog [Bun+19]. The aim is to confirm past discoveries, discover new genetic associations, and flag potentially false reported genetic associations. The envisioned output of this project is a curated catalog of reproducible genetic associations that helps scientists to focus their efforts in revealing the causal mechanisms behind the genetic associations on the most promising and reproducible genetic associations.

2. **Power analysis of the T-Rex selector**: In this dissertation, the TPR (i.e., power) of the proposed *T-Rex* selector has been compared empirically to the benchmark methods. In this comparison, the *T-Rex* selector has shown a higher TPR in almost all numerical experiments. In order to confirm this empirical observation theoretically, a theoretical analysis of the TPR is required. In recent years, similar power analyses have been carried out for knockoff methods [WBC17; WJ20; Wei+23].

3. **Simultaneous TPR control and FDR minimization**: FDR control considers the maximization of the number of selected variables under the condition that a user-defined target FDR level is not exceeded. However, in applications such as emergency response person localization [Sch+24] missed detections might be more

severe than false detections, since a missed detection means failing to rescue a person in a crisis situation (e.g., fire, flood, etc.). That is, it is more important to achieve a high user-defined TPR while keeping the FDR as low as possible. Therefore, this challenge for future research considers solving the problem of minimizing the FDR under the condition that the TPR equals or exceeds the user-defined target level. This future challenge requires the development of a TPR estimator which is difficult because any TPR estimator will depend on the unobservable number of true active variables. Note that this challenge is related to the second challenge, since both are focused on the analysis and estimation of the TPR.

4. **Dummy-free T-Rex selector**: A bottleneck of the *T-Rex* selector has been its high random-access memory (RAM) usage to store the dummies. This problem has been alleviated by using sophisticated memory mapping technologies that allow for an efficient usage of the solid-state drive (SSD) to virtually extend the available RAM on a computer. This has made the *T-Rex* selector scalable to millions of variables on a standard laptop [SMM23]. Nevertheless, an entirely dummy-free *T-Rex* selector would allow for an even further scalability of the framework to potentially billions of variables and, thus, allow for FDR-controlled multivariate association testing using whole genome sequencing data which recently became available in the UK biobank [Sud+15]. The challenge of this future research project is to determine a vector of dummy importance measures for each forward selection step of all random experiments within the *T-Rex* selector and to track the changing distribution of the dummy importance vector along the forward selection paths. This would allow to determine valid dummy importance measures that can be compared against the observed variable importance measures of the candidate variables to terminate the random experiments after the optimal number of dummies, as determined by the *T-Rex* calibration algorithm (see Sections 3.5.3 and 3.5.4), has entered the solution paths.

5. **Complex T-Rex selector**: In many applications that require reliable variable selection tools, the data is complex valued. Some existing complex valued variable selection applications that would greatly benefit from FDR control can be found in many different areas such as direction-of-arrival (DOA) estimation [SPP18], radar-based localization [Sch+24], and dynamic mode decomposition [Gra+20]. Thus, the goal of this work is to extend the proposed FDR-controlling *T-Rex* selector to allow for complex-valued input data. The first challenge of this project lies in effectively leveraging the

existing complex *LARS* [Gra+20] to adapt the base selector of the *T-Rex* framework, i.e., the proposed early-terminating *T-LARS* algorithm (see Section 3.3), to the complex number domain. Then, the dummy generation and the fusion of the candidate sets need to be adapted to the complex number domain in order to obtain the complex *T-Rex* selector.

6. **FDR-controlled estimation of graphs**: Graphical models are widespread and useful tools for modeling the conditional statistical relationships among variables. An exemplary application with high-dimensional data is given by the investigation of the genetic architecture of the human plasma metabolome [Sur+22]. Other important applications exist in biology and healthcare, climate science, and psychology [Bes+21; Iqb+16; Zer+14; Nor+21; Bhu+19]. The challenge is to estimate the often sparse underlying structures of such graphs while keeping the number of false edges low. This allows to not only use the estimated graphical model in follow-up tasks but also enhances the interpretability of the estimated sparse graphical model. A first successful attempt in this direction uses the *Screen-T-Rex* selector to provably control the FDR of the selected edges at a self-estimated level [KMM24]. However, the challenge of adapting the FDR estimator of the ordinary *T-Rex* selector to the graph estimation task still persists. Solving this problem will enable the user to control the FDR at the desired level and not only at the self-estimated level of the *Screen-T-Rex* selector, which might be too high in some low-SNR cases.

7. **FDR control as a tool to meet the European Union (EU) ethics guidelines for trustworthy artificial intelligence (AI)**: The EU has developed and published a set of requirements that any type of AI needs to meet in order to be considered trustworthy [CDT19]. One of these key requirements is that an AI system needs to be technically robust, which is further specified as being reproducible, accurate, and reliable. These are key principles behind the design of the proposed *T-Rex* framework. It reliably (i.e., provably) controls the FDR in variable selection tasks to obtain an accurate set of reproducible discoveries from large-scale high-dimensional data. Thanks to the flexibility of the *T-Rex* selector, any type of heuristic or black box AI forward selection algorithm can be incorporated into the *T-Rex* framework and will automatically be equipped with the FDR control property, which helps satisfying the EU ethics guidelines for trustworthy AI. Thus, the goal of this project is to investigate whether and which existing AI variable selection algorithms can be modified into forward selec-

tion algorithms and to incorporate these into the *T-Rex* framework to allow for reproducible discoveries that meet the EU guidelines for trustworthy AI.

A

Appendix to Chapter 3

This appendix to Chapter 3 is organized as follows: Appendix A.1 presents some technical corollaries and lemmas, and the detailed proofs of Lemma 1, Theorems 4, 5, and Corollary 3. Appendix A.2 provides an intuitive explanation of the deflated relative occurrence from Definition 9. Appendix A.3 discusses the hyperparameter choices for the extended calibration algorithm in Algorithm 3. In Appendix A.4, the computational complexity of the *T-Rex* selector is derived. Appendices A.5 and A.6, respectively, discuss and numerically verify the assumptions used by the state-of-the-art benchmark methods and the proposed approach. Appendix A.7 verifies the assumptions on HAPGEN2 genomics data. In Appendix A.8, additional simulation results for a low-dimensional setting are presented and discussed. Appendix A.9 illustrates Theorem 4. Appendix A.10 discusses the robustness of the *T-Rex* selector in the presence of non-Gaussian noise.

A.1 Proofs

In this appendix, we introduce and prove some technical corollaries and lemmas. Then, the detailed proofs of Lemma 1, Theorem 4 (Dummy generation), Corollary 3, and Theorem 5 (Optimality of Algorithm 2) are presented. The results follow from standard assumptions in FDR control theory (for details and numerical verifications, see Appendices A.5, A.6, and A.7). Table A.1 provides an overview of frequently used expressions.

Table A.1: Overview of frequently used expressions.

Expression	Meaning
$K \in \mathbb{N}_+ \backslash \{1\}$	Number of random experiments.
$L \in \mathbb{N}_+$	Number of dummies.
$T \in \{1, \ldots, L\}$	Number of included dummies after which the forward variable selection process in each random experiment is terminated.
T^*	Optimal value of T as determined by the calibration process.
$v \in [0.5, 1)$	Voting level.
v^*	Optimal value of v as determined by the calibration process.
$\alpha \in [0, 1]$	Target FDR level.
$\mathcal{Z} \coloneqq \{\text{null } j : j \in \{1, \ldots, p\}\}$	Index set of null variables.
$\mathcal{A} \coloneqq \{\text{active } j : j \in \{1, \ldots, p\}\}$	Index set of active variables.
$p_0 \coloneqq \lvert\mathcal{Z}\rvert$	Number of null variables.
$p_1 \coloneqq \lvert\mathcal{A}\rvert$	Number of (true) active variables.
$p = p_0 + p_1$	Total number of variables.
n	Number of data points.
$\widehat{\mathcal{A}}(v) \coloneqq \widehat{\mathcal{A}}_L(v, T)$	Estimator of the active set, i.e., index set of the selected variables.
$\widehat{\mathcal{A}}^0(v) \coloneqq \{\text{null } j : \Phi_{T,L}(j) > v\}$	Index set of the selected null variables.
$\widehat{\mathcal{A}}^1(v) \coloneqq \{\text{active } j : \Phi_{T,L}(j) > v\}$	Index set of the selected active variables.
$\mathcal{C}_{k,L}(T)$	Candidate set of the kth random experiment, i.e., index set of the included variables in the kth random experiment.

A.1.1 Preliminaries: Technical Corollaries and Lemmas

Corollary 3. *Let $\mathcal{Z}_{m,k}$ and $\mathcal{D}_{m,k}$ be the index sets of the non-included null and dummy variables in the mth LARS[1] forward selection step of the kth random experiment, respectively.*

[1]Note that Corollary 3 and subsequent results apply to all forward selection methods that select one (and do not drop any) variable in each forward selection step based on the maximum absolute sample correlations between the predictors and the response or the current residual. Thus, the results hold for the *LARS* algo-

Then, for all $j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}$, the probability of including X_j in the mth step of the kth random experiment (RE) is equal, i.e., for all $j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}$ it holds that

$$\mathbb{P}(\text{“}X_j \textit{ included in } m\textit{th step of } k\textit{th RE”} \mid j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}) = \frac{1}{|\mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}|}. \tag{A.1}$$

Proof. For ease of readability, the proof is deferred to Appendix A.1.4. □

Corollary 4. *The numbers of included null variables at step t of all random experiments are i.i.d. random variables following the negative hypergeometric distribution, i.e., as $n \to \infty$,*

$$\sum_{j \in \mathcal{Z}} \mathbb{1}_k(j, t, L) \sim \text{NHG}(p_0 + L, p_0, t), \tag{A.2}$$

$t = 1, \ldots, T$, $k = 1, \ldots, K$, where $\mathcal{Z}$ is the index set of the null variables.

Proof. Let t be the number of included dummies after which a random experiment is terminated. There exists a *LARS* step m at which t dummies are included. From Corollary 3, we know that the probability of including a null variable and the probability of including a dummy variable are equal in each step of any random experiment. Therefore, it follows from Corollary 3 that the number of included null variables in any random experiment can be described by a process that randomly picks null and dummy variables one at a time, without replacement, and with equal probability from $\mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}$ until the process is terminated after t dummies are included. Since the included active variables in that process do not count towards the number of included null variables, the total number of variables in the process is p_0 instead of p. The described process exactly follows the definition of the negative hypergeometric distribution, i.e., $\text{NHG}(p_0 + L, p_0, t)$ with $p_0 + L$ total elements, p_0 success elements, and t failures after which a random experiment is terminated. □

As a consequence of A-I and A-II (see Appendix A.5), the number of selected null variables (i.e., $V_{T,L}(v)$) conditioned on the number of null variables exceeding the minimum voting level of 50% (i.e., $V_{T,L}(0.5)$) is binomially distributed with $\mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)$ being the selection probability of variable $j_0 \in \widehat{\mathcal{A}}^0(0.5)$. Thus, we obtain the following hierarchical model:

rithm [Efr+04] and approximately hold for the *Lasso* [Tib96], adaptive *Lasso* [Zou06], *elastic net* [ZH05], and many other related methods.

Corollary 5. *The number of selected null variables $V_{T,L}(v)$ follows the hierarchical model*

$$V_{T,L}(v) \mid V_{T,L}(0.5) \tag{A.3}$$

$$\sim \text{Binomial}\Big(V_{T,L}(0.5),\, \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)\Big), \tag{A.4}$$

$$V_{T,L}(0.5) \overset{d}{\leq} \text{NHG}(p_0 + L, p_0, T), \tag{A.5}$$

where $\mathbb{P}\big(\Phi_{T,L}(j_0) > v\big) > 0$ for all $j_0 \in \widehat{\mathcal{A}}^0(0.5)$ and for any $v \in [0.5, 1)$.

Lemma 2. *Let v be any real number in $[0.5, 1)$ and $K \to \infty$. Then, for any $j_0 \in \widehat{\mathcal{A}}^0(0.5)$, the following equation is satisfied:*

$$\mathbb{E}\big[V_{T,L}(v)\big] = \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big) \cdot \mathbb{E}\big[V_{T,L}(0.5)\big]. \tag{A.6}$$

Proof. Using the tower property of the expectation, we can rewrite the expectation of $V_{T,L}(v)$ as follows:

$$\mathbb{E}\big[V_{T,L}(v)\big] = \mathbb{E}\Big[\mathbb{E}\big[V_{T,L}(v) \mid V_{T,L}(0.5)\big]\Big] \tag{A.7}$$

$$= \mathbb{E}\big[V_{T,L}(0.5) \cdot \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)\big] \tag{A.8}$$

$$= \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big) \cdot \mathbb{E}\big[V_{T,L}(0.5)\big]. \tag{A.9}$$

The second equation follows from

$$V_{T,L}(v) \mid V_{T,L}(0.5) \tag{A.10}$$

$$\sim \text{Binomial}\Big(V_{T,L}(0.5),\, \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)\Big) \tag{A.11}$$

in Corollary 5 and the third equation holds because $\Phi_{T,L}(j_0)$, $j_0 \in \widehat{\mathcal{A}}^0(0.5)$, are i.i.d. random variables and, therefore, the selection probability $\mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)$ for any fixed v is the same constant for all j_0. □

Lemma 3. *Let v be any real number in $[0.5, 1)$ and $K \to \infty$. Define*

$$\widehat{V}'_{T,L}(v) \coloneqq \widehat{V}_{T,L}(v) - \sum_{j \in \widehat{\mathcal{A}}(v)} \big(1 - \Phi_{T,L}(j)\big). \tag{A.12}$$

Then, for any $j_0 \in \widehat{\mathcal{A}}^0(0.5)$, the following equation is satisfied:

$$\mathbb{E}\big[\widehat{V}'_{T,L}(v)\big] = \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big) \cdot \widehat{V}'_{T,L}(0.5). \tag{A.13}$$

Proof. Taking the expectation of $\widehat{V}'_{T,L}(v)$ yields

$$\mathbb{E}\big[\widehat{V}'_{T,L}(v)\big] \tag{A.14}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi_{t,L}(q)}{L-(t-1)} \cdot \frac{\sum_{j\in\widehat{\mathcal{A}}(v)} \Delta\Phi_{t,L}(j)}{\sum_{q\in\widehat{\mathcal{A}}(0.5)} \Delta\Phi_{t,L}(q)}\right] \tag{A.15}$$

$$= \sum_{t=1}^{T} \frac{p_0 - \sum_{q\in\mathcal{Z}} \Phi_{t,L}(q)}{L-(t-1)} \tag{A.16}$$

$$\cdot\, \mathbb{E}\left[\frac{\sum_{j\in\widehat{\mathcal{A}}^0(v)} \Delta\Phi_{t,L}(j)}{\sum_{q\in\widehat{\mathcal{A}}^0(0.5)} \Delta\Phi_{t,L}(q)}\right], \tag{A.17}$$

where the first and the second equation follow from Definitions 9, 10, and A-III (see Appendix A.5), respectively. Note that $\sum_{q\in\mathcal{Z}} \Phi_{t,L}(q) = \frac{1}{K}\sum_{k=1}^{K}\sum_{q\in\mathcal{Z}} \mathbb{1}_k(q,t,L)$ is the average number of included null variables when stopping after t dummies have been included. Since $K \to \infty$, the law of large numbers allows replacing the average by its expectation. That is, $\sum_{q\in\mathcal{Z}} \Phi_{t,L}(q) = \mathbb{E}\big[\sum_{q\in\mathcal{Z}} \mathbb{1}_k(q,t,L)\big]$. Therefore, $\sum_{q\in\mathcal{Z}} \Phi_{t,L}(q)$ is deterministic and can be written outside the expectation.

Using the tower property, we can rewrite the expectation in (A.17) as follows:

$$\mathbb{E}\left[\frac{\sum_{j\in\widehat{\mathcal{A}}^0(v)} \Delta\Phi_{t,L}(j)}{\sum_{q\in\widehat{\mathcal{A}}^0(0.5)} \Delta\Phi_{t,L}(q)}\right] \tag{A.18}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{j\in\widehat{\mathcal{A}}^0(v)} \Delta\Phi_{t,L}(j)}{\sum_{q\in\widehat{\mathcal{A}}^0(0.5)} \Delta\Phi_{t,L}(q)} \;\middle|\; \big|\widehat{\mathcal{A}}^0(v)\big|, \big|\widehat{\mathcal{A}}^0(0)\big|\right]\right] \tag{A.19}$$

$$= \mathbb{E}\left[\frac{\big|\widehat{\mathcal{A}}^0(v)\big|}{\big|\widehat{\mathcal{A}}^0(0.5)\big|}\right] \tag{A.20}$$

The last equation follows from $\Delta\Phi_{t,L}(j_0)$, $j_0 \in \widehat{\mathcal{A}}^0(0.5)$, being i.i.d. random variables and the well known fact that $\mathbb{E}[Q_M \,/\, Q_N] = M \,/\, N$, where $Q_B = \sum_{b=1}^{B} Z_b$ with $Z_1, \ldots, Z_B$, $B \in \{M, N\}$, being non-zero i.i.d. random variables and $M \leq N$.

Noting that $|\widehat{\mathcal{A}}^0(v)| = V_{T,L}(v)$ and applying the tower property again, we can rewrite the expectation in (A.20) as follows:

$$\mathbb{E}\left[\frac{|\widehat{\mathcal{A}}^0(v)|}{|\widehat{\mathcal{A}}^0(0.5)|}\right] \tag{A.21}$$

$$= \mathbb{E}\left[\frac{V_{T,L}(v)}{V_{T,L}(0.5)}\right] \tag{A.22}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{V_{T,L}(v)}{V_{T,L}(0.5)} \;\middle|\; V_{T,L}(0.5)\right]\right] \tag{A.23}$$

$$= \mathbb{E}\left[\frac{1}{V_{T,L}(0.5)} \cdot \mathbb{E}\big[V_{T,L}(v) \mid V_{T,L}(0.5)\big]\right] \tag{A.24}$$

$$= \mathbb{E}\left[\frac{1}{V_{T,L}(0.5)} \cdot V_{T,L}(0.5) \cdot \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)\right] \tag{A.25}$$

$$= \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big). \tag{A.26}$$

The last three equations follow from the same arguments as in the proof of Lemma 2. Thus,

$$\mathbb{E}\big[\widehat{V}'_{T,L}(v)\big] = \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big) \cdot \sum_{t=1}^{T} \frac{p_0 - \sum_{q\in\mathcal{Z}} \Phi_{t,L}(q)}{L-(t-1)} \tag{A.27}$$

$$= \mathbb{P}\big(\Phi_{T,L}(j_0) > v\big) \cdot \widehat{V}'_{T,L}(0.5). \tag{A.28}$$

□

Lemma 4. *Let $K \to \infty$. Then,*

$$\mathbb{E}\left[\sum_{q\in\mathcal{Z}} \Phi_{t,L}(q)\right] = \frac{t}{L+1} \cdot p_0. \tag{A.29}$$

Proof. Using Definition 5, we obtain

$$\sum_{q\in\mathcal{Z}} \Phi_{t,L}(q) = \frac{1}{K} \sum_{k=1}^{K} \sum_{q\in\mathcal{Z}} \mathbb{1}_k(q,t,L). \tag{A.30}$$

Then, taking the expectation and noting that

$$\sum_{q\in\mathcal{Z}} \mathbb{1}_k(q,t,L) \sim \text{NHG}(p_0+L, p_0, t), \quad k=1,\dots,K, \tag{A.31}$$

i.e., the number of included null variables in the K random experiments are i.i.d. random variables following the negative hypergeometric distribution as stated in Corollary 4, yields

$$\begin{aligned}
\mathbb{E}\left[\sum_{q\in\mathcal{Z}} \Phi_{t,L}(q)\right] &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\sum_{q\in\mathcal{Z}} \mathbb{1}_k(q,t,L)\right] && \text{(A.32)}\\
&= \frac{1}{K}\cdot K \cdot \frac{t}{L+1}\cdot p_0 && \text{(A.33)}\\
&= \frac{t}{L+1}\cdot p_0. && \text{(A.34)}
\end{aligned}$$

□

Lemma 5. *Let v be any real number in $[0.5, 1)$. Define*

$$\epsilon^*_{T,L}(v) \coloneqq \inf\{\epsilon \in (0,v) : R_{T,L}(v-\epsilon) - R_{T,L}(v) = 1\} \tag{A.35}$$

with the convention that $\epsilon^*_{T,L}(v) = 0$ if the infimum does not exist. Suppose that $V_{T,L}\big(v - \epsilon^*_{T,L}(v)\big) = V_{T,L}(v)+1$, $\mathbb{E}\big[V_{T,L}(v)\big] > 0$, and $\mathbb{E}\big[\widehat{V}'_{T,L}(v)\big] > 0$. Then, for all $j_0 \in \widehat{\mathcal{A}}^0(0.5)$ it holds that

$$\begin{aligned}
&\textit{(i)} \quad \mathbb{E}\big[V_{T,L}\big(v-\epsilon^*_{T,L}(v)\big) \;\big|\; V_{T,L}(v)\big] && \text{(A.36)}\\
&\qquad = V_{T,L}(v)\cdot \frac{\mathbb{P}\big(\Phi_{T,L}(j_0) > v - \epsilon^*_{T,L}(v)\big)}{\mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)} && \text{(A.37)}
\end{aligned}$$

and

$$\begin{aligned}
&\textit{(ii)} \quad \mathbb{E}\big[\widehat{V}'_{T,L}\big(v-\epsilon^*_{T,L}(v)\big) \;\big|\; \widehat{V}'_{T,L}(v)\big] && \text{(A.38)}\\
&\qquad = \widehat{V}'_{T,L}(v)\cdot \frac{\mathbb{P}\big(\Phi_{T,L}(j_0) > v - \epsilon^*_{T,L}(v)\big)}{\mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)}. && \text{(A.39)}
\end{aligned}$$

Proof. (i) Let $\delta \geq 1$ be a constant that satisfies the equation $V_{T,L}\big(v-\epsilon^*_{T,L}(v)\big) = \delta \cdot V_{T,L}(v)$.

Then,

$$\mathbb{E}\big[V_{T,L}\big(v-\epsilon^*_{T,L}(v)\big)\;\big|\;V_{T,L}(v)\big] = \mathbb{E}\big[\delta\cdot V_{T,L}(v)\;\big|\;V_{T,L}(v)\big] \tag{A.40}$$

$$= \delta\cdot V_{T,L}(v). \tag{A.41}$$

We rewrite $\delta\cdot V_{T,L}(v)$ as follows:

$$\delta\cdot V_{T,L}(v) = V_{T,L}(v)\cdot\frac{\delta\cdot\mathbb{E}\big[V_{T,L}(v)\big]}{\mathbb{E}\big[V_{T,L}(v)\big]} \tag{A.42}$$

$$= V_{T,L}(v)\cdot\frac{\mathbb{E}\big[V_{T,L}(v-\epsilon^*_{T,L}(v))\big]}{\mathbb{E}\big[V_{T,L}(v)\big]} \tag{A.43}$$

$$= V_{T,L}(v)\cdot\frac{\mathbb{P}\big(\Phi_{T,L}(j_0)>v-\epsilon^*_{T,L}(v)\big)}{\mathbb{P}\big(\Phi_{T,L}(j_0)>v\big)}. \tag{A.44}$$

The last line follows from Lemma 2. Comparing $\delta\cdot V_{T,L}(v)$ and the last line, we see that

$$\delta = \mathbb{P}\big(\Phi_{T,L}(j_0)>v-\epsilon^*_{T,L}(v)\big)\,/\,\mathbb{P}\big(\Phi_{T,L}(j_0)>v\big) \tag{A.45}$$

and the first part of the lemma follows.

(ii) The proof is analogous to the proof of (i). The only difference is that Lemma 3 instead of Lemma 2 needs to be used for rewriting the expression $\delta\cdot\widehat{V}'_{T,L}(v)$. □

A.1.2 Proof of Lemma 1 (Martingale)

Proof. If there exists a variable with an index, say, j^* that is not selected at the voting level v but at the level $v-\epsilon^*_{T,L}(v)$ and it is a null variable, then we have

$$V_{T,L}(v-\epsilon^*_{T,L}(v)) = V_{T,L}(v)+1. \tag{A.46}$$

However, if j^* is an active variable or if the infimum in (A.35) does not exist, that is, no additional variable is selected at the voting level $v-\epsilon^*_{T,L}(v)$ when compared to the level v, then we obtain

$$V_{T,L}(v-\epsilon^*_{T,L}(v)) = V_{T,L}(v). \tag{A.47}$$

Summarizing both results, we have

$$V_{T,L}(v - \epsilon^*_{T,L}(v)) = \begin{cases} V_{T,L}(v) + 1, & j^* \in \mathcal{Z} \\ V_{T,L}(v), & j^* \in \mathcal{A} \\ & \text{or } \epsilon^*_{T,L}(v) = 0 \end{cases} . \tag{A.48}$$

Thus, using the definition of $H_{T,L}(v)$ within Lemma 1, we obtain

$$\mathbb{E}\big[H_{T,L}\big(v - \epsilon^*_{T,L}(v)\big) \mid \mathcal{F}_v\big] \tag{A.49}$$

$$= \mathbb{E}\left[\frac{V_{T,L}\big(v - \epsilon^*_{T,L}(v)\big)}{\widehat{V}'_{T,L}\big(v - \epsilon^*_{T,L}(v)\big)} \;\middle|\; V_{T,L}(v), \widehat{V}'_{T,L}(v)\right] \tag{A.50}$$

$$= \begin{cases} \mathbb{E}\left[\dfrac{V_{T,L}(v) + 1}{\widehat{V}'_{T,L}\big(v - \epsilon^*_{T,L}(v)\big)} \;\middle|\; V_{T,L}(v), \widehat{V}'_{T,L}(v)\right], & j^* \in \mathcal{Z} \\ \mathbb{E}\left[\dfrac{V_{T,L}(v)}{\widehat{V}'_{T,L}(v)} \;\middle|\; V_{T,L}(v), \widehat{V}'_{T,L}(v)\right], & j^* \in \mathcal{A} \\ & \text{or } \epsilon^*_{T,L}(v) = 0 \end{cases} \tag{A.51}$$

$$= \begin{cases} \dfrac{\mathbb{E}\left[\dfrac{1}{\widehat{V}'_{T,L}\big(v - \epsilon^*_{T,L}(v)\big)} \;\middle|\; V_{T,L}(v), \widehat{V}'_{T,L}(v)\right]}{\big(V_{T,L}(v) + 1\big)^{-1}}, & j^* \in \mathcal{Z} \\ \dfrac{V_{T,L}(v)}{\widehat{V}'_{T,L}(v)}, & j^* \in \mathcal{A} \\ & \text{or } \epsilon^*_{T,L}(v) = 0 \end{cases} . \tag{A.52}$$

Using Lemma 5, we can rewrite the denominator within the first case of Equation (A.52) as follows:

$$V_{T,L}(v) + 1 \tag{A.53}$$

$$= \mathbb{E}\big[V_{T,L}\big(v - \epsilon^*_{T,L}(v)\big) \mid V_{T,L}(v)\big] \tag{A.54}$$

$$= V_{T,L}(v) \cdot \frac{\mathbb{P}\big(\Phi_{T,L}(j_0) > v - \epsilon^*_{T,L}(v)\big)}{\mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)}. \tag{A.55}$$

Next, we rewrite the numerator within the first case of Equation (A.52) as follows:

$$\mathbb{E}\left[\frac{1}{\widehat{V}'_{T,L}\big(v - \epsilon^*_{T,L}(v)\big)} \;\middle|\; V_{T,L}(v), \widehat{V}'_{T,L}(v)\right] \tag{A.56}$$

$$\geq \frac{1}{\mathbb{E}\big[\widehat{V}'_{T,L}\big(v - \epsilon^*_{T,L}(v)\big) \mid V_{T,L}(v), \widehat{V}'_{T,L}(v)\big]} \tag{A.57}$$

$$= \frac{1}{\mathbb{E}\big[\widehat{V}'_{T,L}\big(v - \epsilon^*_{T,L}(v)\big) \mid \widehat{V}'_{T,L}(v)\big]} \tag{A.58}$$

$$= \left(\widehat{V}'_{T,L}(v) \cdot \frac{\mathbb{P}\big(\Phi_{T,L}(j_0) > v - \epsilon^*_{T,L}(v)\big)}{\mathbb{P}\big(\Phi_{T,L}(j_0) > v\big)}\right)^{-1} \tag{A.59}$$

The first inequality follows from Jensen's inequality. The first equation holds because $\widehat{V}'_{T,L}\big(v - \epsilon^*_{T,L}(v)\big)$ and $V_{T,L}(v)$ are conditionally independent given $\widehat{V}'_{T,L}(v)$ and the last line follows from Lemma 5. Plugging (A.55) and (A.59) into (A.52) yields

$$\mathbb{E}\big[H_{T,L}\big(v - \epsilon^*_{T,L}(v)\big) \mid \mathcal{F}_v\big] \geq H_{T,L}(v), \tag{A.60}$$

i.e., $\{H_{T,L}(v)\}_{v\in\mathcal{V}}$, with $\mathcal{V} = \{\Phi_{T,L}(j) : j = 1, \ldots, p\}$, is a backward-running supermartingale with respect to the filtration $\mathcal{F}_v$. □

A.1.3 Proof of Theorem 4 (Dummy generation)

Proof. Since

$$\mathbb{E}\big[D_{n,l,m,k}\big] = \frac{1}{\Gamma_{n,m,k}} \cdot \sum_{i=1}^{n} \gamma_{i,m,k} \cdot \mathbb{E}\big[\mathring{X}_{i,l,k}\big] = 0 \tag{A.61}$$

and

$$\text{Var}\left[D_{n,l,m,k}\right] = \frac{1}{\Gamma^2_{n,m,k}} \cdot \sum_{i=1}^{n} \gamma^2_{i,m,k} \cdot \text{Var}\left[\mathring{X}_{i,l,k}\right] = 1, \tag{A.62}$$

the Lindeberg-Feller central limit theorem can be used to prove that $D_{n,l,m,k} \xrightarrow{d} D$, $D \sim \mathcal{N}(0,1)$. In order to do this, we define

$$\mathring{Q}_{i,l,m,k} \coloneqq \frac{\gamma_{i,m,k} \cdot \mathring{X}_{i,l,k}}{\Gamma_{n,m,k}}, \tag{A.63}$$

and check whether it satisfies the Lindeberg condition, i.e., whether for every $\tau > 0$

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\Big[\mathring{Q}^2_{i,l,m,k} \cdot I\big(\big|\mathring{Q}_{i,l,m,k}\big| > \tau\big)\Big] = 0 \tag{A.64}$$

holds. Rewriting the Lindeberg condition using the definition of $\mathring{Q}_{i,l,m,k}$ yields

$$\lim_{n\to\infty} \sum_{i=1}^{n} \left(\frac{\gamma_{i,m,k}}{\Gamma_{n,m,k}}\right)^2 \mathbb{E}\left[\mathring{X}^2_{i,l,k} \cdot I\left(\big|\mathring{X}_{i,l,k}\big| > \frac{\tau\Gamma_{n,m,k}}{|\gamma_{i,m,k}|}\right)\right] = 0, \tag{A.65}$$

where $I(\cdot)$ denotes the indicator function, i.e., $I(A > B)$ is equal to one if $A > B$ and equal to zero if $A \leq B$. Since

$$\lim_{n\to\infty} \max_{1\leq i\leq n} \left(\frac{\gamma_{i,m,k}}{\Gamma_{n,m,k}}\right)^2 = 0 \tag{A.66}$$

and

$$\lim_{n\to\infty} \min_{1\leq i\leq n} \left(\frac{\Gamma_{n,m,k}}{|\gamma_{i,m,k}|}\right) \to \infty, \tag{A.67}$$

the Lindeberg condition is satisfied and the theorem follows. □

Remark 13. *Loosely speaking, Theorem 4 states that regardless of the distribution from which the dummies are sampled, the dummy correlation variables follow the standard normal distribution as $n \to \infty$. That is, the distribution of the dummies has no influence on the resulting distribution of the dummy correlation variables. Since the realizations of the dummy correlation variables determine which dummies are included along the LARS solution path, we can conclude that the decisions of which variable enters next along the solution path is independent of the distribution of the dummies. Thus, the dummies can be sampled from any univariate probability distribution with finite expectation and finite non-zero variance to serve as flagged null variables within the T-Rex selector.*

A.1.4 Proof of Corollary 3

Proof. Similarly to Theorem 4, we consider the predictors $\boldsymbol{x}_j = [x_{1,j}, \ldots, x_{n,j}]$ as n i.i.d. realizations of X_j, which can also be considered as one realization from each of the i.i.d. random variables $X_{1,j}, \ldots, X_{n,j}$. Replacing

(i) $\mathring{X}_{i,l,k}, i \in \{1, \ldots, n\}, l \in \mathcal{D}_{m,k}, k \in \{1, \ldots, K\}$,

in Theorem 4 with

(ii) $X_{i,j}, i \in \{1, \ldots, n\}, j \in \mathcal{Z}_{m,k}$,

and using A-I (see Appendix A.5), the conditions in Theorem 4 are satisfied. Thus, it follows that, as $n \to \infty$,

$$D_{n,j,m,k} \xrightarrow{d} D, \quad D \sim \mathcal{N}(0,1), \tag{A.68}$$

i.e., the null correlation variables $\{G_{j,m,k} : j \in \mathcal{Z}_{m,k}\}$ are identically distributed.[2] Since the non-included null random variables $\{X_j : j \in \mathcal{Z}_{m,k}\}$ are independent of the true active variables and mutually independent, the null correlation variables are also independently distributed. Thus, in combination with Theorem 4, the null and dummy correlation variables $\{G_{j,m,k} : j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}\}$ are i.i.d.

We define

$$g^*(j) \coloneqq \underset{g \in (\mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}) \setminus \{j\}}{\arg\max} \big\{|G_{g,m,k}|\big\}, \tag{A.69}$$

i.e., the largest absolute correlation with the current residual among all non-included nulls and dummies (except for variable j) in the mth *LARS* step. Since in each step m, the *LARS* algorithm includes the variable with the largest absolute correlation with the current residual, we have

$$\begin{aligned} &\mathbb{P}(\text{“}X_j \text{ included in } m\text{th step of } k\text{th RE”} \mid j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}) && \text{(A.70)} \\ &= \mathbb{P}\big(|G_{j,m,k}| \geq |G_{g^*(j),m,k}| \;\big|\; j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}\big). && \text{(A.71)} \end{aligned}$$

[2] Note that, as in Theorem 4, the constant factor $\Gamma_{n,m,k}$ in $D_{n,j,m,k} = (1/\Gamma_{n,m,k})G_{j,m,k}$ is equal for all $j \in \mathcal{Z}_{m,k}$ and does not affect the distribution of $G_{j,m,k}$.

Summing up the probabilities in (A.71) over all $j' \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}$ yields

$$1 = \sum_{j' \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}} \mathbb{P}\big(|G_{j',m,k}| \geq |G_{g^*(j'),m,k}| \;\big|\; j' \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}\big) \tag{A.72}$$

$$= |\mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}| \cdot \mathbb{P}\big(|G_{j,m,k}| \geq |G_{g^*(j),m,k}| \;\big|\; j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}\big) \tag{A.73}$$

for all $j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}$. The second line follows from the fact that the $G_{j,m,k}$'s are exchangeable because they are i.i.d. Exchangeability is meant in the sense that

$$\mathbb{P}\big(|G_{j_1,m,k}| \geq |G_{g^*(j_1),m,k}| \;\big|\; j_1 \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}\big) \tag{A.74}$$

$$= \mathbb{P}\big(|G_{j_2,m,k}| \geq |G_{g^*(j_2),m,k}| \;\big|\; j_2 \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}\big) \tag{A.75}$$

for all $j_1, j_2 \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}$.

Combining (A.71) and (A.73), we obtain

$$\mathbb{P}(\text{“}X_j \text{ included in } m\text{th step of } k\text{th RE”} \mid j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}) = \frac{1}{|\mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}|} \tag{A.76}$$

for all $j \in \mathcal{Z}_{m,k} \cup \mathcal{D}_{m,k}$. □

A.1.5 Proof of Theorem 5 (Optimality of Algorithm 2)

Proof. First, note that for all triples (v, T, L) that satisfy $\widehat{\text{FDP}}(v, T, L) \leq \alpha$, the objective functions in Step 4 of Algorithm 2 and in the optimization problem in (3.15) are equivalent, i.e., $\big|\widehat{\mathcal{A}}_L(v, T)\big| = R_{T,L}(v)$. Thus, in order to prove that (v^*, T^*) is an optimal solution of (3.15), it must be shown that the set of feasible tuples obtained by the algorithm contains the feasible set of (3.15). This also proves that (v^*, T^*) is a feasible solution of (3.13) and (3.14) because the conditions of the optimization problems in (3.13), (3.14), and (3.15) are equivalent.

Since, ceteris paribus, $\widehat{\text{FDP}}(v, T, L)$ is monotonically decreasing in v and monotonically increasing in T, the minimum of $\widehat{\text{FDP}}(v, T, L)$ is attained at $v = 1 - \Delta v$, $\Delta v = 1/K$, for any $T = T_{\text{fin}}$ that satisfies the inequalities $\widehat{\text{FDP}}(v = 1 - \Delta v, T = T_{\text{fin}}, L) \leq \alpha$ and $\widehat{\text{FDP}}(v = 1 - \Delta v, T = T_{\text{fin}} + 1, L) > \alpha$. All in all, and since $v = 1 - \Delta v$ asymptotically ($K \to \infty$) coincides with the supremum of the interval $[0.5, 1)$, the feasible set of (3.15) can

be rewritten as follows:

$$\left\{(v,T) : \widehat{\text{FDP}}(v,T,L) \leq \alpha\right\} \tag{A.77}$$

$$= \Big\{(v,T) : v \in [0.5, 1-\Delta v],\, T \in \{1,\ldots,T_{\text{fin}}\}, \tag{A.78}$$

$$\widehat{\text{FDP}}(v,T,L) \leq \alpha\Big\}. \tag{A.79}$$

Note that the v-grid in Algorithm 2 is adapted to the number of random experiments K and, therefore, all values of the objective function (i.e., $|\widehat{\mathcal{A}}_L(v,T)|$) that can be attained by off-grid solutions can also be attained by at least one on-grid solution. Therefore, we can replace the right side of Equation (A.79) by

$$\Big\{(v,T) : v \in \{0.5, 0.5+\Delta v, 0.5+2\cdot\Delta v, \ldots, 1-\Delta v\}, \tag{A.80}$$

$$T \in \{1,\ldots,T_{\text{fin}}\},\, \widehat{\text{FDP}}(v,T,L) \leq \alpha\Big\}. \tag{A.81}$$

The "while"-loop in Step 3 of Algorithm 2 is terminated when $T = T_{\text{fin}}+1$. Thus, the feasible set of the optimization problem in Step 4 of Algorithm 2 can be written as follows:

$$\Big\{(v,T) : v \in \{0.5, 0.5+\Delta v, 0.5+2\cdot\Delta v, \ldots, 1-\Delta v\}, \tag{A.82}$$

$$T \in \{1,\ldots,T_{\text{fin}}\},\, \widehat{\text{FDP}}(v,T,L) \leq \alpha\Big\}. \tag{A.83}$$

Since (A.81) is equal to (A.83), the theorem follows. □

A.2 The Deflated Relative Occurrence

In order to provide an intuitive explanation of the deflated relative occurrence, we rewrite the expression as follows:

$$\Phi'_{T,L}(j) = \sum_{t=1}^{T} \left(1 - \frac{\dfrac{1}{L-(t-1)}}{\dfrac{\sum\limits_{q \in \widehat{\mathcal{A}}(0.5)} \Delta\Phi_{t,L}(q)}{p - \sum\limits_{q=1}^{p} \Phi_{t,L}(q)}} \right) \Delta\Phi_{t,L}(j) \tag{A.84}$$

$$= \sum_{t=1}^{T} \left(1 - \underbrace{\frac{\overbrace{\dfrac{t-(t-1)}{L-(t-1)}}^{\text{(i)}}}{\dfrac{\frac{1}{K}\sum\limits_{k=1}^{K}\left(\sum\limits_{q \in \widehat{\mathcal{A}}(0.5)} \mathbb{1}_k(q,t,L) - \sum\limits_{q \in \widehat{\mathcal{A}}(0.5)} \mathbb{1}_k(q,t-1,L)\right)}{p - \frac{1}{K}\sum\limits_{k=1}^{K}\sum\limits_{q=1}^{p} \mathbb{1}_k(q,t,L)}}}_{\text{(ii)}} \right) \Delta\Phi_{t,L}(j). \tag{A.85}$$

The last equation follows by rewriting the expression in the denominator within the first expression using Definition 5. In the last expression, each element of the sum consists of $\Delta\Phi_{t,L}(j)$ multiplied with what we call the *deflation factor*. That factor is computed by subtracting from one the fraction of

(i) the number of included dummies at step t, which is always one, divided by the number of non-included dummies up until step $t-1$ and

(ii) the average number of included candidates at step t divided by the average number of non-included candidates up until step t.

That is, the larger (smaller) the fraction of included candidates at step t compared to the fraction of included dummies at step t, the more (less) weight is given to the change in relative occurrence in that step. Loosely speaking, if the number of non-included null variables and dummies is equal in step $t-1$ of the kth random experiment, then allowing one more dummy to enter the solution path leads, on average, to the inclusion of one more null variable. Thus, if going from step $t-1$ to t leads to the inclusion of many variables, then still only one null variable is expected to be among them and, therefore, the deflation factor for that step is close to one.

Remark 14. *The reader might wonder whether the deflation factors affect not only the inflated $\Delta\Phi_{t,L}(j)$'s of the null variables but also those of the active variables. In the following, we shall give an intuitive explanation of why the deflation factors have only a negligible effect on the $\Delta\Phi_{t,L}(j)$'s of the active variables: Since usually most active variables enter the solution paths early, i.e., at low values of t and because they are accompanied by very few null variables, the deflation factor is close to one. For this reason, the $\Delta\Phi_{t,L}(j)$'s of the active variables are relatively unaffected. With increasing values of t, the $\Delta\Phi_{t,L}(j)$'s of the active variables are close to zero, because for active variables the increases in relative occurrence are usually high for low values of t and, consequently, low (or even zero) at higher values of t. Summarizing, the deflation factors have little or no effect on the $\Delta\Phi_{t,L}(j)$'s of the active variables because for low values of t they are close to one and for large values of t the $\Delta\Phi_{t,L}(j)$'s of the active variables are close to zero or zero.*

A.3 Hyperparameter Choices for the Extended Calibration Algorithm

In this appendix, we discuss the choices of the reference voting level $\tilde{v}$ and the maximum values of L and T, namely $L_{\max}$ and $T_{\max}$ for the extended calibration algorithm in Algorithm 3:

1. $\tilde{v} = 0.75$: The choice of $\tilde{v}$ is a compromise between the 50% and 100% voting levels. Setting $\tilde{v} = 0.5$ would require low values of L to push $\widehat{\text{FDP}}(v = \tilde{v}, T, L)$ below the target FDR level while setting $\tilde{v} = 1$ would require very high values of L. Thus, $\tilde{v} = 0.75$ is a compromise between tight FDR control and memory consumption. Note that the FDR control property holds for any choice of $\tilde{v} \in [0.5, 1)$.

2. $L_{\max} = 10p$: In order to allow for sufficiently large values of L such that tight FDR control is possible while not running out of memory, setting $L_{\max} = 10p$ has proven

to be a practical choice. Note that the FDR control property in Theorem 3 holds for any choice of L. However, we can achieve tighter FDR control with larger values of L.

3. $T_{\text{max}} = \lceil n/2 \rceil$: As discussed for the *T-LARS* algorithm in the caption of Figure 3.5, the *LARS* algorithm includes at most $\min\{n, p\}$ variables and in high-dimensional settings ($p > n$), the maximum number of included variables in each random experiment is n. Since for $L = p$ we expect roughly as many null variables as dummies in very sparse settings, choosing $T_{\text{max}} = \lceil n/2 \rceil$ ensures that the *LARS* algorithm could potentially run until (almost) the end of the solution path. In contrast, for $L = 10p$ we expect 10 times as many dummies as null variables in very sparse settings. Thus, for $L = p$ we allow the solution paths to potentially run until the end, although this might only happen in rare cases, while for $L = 2p, \ldots, 10p$ we restrict the run length. This is a compromise between a higher computation time and a higher TPR (i.e., power) that are both associated with larger values of T_{max}.

A.4 Computational Complexity

The computational complexities of sampling dummies from the univariate standard normal distribution and fusing the candidate sets are negligible compared to the computational complexity of the utilized forward selection method. Therefore, it is sufficient to analyze the computational complexities of the early terminated forward selection processes. We restrict the following analysis to the *LARS* algorithm [Efr+04], which also applies to the *Lasso* [Tib96].[3] The κth step of the *LARS* algorithm has the complexity $\mathcal{O}\big((p-\kappa)\cdot n+\kappa^2\big)$, where the terms $(p-\kappa)\cdot n$ and κ^2 account for the complexity of determining the variable with the highest absolute correlation with the current residual (i.e., the next to be included variable) and the so-called equiangular direction vector, respectively. Replacing p by $p + L$, since the original predictor matrix is replaced by the enlarged predictor matrix, and summing up the complexities of all steps until termination yields the computational complexity of the *T-Rex* selector. First, we define the run lengths as the cardinalities of the respective candidate sets, i.e.,

$$\kappa_{T,L}(k) \coloneqq \left|\mathcal{C}_{k,L}(T)\right|, \quad k = 1, \ldots, K, \tag{A.86}$$

[3] Since the *Lasso* solution path can be computed by a slightly modified *LARS* algorithm, the *Lasso* and the *LARS* algorithm have the same computational complexity.

and assume $L \geq p$. Then, the sum over all steps until the termination of the kth random experiment is given by

$$\sum_{\kappa=1}^{\kappa_{T,L}(k)} \left((p + L - \kappa) \cdot n + \kappa^2\right) \tag{A.87}$$

$$= n \cdot \kappa_{T,L}(k) \cdot (p + L) - n \cdot \sum_{\kappa=1}^{\kappa_{T,L}(k)} \kappa + \sum_{\kappa=1}^{\kappa_{T,L}(k)} \kappa^2 \tag{A.88}$$

$$\leq n \cdot \kappa_{T,L}(k) \cdot (p + L) + \left(\kappa_{T,L}(k)\right)^3 \tag{A.89}$$

$$\leq 2 \cdot n \cdot \kappa_{T,L}(k) \cdot (p + L). \tag{A.90}$$

We can write $L = \lceil \eta \cdot p \rceil$, $\eta > 0$, and the expected run length can be upper bounded as follows:

$$\mathbb{E}\left[\kappa_{T,L}(k)\right] \leq p_1 + T + \mathbb{E}\left[\Psi\right] = p_1 + T + \frac{T}{L+1} \cdot p_0 \leq p_1 + 2T, \tag{A.91}$$

where the first equation follows from $\Psi \sim \text{NHG}(p_0 + L, p_0, T)$ and the second inequality holds because $L \geq p$. So, the expected computational complexity of one random experiment of the proposed *T-Rex* selector is $\mathcal{O}(np)$. Although the theoretical FDR control result requires $K \to \infty$, as stated in Section 3.2, choosing $K \geq 20$ provides excellent empirical results and we did not observe any notable improvements for $K \geq 100$. Therefore, with fixed K (e.g., $K = 20$), the overall expected computational complexity of the *T-Rex* selector is $\mathcal{O}(np)$. The computational complexity of the original (i.e., non-terminated) *LARS* algorithm in high-dimensional settings is $\mathcal{O}(p^3)$. Thus, on average the high computational complexity of the *LARS* algorithm does not carry over to the *T-Rex* selector because within the *T-Rex* selector the solution paths of the random experiments are early terminated. Moreover, the computational complexity of the *T-Rex* selector is the same as that of the pathwise coordinate descent algorithm [FHT10].

A.5 General Assumptions

It is important to note that existing theory for FDR control in high-dimensional settings, i.e., the *model-X* knockoff methods [Can+18], relies on an accurate estimation of the covariance matrix of the predictors, which is known to not be possible, in general, when $p \gg n$ (see, e.g., Figure 7 in [Can+18]). Further, the knockoff generation algorithm in [Can+18] is practically

infeasible due to its exponential complexity in p and the authors resort to second-order *model-X* knockoffs for which no FDR control proof exists. In contrast, the *T-Rex* selector does not rely on an accurate estimate of a high-dimensional covariance matrix and does not resort to an approximation of its theory to obtain a feasible algorithm.

Instead, to establish the FDR control theory for the *T-Rex* selector, we will introduce two general and mild assumptions that are thoroughly verified in relevant use-cases and especially for non-Gaussian simulated genomics data using the software HAPGEN2 [SMD11] (see Appendices A.6 and A.7).

Knockoff methods [BC15], as well as many popular FDR-controlling methods (i.e., [BH95; STS04; GBS09]) assume that the null p-values are i.i.d. and uniformly distributed between 0 and 1. In particular, to prove the FDR control property of the knockoff methods in [BC15; Can+18], the authors assume that the null p-values

1. are i.i.d.,
2. are independent of the p-values corresponding to the true active variables, and
3. stochastically dominate a random variable following the uniform distribution with support between 0 and 1.

Since we do not use p-values, we make a different assumption and explain how our weaker assumption is implied by the aforementioned standard assumptions.

A-I. *Let $\mathcal{A}$ and $\mathcal{Z}$ be the index sets of the true active and the null variables, respectively, and let the candidate variables $X_1, \ldots, X_p$ be standardized (i.e., $\mathbb{E}[X_j] = 0$ and $\mathrm{Var}[X_j] = 1$ for all $j \in \{1, \ldots, p\}$) and follow probability distributions with finite mean and finite non-zero variance. Then,*

(i) X_j is independent of $\{X_g : g \in \mathcal{A}\}$ for all $j \in \mathcal{Z}$, i.e., the null variables are independent of the true active variables,

(ii) $\{X_j : j \in \mathcal{Z}\}$ is a set of independent random variables, i.e., the nulls are mutually independent.

Remark 15. *Points 1 and 2 of the above standard assumption in FDR control theory state that the null p-values are i.i.d. and independent of the p-values corresponding to the true active variables. This can also be stated in terms of test statistics. That is, the test statistics corresponding to the null variables are i.i.d. and independent of the test statistics corresponding*

to the true active variables. Null p-values are defined by

$$P_j = 1 - F_0(T_j), \tag{A.92}$$

where P_j and T_j are the null p-value and the null test statistic, respectively, corresponding to the jth null variable and $F_0(\cdot)$ is the distribution of the null test statistics [SS99]. From this definition of p-values, it is obvious that this assumption can be stated equivalently either in terms of p-values or test statistics, which is frequently done [BH95; BY01; STS04; BC15]. As stated in [BC15] (p. 2075), especially in the case where the test statistics stem from the regression coefficient estimates $\hat{\boldsymbol{\beta}} = \left[\hat{\beta}_1 \cdots \hat{\beta}_p\right]^\top \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right)$, the coefficient estimates (and the test statistics) are mutually independent if and only if $\boldsymbol{X}^\top \boldsymbol{X}$ is a diagonal matrix (i.e., orthogonal design). This implies that the null test statistics are mutually independent and independent of the test statistics corresponding to the true active variables if and only if the null variables are mutually independent and independent of the true active variables. Note that this is what we are stating in our A-I. Thus, the standard assumption in FDR control theory implies A-I and, since this implication does not require Point 3 of the above standard assumption in FDR control theory, A-I is weaker.

For a numerical verification of A-I in relevant use-cases and especially for non-Gaussian simulated genomics data using the software HAPGEN2 [SMD11], see Appendices A.6 and A.7.

Remark 16. *In the genomics literature, it is well-known that SNPs (i.e., variables) form groups of highly correlated SNPs. The biological phenomenon that leads to such dependency structures is called linkage disequilibrium [Rei+01]. It is common in genomics research to use pruning methods to group SNPs and to keep only one representative SNP from each group and, thus, drastically reduce the dependencies among the SNPs before applying any variable selection procedure (see, e.g., [SSC19] and references therein). Therefore, SNP pruning is a valid method to satisfy A-I in practice. When choosing the amount of pruning (i.e., the number of groups that the SNPs are grouped into) one must consider the trade-off between*

1. *the reduction of dependencies among SNPs (by creating few SNP groups) and*
2. *the increase of the resolution of the to be detected regions on the genome (by creating many SNP groups).*

For details on how this trade-off is commonly tuned for GWAS, see Section 6.1.3.

As shown in Figure 3.2, the estimator of the active set $\widehat{\mathcal{A}}(v)$ results from fusing the candidate

sets $\mathcal{C}_{1,L}(T), \ldots, \mathcal{C}_{K,L}(T)$ based on a voting level that is applied to the relative occurrences of the candidate variables. Therefore, the number of selected null variables $V_{T,L}(v)$ is related to the distribution of the number of included null variables in the terminal step $t = T$. We state this relationship as an assumption:

A-II. *For any $v \in [0.5, 1)$, the number of selected null variables is stochastically dominated by a random variable following the negative hypergeometric distribution with parameters specified in Corollary 3, i.e.,*

$$V_{T,L}(v) \overset{d}{\leq} \mathrm{NHG}(p_0 + L, p_0, T). \tag{A.93}$$

For a numerical verification of A-II in relevant use-cases and especially for non-Gaussian simulated genomics data using the software HAPGEN2 [SMD11], see Appendices A.6 and A.7.

The expression for $\widehat{V}'_{T,L}(v)$ from Remark 4 can be rewritten as follows:

$$\widehat{V}'_{T,L}(v) \tag{A.94}$$

$$= \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi_{t,L}(q)}{L - (t-1)} \cdot \frac{\sum_{j \in \widehat{\mathcal{A}}(v)} \Delta\Phi_{t,L}(j)}{\sum_{q \in \widehat{\mathcal{A}}(0.5)} \Delta\Phi_{t,L}(q)} \tag{A.95}$$

$$= \sum_{t=1}^{T} \frac{p_0 - \sum_{q \in \mathcal{Z}} \Phi_{t,L}(q) + \overline{p_1 - \sum_{q \in \mathcal{A}} \Phi_{t,L}(q)}}{L - (t-1)} \tag{A.96}$$

$$\cdot \frac{\sum_{j \in \widehat{\mathcal{A}}^0(v)} \Delta\Phi_{t,L}(j) + \overline{\sum_{j \in \widehat{\mathcal{A}}^1(v)} \Delta\Phi_{t,L}(j)}}{\sum_{q \in \widehat{\mathcal{A}}^0(0.5)} \Delta\Phi_{t,L}(q) + \overline{\sum_{q \in \widehat{\mathcal{A}}^1(0.5)} \Delta\Phi_{t,L}(q)}} \tag{A.97}$$

$$\approx \sum_{t=1}^{T} \frac{p_0 - \sum_{q \in \mathcal{Z}} \Phi_{t,L}(q)}{L - (t-1)} \cdot \frac{\sum_{j \in \widehat{\mathcal{A}}^0(v)} \Delta\Phi_{t,L}(j)}{\sum_{q \in \widehat{\mathcal{A}}^0(0.5)} \Delta\Phi_{t,L}(q)} \tag{A.98}$$

The marked terms consider only the relative occurrences of the active variables. Recall that, assuming that the variable selection method is better than random selection, almost all active variables are selected early, i.e., terminating the *T-Rex* selector after a small number of T dummies have been included allows to select almost all active variables (see Figure 3.5a). Thus, the relative occurrences of the active variables are approximately one for a sufficient number of included dummies. In consequence, and since $\Delta\Phi_{t,L} = \Phi_{t,L} - \Phi_{t-1,L}$, $t \in \{1, \ldots, T\}$, the $\Delta\Phi_{t,L}$'s of the active variables are approximately zero for a sufficiently large t and T. This

motivates the assumption that the marked terms can be neglected.

A-III. *For sufficiently large* $T \in \{1, \ldots, L\}$ *it holds that*

$$\widehat{V}'_{T,L}(v) = \sum_{t=1}^{T} \frac{p_0 - \sum_{q \in \mathcal{Z}} \Phi_{t,L}(q)}{L - (t-1)} \cdot \frac{\sum_{j \in \widehat{\mathcal{A}}^0(v)} \Delta\Phi_{t,L}(j)}{\sum_{q \in \widehat{\mathcal{A}}^0(0.5)} \Delta\Phi_{t,L}(q)}. \tag{A.99}$$

See Appendices A.6 and A.7 for extensive numerical verifications of A-III.

A.6 Exemplary Numerical Verification of A-I, A-II, and A-III

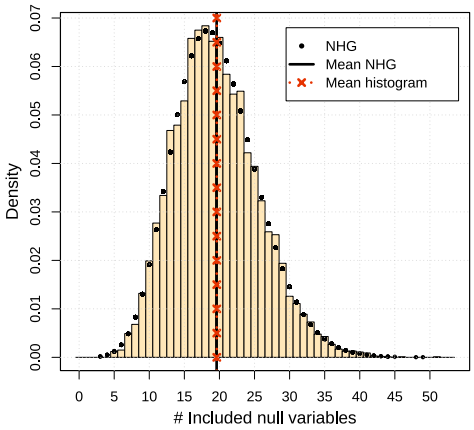
In this section, A-I, A-II, and A-III from Appendix A.5 are verified. The general setup for the exemplary numerical verification of all assumptions is as described in Section 3.6.1. The specific values of the generic high-dimensional simulation setting in Section 3.6.1 and the parameters of the proposed *T-Rex* selector and the proposed extended calibration algorithm in Algorithm 3, i.e., the values of n, p, p_1, v, T, L, K, and SNR are specified in the figure captions. All results are averaged over $MC = 500$ Monte Carlo realizations. An additional verification for our use-case of GWAS is provided in Appendix A.7.

A.6.1 Exemplary Numerical Verification of A-I

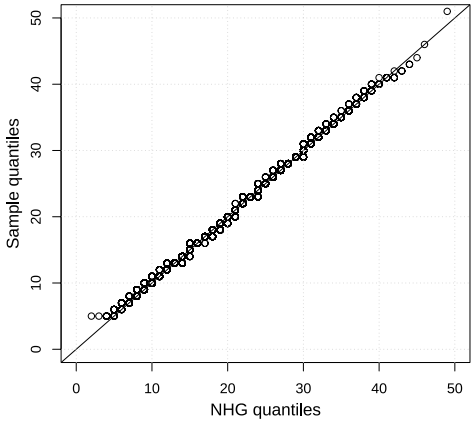
Figure A.1a shows the histogram of the number of included null variables for $T = 20$ and for 500 Monte Carlo replications consisting of $K = 20$ candidate sets each while Figure A.1b shows the corresponding Q-Q plot. The histogram closely approximates the probability mass function (PMF) of the negative hypergeometric distribution with the parameters specified in Corollary 4. Moreover, the points in the Q-Q plot closely approximate the ideal line. Thus, Figure A.1 provides an exemplary numerical verification of Corollary 4 and, therewith, an implicit exemplary verification of A-I.

A.6.2 Exemplary Numerical Verification of A-II

Figure A.2 shows the empirical cumulative distribution function (CDF) of $V_{T,L}(v)$ for $T = 20$ and different values of the voting level v and the CDF of the negative hypergeometric distribution. The empirical CDFs are based on 500 Monte Carlo replications. Already for a small number of random experiments, i.e., $K = 20$, the CDF of the negative hypergeometric distri-



(a) Histogram and theoretical distribution for $t = T = 20$. Note that the histogram is based on $K = 20$ random experiments for each of the 500 Monte Carlo realizations.



(b) Q-Q plot corresponding to Figure (a).

Figure A.1: Exemplary numerical verification of Corollary 4 and A-I: The histogram of the number of included null variables in Figure (a) approximates the theoretical probability mass function (PMF). The expected value of a random variable following the negative hypergeometric distribution with the parameters specified in the last sentence of this caption is given by $T \cdot p_0 \,/\, (L+1) = 20 \cdot 290 \,/\, (300+1) \approx 19.27$, which fits the mean of the histogram. The Q-Q plot in Figure (b) confirms that the number of included null variables follows the negative hypergeometric distribution. Setup: $n = 150$, $p = 300$, $p_1 = 5$, $T = 20$, $L = p$, $K = 20$, SNR $= 1$, $MC = 500$.

bution with its parameters being as specified in A-II lies below the empirical CDFs of $V_{T,L}(v)$ for all $v \geq 0.5$ at almost all values of $V_{T,L}(v)$. For values of $V_{T,L}(v)$ between 6 and 12, we observe that the CDF of the negative hypergeometric distribution lies slightly above the empirical CDF for $v = 0.5$. All in all, we conclude that a random variable following the negative hypergeometric distribution stochastically dominates $V_{T,L}(v)$ at almost all values and for all $v \geq 0.5$, which exemplarily verifies A-II.

A.6.3 Exemplary Numerical Verification of A-III

An exemplary numerical verification of A-III is given in Figure A.3, where we see that approximations and true values are almost identical for different choices of v and T.

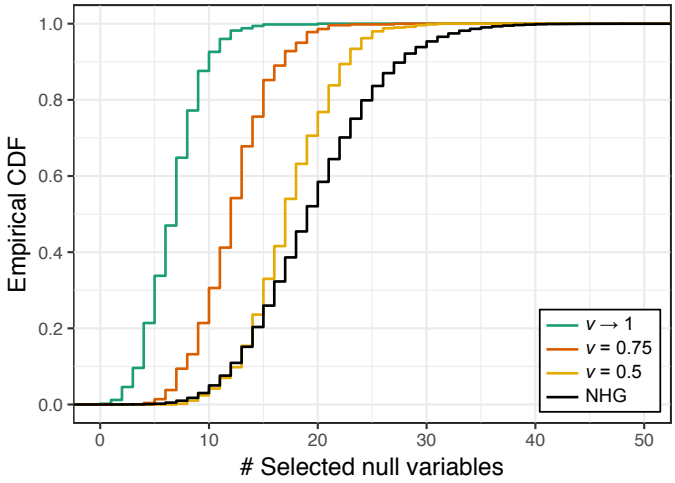


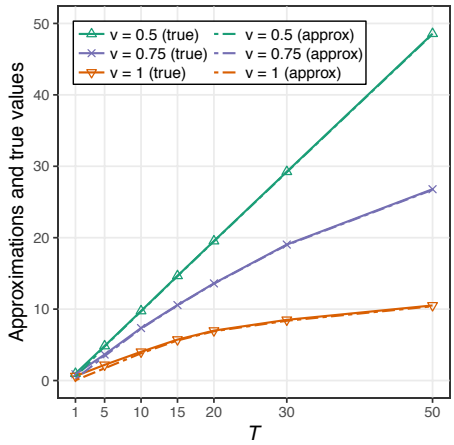
Figure A.2: Exemplary numerical verification of A-II: For $v \geq 0.5$, a random variable following the negative hypergeometric distribution stochastically dominates the random variable $V_{T,L}(v)$ (i.e., the number of selected null variables) at almost all values of $V_{T,L}(v)$. Setup: $n = 150$, $p = 300$, $p_1 = 5$, $T = 20$, $L = p$, $K = 20$, SNR $= 1$, $MC = 500$.

A.7 Verification of A-I, A-II, and A-III on HAPGEN2 Genomics Data

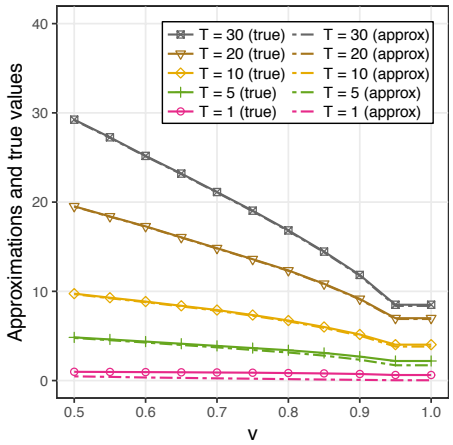
Figures A.4, A.5, and A.6 show that for the genomics data analyzed in Section 6.1 and with the preprocessing (i.e., SNP pruning, etc.) described above, A-I, A-II, and A-III are surprisingly well satisfied. For our verifications here, we have only made one necessary minor adjustment to the preprocessing described in the previous section. The reason is that for each of the 100 data sets, that have been generated using HAPGEN2 [SMD11], the SNP pruning procedure outputs pruned SNP sets with slightly different sizes. For the verification of the assumptions, it is necessary to have a constant number of SNPs. Therefore, we have removed very few randomly selected SNPs from all sets in order to match the size of the smallest SNP set, which contains 8,120 out of originally 20,000 SNPs after the preprocessing.

A.8 Additional Simulation Results

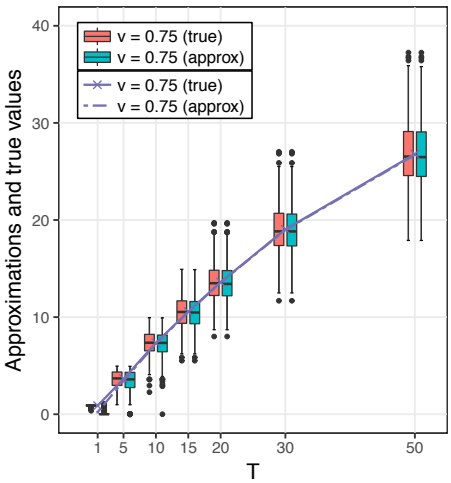
For the sake of completeness, we present additional simulation results for the classical low-dimensional setting, i.e., $p \leq n$. The data is generated as described in Section 3.6.1. The specific values of the generic simulation setting in Section 3.6.1 and the parameters of the proposed *T-Rex* selector and the proposed extended calibration algorithm in Algorithm 3, i.e.,



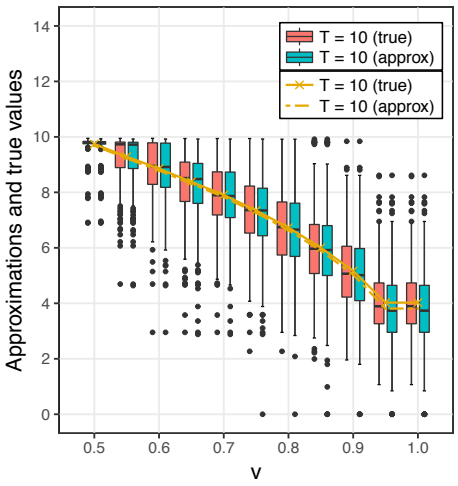
(a) Approximations and true values for different choices of v averaged over 500 Monte Carlo realizations.



(b) Approximations and true values for different choices of T averaged over 500 Monte Carlo realizations.

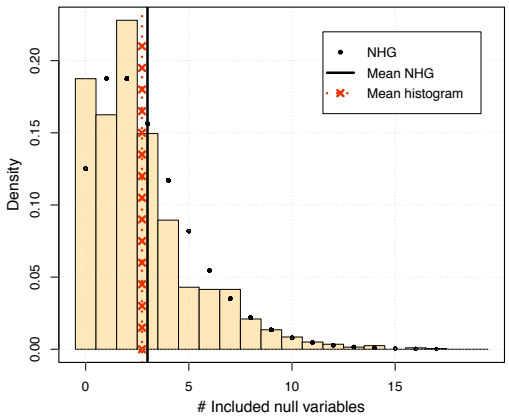


(c) Box plots of approximations and true values corresponding to the lines for $v = 0.75$ in Figure (a).

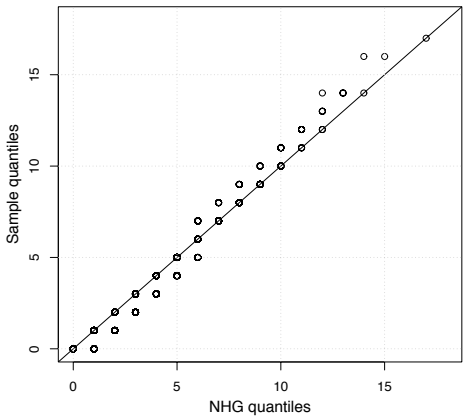


(d) Box plots of approximations and true values corresponding to the lines for $T = 10$ in Figure (b).

Figure A.3: Exemplary numerical verification of A-III: In Figures (a) and (b), we see that the approximations and the true values are almost identical for different values of v and T. The corresponding box plots in Figures (c) and (d) show that also the distributions of approximations and true values are very similar. Setup: $n = 150, p = 300, p_1 = 5, L = p, K = 20$, SNR $= 1, MC = 500$.



(a) Histogram and theoretical distribution for $t = T = 3$. Note that the histogram is based on $K = 20$ random experiments for each of the 100 HAPGEN2 genomics data sets.



(b) Q-Q plot corresponding to Figure (a).

Figure A.4: Exemplary numerical verification of Corollary 4 and A-I for HAPGEN2 genomics data: The histogram of the number of included null variables in Figure (a) approximates the theoretical probability mass function (PMF). The expected value of a random variable following the negative hypergeometric distribution with the parameters specified in the last sentence of this caption is given by $T \cdot p_0 \,/\, (L+1) = 3 \cdot 8{,}110 \,/\, (8{,}120+1) \approx 2.996$, which fits the mean of the histogram. The Q-Q plot in Figure (b) confirms that the number of included null variables follows the negative hypergeometric distribution. Setup after preprocessing: $n = 1{,}000, p = 8{,}120, p_1 = 10, T = 3, L = p, K = 20$.

the values of n, p, p_1, T_{max}, L_{max}, K, and SNR are specified in the captions of Figure A.7. All results are averaged over 955 Monte Carlo realizations. The simulations were conducted using the R packages *TRexSelector* [Mac+24c] and *tlars* [Mac+24b].

Summarizing in brief, the proposed *T-Rex* selector controls the FDR at the target level of 10% while, in terms of power, outperforming the *fixed-X* knockoff method, the *fixed-X* knockoff+ method, and the *BY* method and showing a comparable performance to the *BH* method. A detailed discussion of the simulation results is given in the captions of Figure A.7 and its subfigures.

A.9 Illustration of Theorem 4 (Dummy Generation)

Theorem 4 is an asymptotic result that, loosely speaking, tells us that the FDR control property of the *T-Rex* selector remains intact regardless of the distribution that the dummies are sampled from. In order to exemplify the somehow surprising results of Theorem 4, we have conducted simulations to show that the FDR control property of the *T-Rex* selector remains

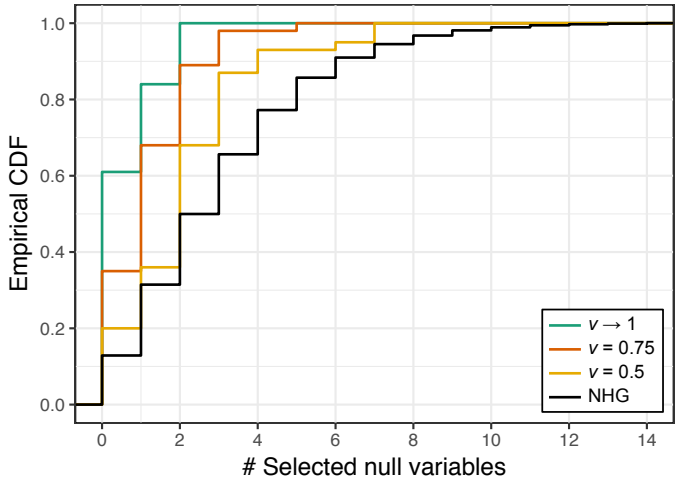


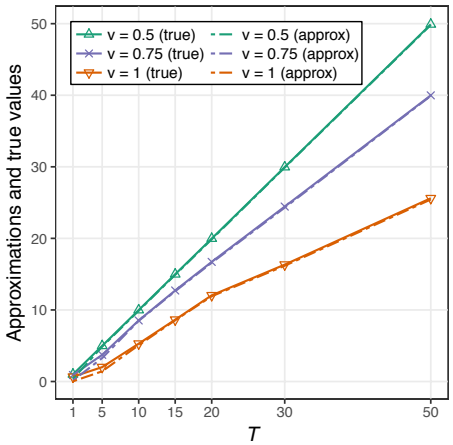
Figure A.5: Exemplary numerical verification of A-II for HAPGEN2 genomics data: For $v \geq 0.5$, a random variable following the negative hypergeometric distribution stochastically dominates the random variable $V_{T,L}(v)$ (i.e., the number of selected null variables) at all values of $V_{T,L}(v)$. Setup after preprocessing: $n = 1{,}000, p = 8{,}120, p_1 = 10, T = 3, L = p, K = 20$.

intact for dummies sampled from the standard normal, uniform, t-, and Gumbel distribution, while the original predictors are sampled from the standard normal distribution. In Figure A.8, we see that the results remain almost unchanged regardless of the choice of the dummy distribution.

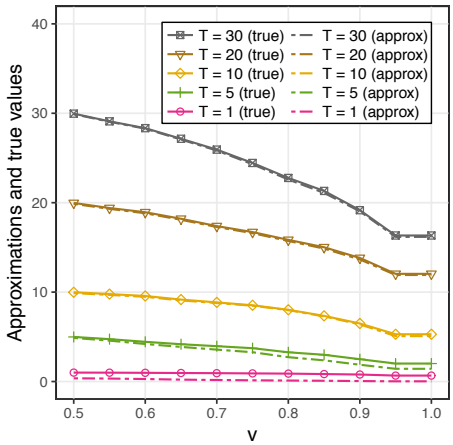
In order to also verify that the FDR control property holds for different distributions (with finite mean and finite non-zero variance) of the original predictors, we have conducted simulations in which the dummies are sampled from a standard normal distribution, while the original predictors are sampled from non-Gaussian heavy-tailed (i.e., Student's $t(3)$, $t(2.1)$, and $t(2.01)$) and skewed (i.e., Gumbel$(0, 1)$) distributions. Figure A.9 shows that, regardless of the mismatch between the distribution of the original variables and the dummies, the FDR control property holds for all these different distributions.

A.10 Robustness of The T-Rex Selector

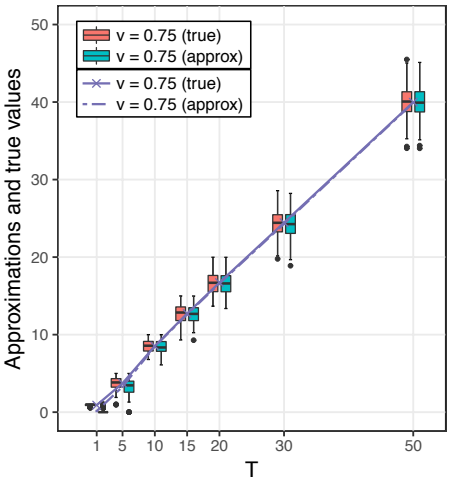
In this appendix, we investigate the robustness of the proposed *T-Rex* selector in the presence of non-Gaussian noise. We have conducted simulations with heavy-tailed noise following the t-distribution with three degrees of freedom. Figure A.10 shows that the proposed method performs well, even in the presence of heavy-tailed noise and, most importantly, maintains its FDR control property.



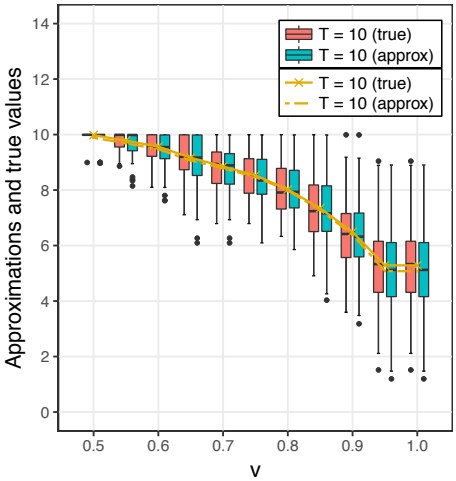
(a) Approximations and true values for different choices of v averaged over 100 HAPGEN2 genomics data sets.



(b) Approximations and true values for different choices of T averaged over 100 HAPGEN2 genomics data sets.

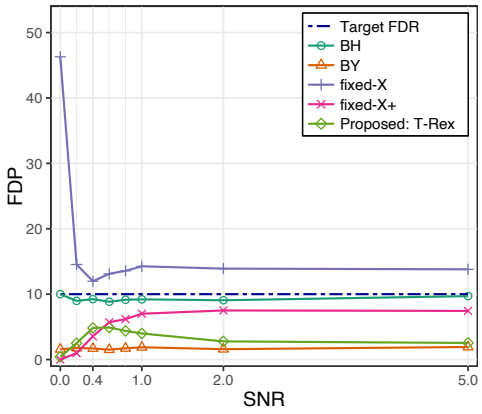


(c) Box plots of approximations and true values corresponding to the lines for $v = 0.75$ in Figure (a).

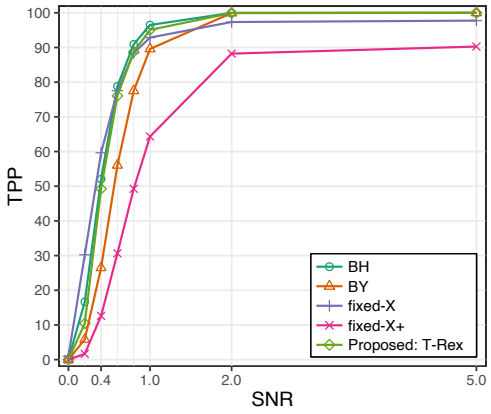


(d) Box plots of approximations and true values corresponding to the lines for $T = 10$ in Figure (b).

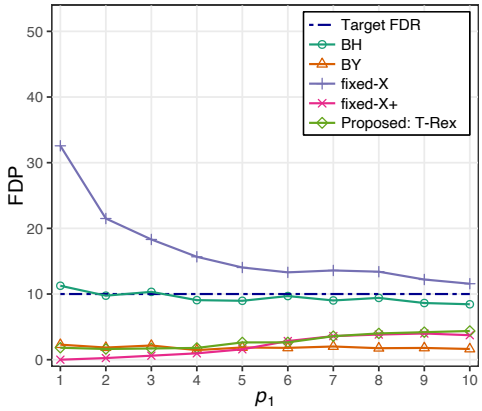
Figure A.6: Exemplary numerical verification of A-III for HAPGEN2 genomics data: In Figures (a) and (b), we see that the approximations and the true values are almost identical for different values of v and T. The corresponding box plots in Figures (c) and (d) show that also the distributions of approximations and true values are very similar. Setup after preprocessing: $n = 1{,}000$, $p = 8{,}120$, $p_1 = 10$, $L = p$, $K = 20$.



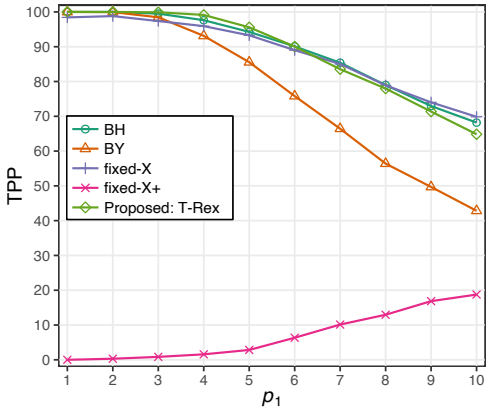
(a) Setup: $n = 300, p = 100, p_1 = 10, T_{\max} = \lceil n/2 \rceil, L_{\max} = 10p, K = 20, MC = 955$.



(b) Setup: Same as in Figure (a).

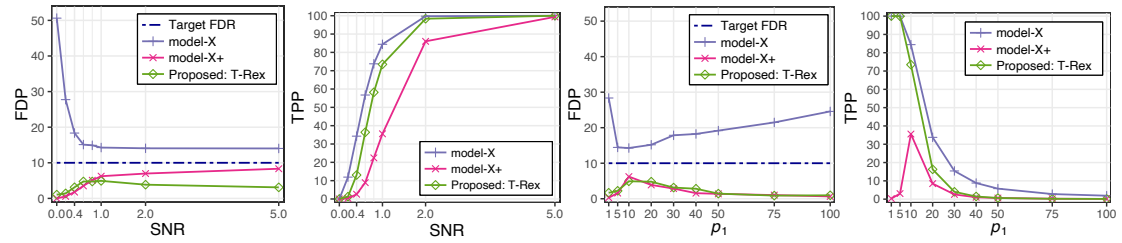


(c) Setup: $n = 300, p = 100, T_{\max} = \lceil n/2 \rceil, L_{\max} = 10p, K = 20, \text{SNR} = 1, MC = 955$.

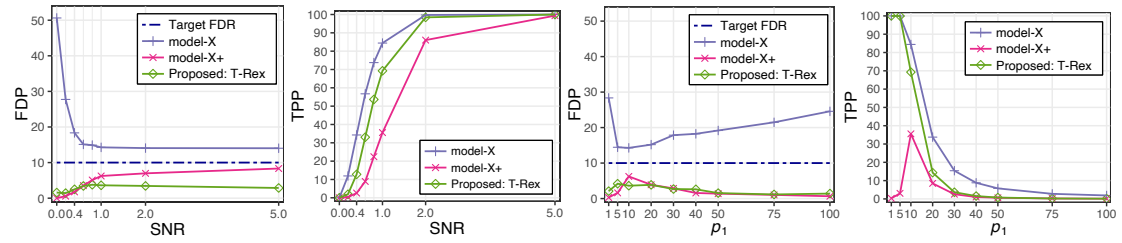


(d) Setup: Same as in Figure (c).

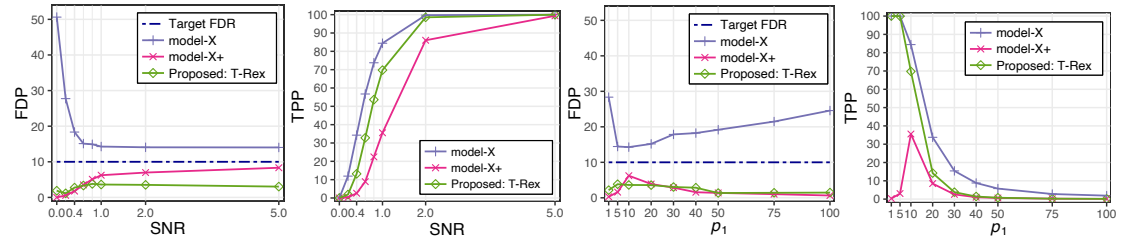
Figure A.7: General: The *fixed-X* knockoff method fails to control the FDR. In terms of power, the proposed *T-Rex* selector outperforms the *fixed-X* knockoff method, the *fixed-X* knockoff+ method, and the *BY* method and shows a comparable performance to the *BH* method. **Details**: **(a)** All methods except for the *fixed-X* knockoff method control the FDR at a target level of 10% for the whole range of SNR values. The *fixed-X* knockoff method fails to control the FDR and performs poorly at low SNR values. **(b)** As expected, the TPR (i.e., power) increases with respect to the SNR. It is remarkable that the TPP (i.e., power) of the proposed *T-Rex* selector is comparable to that of the *BH* method, although the FDR of the *T-Rex* selector is less than half of the achieved FDR of the *BH* method (see Figure (a)). The high power of the *fixed-X* knockoff method cannot be interpreted as an advantage because the method does not control the FDR. **(c)** The proposed *T-Rex* selector, the *fixed-X* knockoff+ method, and the *BY* method control the FDR at a target level of 10%, while the *BH* method exceeds the target level for some low values of p_1 and the curve of the *fixed-X* knockoff method never falls below the target level. **(d)** Among the methods that control the FDR for all considered values of p_1, the proposed *T-Rex* selector has the highest power. It is remarkable that the TPP (i.e., power) of the proposed *T-Rex* selector is comparable to that of the *BH* method, although the FDR of the *T-Rex* selector is approximately only half of the achieved FDR of the *BH* method (see Figure (c)).



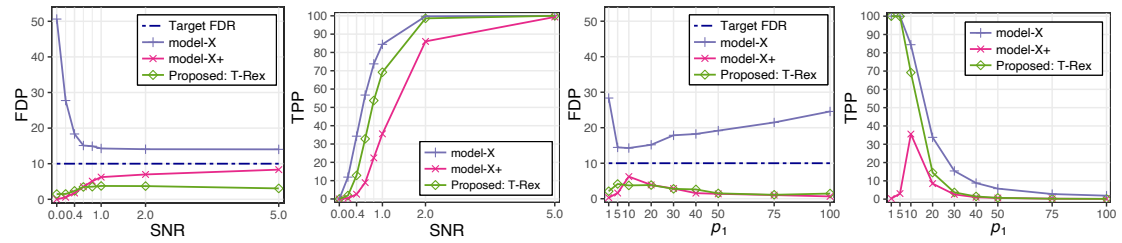
(a) Gaussian dummies. **(b)** Gaussian dummies. **(c)** Gaussian dummies. **(d)** Gaussian dummies.



(e) Uniform dummies. **(f)** Uniform dummies. **(g)** Uniform dummies. **(h)** Uniform dummies.

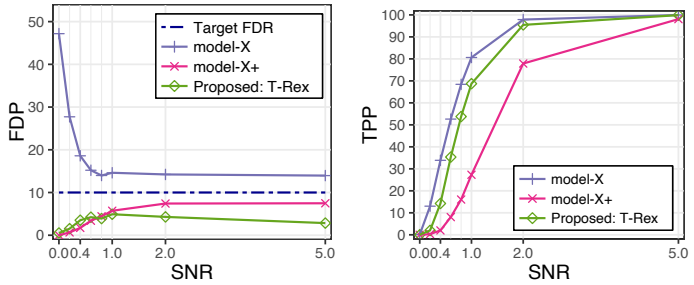


(i) Student t dummies. **(j)** Student t dummies. **(k)** Student t dummies. **(l)** Student t dummies.

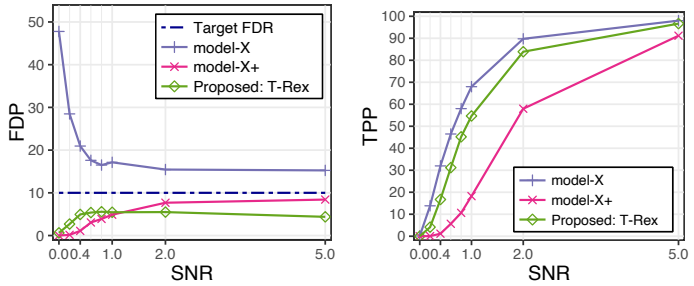


(m) Gumbel dummies. **(n)** Gumbel dummies. **(o)** Gumbel dummies. **(p)** Gumbel dummies.

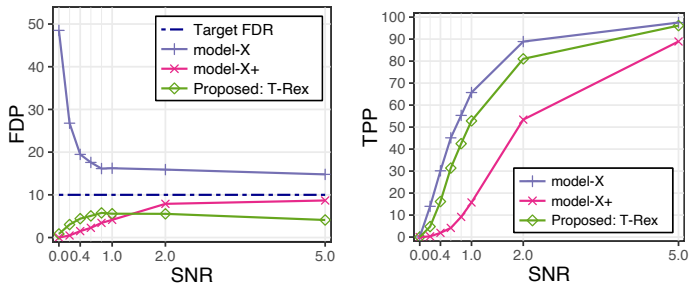
Figure A.8: Illustration of Theorem 4 (Dummy generation): The average FDP and TPP of the *T-Rex* selector remain almost unchanged regardless of the distribution that the dummies are sampled from: (a) - (d) standard normal distribution, (e) - (h) uniform distribution with support between 0 and 100, (i) - (l) Student's t-distribution with 3 degrees of freedom, (m) - (p) Gumbel distribution with its location and scale being 0 and 1, respectively. Setup: $n = 300$, $p = 1{,}000$, $p_1 = 10$, $T_{\max} = \lceil n/2 \rceil$, $L_{\max} = 10p$, $K = 20$, SNR $= 1$, $MC = 955$.



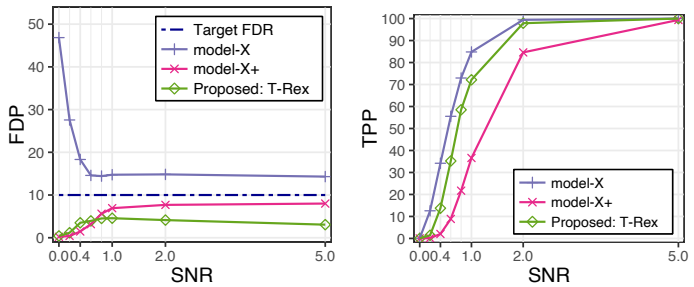
(a) $t(3)$ distributed $\boldsymbol{X}$. **(b)** $t(3)$ distributed $\boldsymbol{X}$.



(c) $t(2.1)$ distributed $\boldsymbol{X}$. **(d)** $t(2.1)$ distributed $\boldsymbol{X}$.



(e) $t(2.01)$ distributed $\boldsymbol{X}$. **(f)** $t(2.01)$ distributed $\boldsymbol{X}$.



(g) Gumbel$(0, 1)$ distributed $\boldsymbol{X}$. **(h)** Gumbel$(0, 1)$ distributed $\boldsymbol{X}$.

Figure A.9: Average FDP and TPP in the case of non-Gaussian predictors in $\boldsymbol{X}$: The FDR is controlled by the *T-Rex* selector and the *model-X* knockoff+ method while the *model-X* knockoff method does not control the FDR. The predictors in $\boldsymbol{X}$ were sampled from (a) - (f) the Student's t distribution with 3, 2.1, and 2.01 degrees of freedom (i.e., $t(3)$, $t(2.1)$, and $t(2.01)$) and (g) - (h) the Gumbel distribution with location and scale being zero and one (i.e., Gumbel$(0, 1)$), respectively. The response was generated according to the linear model in (2.1). Setup: $n = 300$, $p = 1{,}000$, $p_1 = 10$, $T_{\max} = \lceil n/2 \rceil$, $L_{\max} = 10p$, $K = 20$, $MC = 955$.

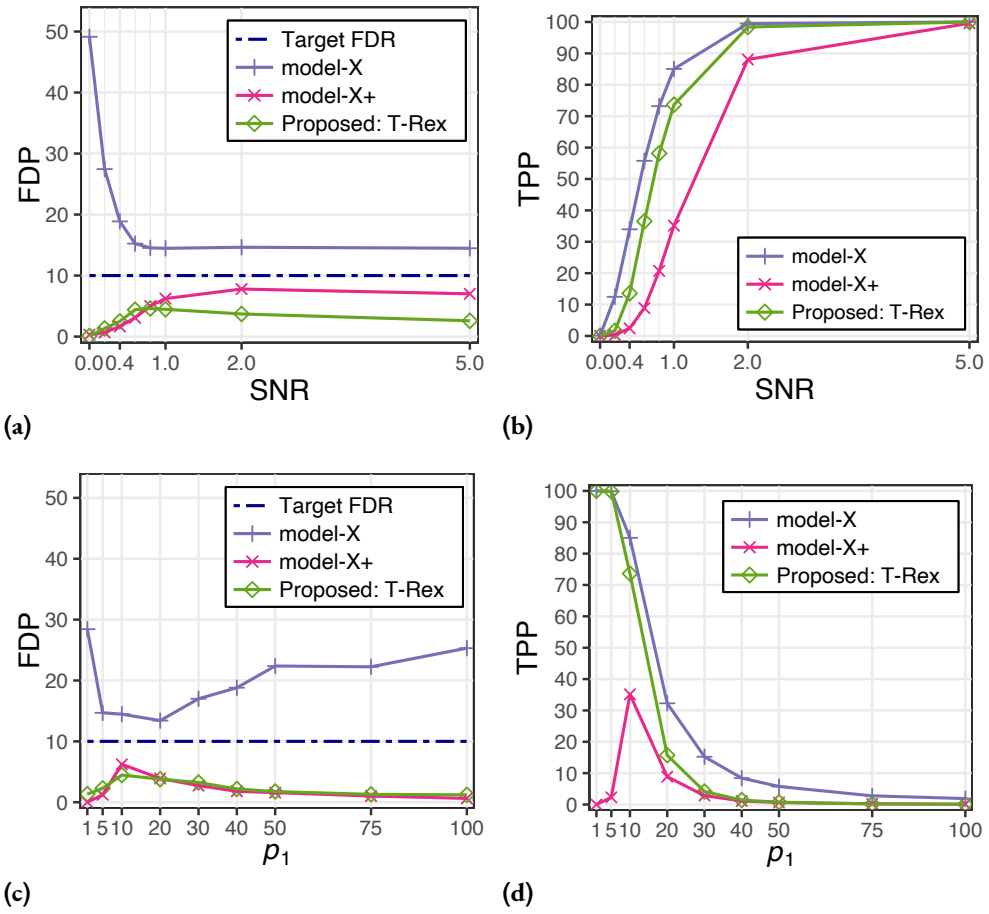


Figure A.10: Average FDP and TPP in the case of non-Gaussian noise: The results are similar to the results of the Gaussian noise case in Figure 3.10. That is, all considered methods appear to be robust against deviations from the Gaussian noise assumption for the case of heavy-tailed (t-distributed) noise. The predictors in $\boldsymbol{X}$ were sampled from a univariate standard normal distribution and the response was generated according to the linear model in (2.1) with the noise vector ϵ being sampled from the t-distribution with 3 degrees of freedom. Setup: $n = 300, p = 1{,}000, p_1 = 10, T_{\text{max}} = \lceil n/2 \rceil, L_{\text{max}} = 10p, K = 20, \text{SNR} = 1, MC = 955$.

B

Appendix to Chapter 4

This appendix to Chapter 4 is organized as follows: Appendix B.1 contains some preliminaries for Theorem 8, the proofs of Theorems 8, 9, 10, 11, and 12, and some technical lemmas. Appendix B.2 presents additional simulation results for settings with heavy-tailed data.

B.1 Proofs and Technical Lemmas

First, some preliminary results for the proof of Theorem 8 are stated. Then, the proofs of Theorems 8, 9, 10, 11, and 12 are presented. Finally, some technical lemmas, which are used to prove the finite sample FDR control property of the *T-Rex+DA* selector, are stated and proved.

B.1.1 Preliminaries for Theorem 8

Lemma 6. *Let* $\rho_{j,j'} := \boldsymbol{x}_j^\top \boldsymbol{x}_{j'}$, $j, j' \in \{1, \dots, p\}$, *be the sample correlation coefficient of the standardized variables* j *and* j', $\hat{\boldsymbol{r}}_k := \boldsymbol{y} - \widetilde{\boldsymbol{X}}_k \hat{\boldsymbol{\beta}}_k$, *and* $\operatorname{sign}\left(\hat{\beta}_{j,k}\right)$, $\operatorname{sign}\left(\hat{\beta}_{j',k}\right)$ *be the signs of the* j*th and* j'*th Lasso coefficient estimates of the* k*th random experiment, respectively.*

Suppose that $\hat{\beta}_{j,k}, \hat{\beta}_{j',k} \neq 0$. *Then, it holds that*

$$\left| \left| \operatorname{sign}\left(\hat{\beta}_{j,k}\right) \right| - \left| \operatorname{sign}\left(\hat{\beta}_{j',k}\right) \right| \right| \leq \frac{\|\hat{\boldsymbol{r}}_k\|_2 \sqrt{2\left(1-\rho_{j,j'}\right)}}{\lambda_k(T,L)}. \tag{B.1}$$

Proof. The *Lasso* optimization problem is solved by the coefficient vector $\boldsymbol{\beta}_k = \hat{\boldsymbol{\beta}}_k$ that minimizes the function

$$\mathcal{L}\left(\boldsymbol{\beta}_k, \lambda_k(T,L)\right) \coloneqq \frac{1}{2}\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}_k \boldsymbol{\beta}_k\|_2^2 + \lambda_k(T,L)\|\boldsymbol{\beta}_k\|_1. \tag{B.2}$$

Taking the first derivative of (B.2) and setting it equal to zero yields

$$\left.\frac{\partial \mathcal{L}\left(\boldsymbol{\beta}_k, \lambda_k(T,L)\right)}{\partial \boldsymbol{\beta}_k}\right|_{\boldsymbol{\beta}_k = \hat{\boldsymbol{\beta}}_k} = -\widetilde{\boldsymbol{X}}_k^\top \hat{\boldsymbol{r}}_k + \lambda_k(T,L)\frac{\partial \|\hat{\boldsymbol{\beta}}_k\|_1}{\partial \hat{\boldsymbol{\beta}}_k} \stackrel{!}{=} 0, \tag{B.3}$$

which is a system of equations whose jth and j'th equation are given by

$$-\boldsymbol{x}_j^\top \hat{\boldsymbol{r}}_k + \lambda_k(T,L) \cdot \operatorname{sign}\left(\hat{\beta}_{j,k}\right) = 0, \tag{B.4}$$

$$-\boldsymbol{x}_{j'}^\top \hat{\boldsymbol{r}}_k + \lambda_k(T,L) \cdot \operatorname{sign}\left(\hat{\beta}_{j',k}\right) = 0. \tag{B.5}$$

Subtracting Equation (B.5) from Equation (B.4) and rearranging the resulting equation yields

$$\operatorname{sign}\left(\hat{\beta}_{j,k}\right) - \operatorname{sign}\left(\hat{\beta}_{j',k}\right) = \frac{1}{\lambda_k(T,L)}\left(\boldsymbol{x}_j - \boldsymbol{x}_{j'}\right)^\top \hat{\boldsymbol{r}}_k. \tag{B.6}$$

Now, we can rewrite the left-hand side of (B.1) as follows:

$$\begin{aligned}
&\left| \left| \operatorname{sign}\left(\hat{\beta}_{j,k}\right) \right| - \left| \operatorname{sign}\left(\hat{\beta}_{j',k}\right) \right| \right| && \text{(B.7)}\\
&\qquad \leq \left| \operatorname{sign}\left(\hat{\beta}_{j,k}\right) - \operatorname{sign}\left(\hat{\beta}_{j',k}\right) \right| && \text{(B.8)}\\
&\qquad = \frac{1}{\lambda_k(T,L)} \left| \left(\boldsymbol{x}_j - \boldsymbol{x}_{j'}\right)^\top \hat{\boldsymbol{r}}_k \right| && \text{(B.9)}\\
&\qquad \leq \frac{1}{\lambda_k(T,L)} \|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|_2 \cdot \|\hat{\boldsymbol{r}}_k\|_2 && \text{(B.10)}
\end{aligned}$$

$$= \frac{\|\hat{\boldsymbol{r}}_k\|_2 \sqrt{2\big(1-\rho_{j,j'}\big)}}{\lambda_k(T,L)}. \tag{B.11}$$

The inequality in the second line follows from the reverse triangle inequality, the equation in the third line follows from (B.6), the inequality in the fourth line follows from the Cauchy-Schwartz inequality, and the equation in the last line is a consequence of

$$\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|_2^2 = (\boldsymbol{x}_j - \boldsymbol{x}_{j'})^\top (\boldsymbol{x}_j - \boldsymbol{x}_{j'}) \tag{B.12}$$

$$= \|\boldsymbol{x}_j\|_2^2 + \|\boldsymbol{x}_{j'}\|_2^2 - 2\boldsymbol{x}_j^\top \boldsymbol{x}_{j'} \tag{B.13}$$

$$= 2(1-\rho_{j,j'}), \tag{B.14}$$

where the last line follows from the fact that the variables are standardized. □

B.1.2 Proof of Theorem 8

Proof. Note that for any $\hat{\beta}_{j,k} \neq 0$, the indicator function in (3.4) can be written as follows:

$$\mathbb{1}_k(j,T,L) = \big|\operatorname{sign}\big(\hat{\beta}_{j,k}\big)\big|, \quad j = 1,\dots,p. \tag{B.15}$$

Thus, we can rewrite the left-hand side of the inequality in Theorem 8 as follows:

$$\frac{\big|\Phi_{T,L}(j) - \Phi_{T,L}(j')\big|}{\|\boldsymbol{y}\|_2} \tag{B.16}$$

$$= \frac{1}{\|\boldsymbol{y}\|_2 \cdot K} \left| \sum_{k=1}^{K} \big(\mathbb{1}_k(j,T,L) - \mathbb{1}_k(j',T,L)\big) \right| \tag{B.17}$$

$$\leq \frac{1}{\|\boldsymbol{y}\|_2 \cdot K} \sum_{k=1}^{K} \Big|\mathbb{1}_k(j,T,L) - \mathbb{1}_k(j',T,L)\Big| \tag{B.18}$$

$$= \frac{1}{\|\boldsymbol{y}\|_2 \cdot K} \sum_{k=1}^{K} \Big|\big|\operatorname{sign}\big(\hat{\beta}_{j,k}\big)\big| - \big|\operatorname{sign}\big(\hat{\beta}_{j',k}\big)\big|\Big| \tag{B.19}$$

$$\leq \sqrt{2(1-\rho_{j,j'})} \cdot \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\lambda_k(T,L)} \cdot \frac{\|\hat{\boldsymbol{r}}_k\|_2}{\|\boldsymbol{y}\|_2} \tag{B.20}$$

$$\leq \sqrt{2(1-\rho_{j,j'})} \cdot \overline{\Lambda}, \tag{B.21}$$

where $\overline{\Lambda} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{\lambda_k(T,L)}$. The equation in the second line follows from the definition of the relative occurrences in (3.4), the inequality in the third line is a consequence of the triangle inequality, the equation in the fourth line follows from (B.15), and the inequality in the fifth line follows from Lemma 6. The inequality in the last line holds since $\hat{\boldsymbol{\beta}}_k$ is by definition the minimizer of (B.2) and, therefore,

$$\mathcal{L}\big(\boldsymbol{\beta}_k = \hat{\boldsymbol{\beta}}_k, \lambda_k(T,L)\big) \leq \mathcal{L}\big(\boldsymbol{\beta}_k = \mathbf{0}, \lambda_k(T,L)\big) \tag{B.22}$$

and, equivalently,

$$\frac{1}{2}\|\hat{\boldsymbol{r}}_k\|_2^2 + \lambda_k(T,L)\|\hat{\boldsymbol{\beta}}_k\|_1 \leq \frac{1}{2}\|\boldsymbol{y}\|_2^2, \tag{B.23}$$

which yields $\|\hat{\boldsymbol{r}}_k\|_2 \leq \|\boldsymbol{y}\|_2$. Finally, we obtain

$$\big|\Phi_{T,L}(j) - \Phi_{T,L}(j')\big| \leq \overline{\Lambda}\|\boldsymbol{y}\|_2 \cdot \sqrt{2(1-\rho_{j,j'})}. \tag{B.24}$$

□

B.1.3 Proof of Theorem 9

Proof. Rewriting the expression for the FDP in Definition 14, we obtain

$$\begin{aligned}
\mathrm{FDP}(v, \rho_{\mathrm{thr}}(u_c), T, L) &= \frac{V_{T,L}(v, \rho_{\mathrm{thr}}(u_c))}{R_{T,L}(v, \rho_{\mathrm{thr}}(u_c)) \vee 1} && \text{(B.25)}\\
&= \frac{\widehat{V}_{T,L}(v, \rho_{\mathrm{thr}}(u_c))}{R_{T,L}(v, \rho_{\mathrm{thr}}(u_c)) \vee 1} \cdot \frac{V_{T,L}(v, \rho_{\mathrm{thr}}(u_c))}{\widehat{V}_{T,L}(v, \rho_{\mathrm{thr}}(u_c))} && \text{(B.26)}\\
&\leq \alpha \cdot \frac{V_{T,L}(v, \rho_{\mathrm{thr}}(u_c))}{\widehat{V}_{T,L}(v, \rho_{\mathrm{thr}}(u_c))} && \text{(B.27)}\\
&\leq \alpha \cdot \frac{V_{T,L}(v, \rho_{\mathrm{thr}}(u_c))}{\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}}(u_c))} && \text{(B.28)}\\
&=: \alpha \cdot H_{T,L}(v, \rho_{\mathrm{thr}}(u_c)), && \text{(B.29)}
\end{aligned}$$

where the inequality in the third line follows from the condition in (4.33) that all considered quadruples $(T, L, \rho_{\mathrm{thr}}(u_c), v)$ must satisfy. Taking the expectation of the FDP, we obtain

$$\mathrm{FDR}(v, \rho_{\mathrm{thr}}(u_c), T, L) \leq \alpha \cdot \mathbb{E}\big[H_{T,L}(v, \rho_{\mathrm{thr}}(u_c))\big] \tag{B.30}$$

and, thus, it remains to prove that $\mathbb{E}[H_{T,L}(v, \rho_{\text{thr}}(u_c))] \leq 1$. Since (4.33) is a stopping time that is adapted to the filtration $\mathcal{F}_v$ in Lemma 7 (i.e., v is $\mathcal{F}_v$-measurable) and $H_{T,L}(v, \rho_{\text{thr}}(u_c))$ is bounded, we can apply Doob's optional stopping theorem (i.e., Theorem 1 in Section 2.3.2) to obtain an upper bound for $\mathbb{E}[H_{T,L}(v, \rho_{\text{thr}}(u_c))]$, i.e.,

$$\mathbb{E}\big[H_{T,L}(v, \rho_{\text{thr}}(u_c))\big] \leq \mathbb{E}\big[H_{T,L}(0.5, \rho_{\text{thr}}(u_c))\big]. \tag{B.31}$$

Defining

$$\Psi^{+}_{t,L}(j, \rho_{\text{thr}}(u_c)) \tag{B.32}$$

$$:= \begin{cases} \dfrac{1}{2 \;-\; \min\limits_{\substack{j' \in \\ \text{Gr}(j, \rho_{\text{thr}}(u_c))}} \big\{\big|\Phi_{T,L}(j) - \Phi_{T,L}(j')\big|\big\}}, & \text{Gr}(j, \rho_{\text{thr}}(u_c)) \\ & \neq \varnothing \\ 1, & \text{Gr}(j, \rho_{\text{thr}}(u_c)) \\ & = \varnothing \end{cases} \tag{B.33}$$

$$\geq \begin{cases} \dfrac{1}{2 \;-\; \min\limits_{\substack{j' \in \\ \text{Gr}(j, \rho_{\text{thr}}(u_c))}} \big\{\big|\Phi_{T,L}(j) - \Phi_{T,L}(j')\big|\big\}}, & \text{Gr}(j, \rho_{\text{thr}}(u_c)) \\ & \neq \varnothing \\ 1/2, & \text{Gr}(j, \rho_{\text{thr}}(u_c)) \\ & = \varnothing \end{cases}, \tag{B.34}$$

$$= \Psi_{t,L}(j, \rho_{\text{thr}}(u_c)), \tag{B.35}$$

$(t, L) \in \{1, \ldots, T\} \times \mathbb{N}_+$, $H_{T,L}(0.5, \rho_{\text{thr}}(u_c))$ can be upper bounded as follows:

$$H_{T,L}(0.5, \rho_{\text{thr}}(u_c)) = \frac{V_{T,L}(0.5, \rho_{\text{thr}}(u_c))}{\widehat{V}'_{T,L}(0.5, \rho_{\text{thr}}(u_c))} \tag{B.36}$$

$$\leq \frac{V^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c))}{\widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c))} \tag{B.37}$$

$$=: H^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c)). \tag{B.38}$$

The inequality in the second line follows from

$$\text{(i) } V^{+}_{T,L}(v, \rho_{\text{thr}}(u_c)) := \big|\{\text{null } j : \Psi^{+}_{T,L}(j, \rho_{\text{thr}}(u_c)) \cdot \Phi_{T,L}(j) > v\}\big| \tag{B.39}$$

$$\geq \left|\{\text{null } j : \Psi_{T,L}(j, \rho_{\text{thr}}(u_{\text{c}})) \cdot \Phi_{T,L}(j) > v\}\right| \tag{B.40}$$

$$= V_{T,L}(v, \rho_{\text{thr}}(u_{\text{c}})), \tag{B.41}$$

$$\text{(ii) } \widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}})) \coloneqq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Psi^{+}_{t,L}(q, \rho_{\text{thr}}(u_{\text{c}})) \cdot \Phi_{t,L}(q)}{L - (t-1)} \tag{B.42}$$

$$\leq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Psi_{t,L}(q, \rho_{\text{thr}}(u_{\text{c}})) \cdot \Phi_{t,L}(q)}{L - (t-1)} \tag{B.43}$$

$$= \widehat{V}'_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}})). \tag{B.44}$$

Next, we show that $H^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}}))$ is monotonically increasing in $\rho_{\text{thr}}(u_{\text{c}})$, i.e.,

$$H^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}}+1)) = \frac{V^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}}+1))}{\widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}}+1))} \tag{B.45}$$

$$\geq \frac{V^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}}))}{\widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}}))} \tag{B.46}$$

$$= H^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}})), \tag{B.47}$$

where the inequality in the second line follows from Lemmas 8 and 9. Combining these preliminaries and noting that $\Psi^{+}_{t,L}(j, \rho_{\text{thr}}(u_{\text{c}})) = 1$ yields

$$V_{T,L}(0.5) \coloneqq \left|\{\text{null } j : \Phi_{T,L}(j) > 0.5\}\right|, \tag{B.48}$$

$$\widehat{V}'_{T,L}(0.5) \coloneqq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi_{t,L}(q)}{L - (t-1)} \tag{B.49}$$

$$H_{T,L}(0.5) \coloneqq \frac{V_{T,L}(0.5)}{\widehat{V}'_{T,L}(0.5)} \tag{B.50}$$

i.e., the ordinary counterparts of the dependency-aware expressions $V^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_{\text{c}}))$,

$V'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c))$, and $H^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c))$, respectively, we finally obtain

$$\begin{aligned}
\mathbb{E}\big[H_{T,L}(v, \rho_{\text{thr}}(u_c))\big] &\leq \mathbb{E}\big[H_{T,L}(0.5, \rho_{\text{thr}}(u_c))\big] && \text{(B.51)}\\
&\leq \mathbb{E}\big[H^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c))\big] && \text{(B.52)}\\
&\leq \mathbb{E}\big[H^{+}_{T,L}(0.5, \rho_{\text{thr}}(p))\big] && \text{(B.53)}\\
&= \mathbb{E}\big[H^{+}_{T,L}(0.5, 1)\big] && \text{(B.54)}\\
&= \mathbb{E}\big[H_{T,L}(0.5)\big] && \text{(B.55)}\\
&\leq 1, && \text{(B.56)}
\end{aligned}$$

where the first inequality is the same as in (B.31), the second and third inequalities follow from (B.38) and (B.47), respectively, and the equation in the fourth line follows from (4.11) (i.e., $\rho_{\text{thr}}(p) = \sum_{u=1}^{p} \Delta\rho_{\text{thr},u} = 1$). The equation in the fifth line is a consequence of the following: For $\rho_{\text{thr}}(p) = 1$, we have $\text{Gr}(j, 1) = \varnothing$ for all $j \in \{1, \ldots, p\}$, and, therefore, it holds that $\Psi^{+}_{t,L}(j, 1) = 1$ for all $j \in \{1, \ldots, p\}$. Thus, $H^{+}_{T,L}(0.5, 1)$ boils down to its ordinary counterpart $H_{T,L}(0.5)$ in (B.50), i.e., $H^{+}_{T,L}(0.5, 1) = H_{T,L}(0.5)$. Finally, the proof of Inequality (B.56) is omitted because it is exactly as in the proof of Theorem 3 (FDR control - *T-Rex* selector). □

B.1.4 Proof of Theorem 10

Proof. The FDR control property in Theorem 9 holds if Lemmas 8 and 9 hold. Lemmas 8 and 9 hold for any definition of $\text{Gr}(j, \rho_{\text{thr}})$ that satisfies the group design principle in Theorem 10. □

B.1.5 Proof of Theorem 11

Proof. From Equation (4.24), it follows that for all quadruples $(v, \rho_{\text{thr}}(u_c), T, L)$ that satisfy $\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \leq \alpha$, the objective functions in Step 6 of Algorithm 5 and in the optimization problem in (4.32) are equal, i.e., $|\widehat{\mathcal{A}}_L(v', \rho_{\text{thr}}(u'_c), T')| = R_{T',L}(v', \rho_{\text{thr}}(u'_c))$. Therefore, and since all attainable values of u_c are considered in Step 6 of Algorithm 5, it suffices to show that for fixed $\rho_{\text{thr}}(u_c)$ and L, the set of feasible tuples (v, T) of (4.32) is a subset of or equal to the set of feasible tuples obtained by Algorithm 5. Since, ceteris paribus, $\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L)$ is monotonically decreasing in v and monotonically increasing in T, for $T = T_{\text{fin}}$, $\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L)$ attains its minimum value at $(v, T) = (1 - \Delta v, T_{\text{fin}})$,

where $T_{\text{fin}} \in \{1, \dots, L\}$ is implicitly defined through the inequalities

$$\widehat{\text{FDP}}(1 - \Delta v, \rho_{\text{thr}}(u_c), T_{\text{fin}}, L) \leq \alpha \tag{B.57}$$

and

$$\widehat{\text{FDP}}(1 - \Delta v, \rho_{\text{thr}}(u_c), T_{\text{fin}} + 1, L) > \alpha. \tag{B.58}$$

Thus, the feasible set of the optimization problem in (4.32) is given by

$$\big\{(v,T) : \widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\big\} \tag{B.59}$$

$$= \{(v, T) : v \in [0.5, 1 - \Delta v], \tag{B.60}$$

$$T \in \{1, \dots, T_{\text{fin}}\}, \tag{B.61}$$

$$\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\big\}. \tag{B.62}$$

Since $\Delta v = 1/K$, the upper endpoint of the interval $[0.5, 1 - \Delta v]$ asymptotically (i.e., $K \to \infty$) coincides with the supremum of the interval $[0.5, 1)$. That is, the set in (B.62) contains all feasible solutions of (4.32). However, since the v-grid in Algorithm 5 is, as in Algorithms 2 and 3, adapted to K, all values of $R_{T,L}(v, \rho_{\text{thr}}(u_c))$ that are attained by off-grid solutions can also be attained by on-grid solutions. Thus, instead of (B.62) only the following fully discrete feasible set of (4.32) needs to be considered:

$$\big\{(v, T) : v \in \{0.5, 0.5 + \Delta v, 0.5 + 2\Delta v, \dots, 1 - \Delta v\}, \tag{B.63}$$

$$T \in \{1, \dots, T_{\text{fin}}\}, \tag{B.64}$$

$$\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\big\}. \tag{B.65}$$

Since the "while"-loop in Step 5 of Algorithm 5 is terminated when $T = T_{\text{fin}} + 1$, the feasible set of the optimization problem in Step 6 of Algorithm 5 is given by

$$\big\{(v, T) : v \in \{0.5, 0.5 + \Delta v, 0.5 + 2\Delta v, \dots, 1 - \Delta v\}, \tag{B.66}$$

$$T \in \{1, \dots, T_{\text{fin}}\} \tag{B.67}$$

$$\widehat{\text{FDP}}(v, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\big\}, \tag{B.68}$$

which is equal to (B.65). □

B.1.6 Proof of Theorem 12

Proof. An upper bound on the FDR is given by

$$\begin{aligned}
\mathrm{FDP}(v, \rho_{\mathrm{thr}}, T, L) &= \frac{V_{T,L}(v, \rho_{\mathrm{thr}})}{\max\left\{1, R_{T,L}(v, \rho_{\mathrm{thr}})\right\}} && \text{(B.69)}\\
&= \widehat{\mathrm{FDP}}(v, \rho_{\mathrm{thr}}, T, L) \cdot \frac{V_{T,L}(v, \rho_{\mathrm{thr}})}{\widehat{V}_{T,L}(v, \rho_{\mathrm{thr}})} && \text{(B.70)}\\
&\leq \alpha \cdot \frac{V_{T,L}(v, \rho_{\mathrm{thr}})}{\widehat{V}_{T,L}(v, \rho_{\mathrm{thr}})}, && \text{(B.71)}\\
&\leq \alpha \cdot \frac{V_{T,L}(v, \rho_{\mathrm{thr}})}{\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})}, && \text{(B.72)}
\end{aligned}$$

where the second, third, and fourth line follow from (4.42), (4.45), and (4.43), respectively. Taking the expected value yields

$$\mathrm{FDR}(v, \rho_{\mathrm{thr}}, T, L) \leq \alpha \cdot \mathbb{E}\left[\frac{V_{T,L}(v, \rho_{\mathrm{thr}})}{\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})}\right]. \tag{B.73}$$

It remains to prove that $\mathbb{E}[V_{T,L}(v, \rho_{\mathrm{thr}})/\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})] \leq 1$. Along the lines of the proof of Lemma 7, it can be shown that $V_{T,L}(v, \rho_{\mathrm{thr}})/\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})$ is a backward-running supermartingale with respect to the filtration

$$\mathcal{F}_v \coloneqq \sigma(\{R_{T,L}(v, \rho_{\mathrm{thr}})\}_{u \geq v}, \{V_{T,L}(v, \rho_{\mathrm{thr}})\}_{u \geq v}, \{\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})\}_{u \geq v}). \tag{B.74}$$

Therefore, and since v in (4.45) is $\mathcal{F}_v$-measurable and $V_{T,L}(v, \rho_{\mathrm{thr}})/\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})$ is bounded, an upper bound on

$$\mathbb{E}\left[\frac{V_{T,L}(v, \rho_{\mathrm{thr}})}{\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})}\right] \tag{B.75}$$

can be derived using Doob's optional stopping theorem (i.e., Theorem 1 in Section 2.3.2). First, using the definition of $V_{T,L}(v)$ in Definition 6 and defining $\widehat{V}'_{T,L}(0.5) \coloneqq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi_{t,L}(q)}{L-(t-1)}$, we obtain

$$\mathbb{E}\left[\frac{V_{T,L}(v, \rho_{\mathrm{thr}})}{\widehat{V}'_{T,L}(v, \rho_{\mathrm{thr}})}\right] \leq \mathbb{E}\left[\frac{V_{T,L}(0.5, \rho_{\mathrm{thr}})}{\widehat{V}'_{T,L}(0.5, \rho_{\mathrm{thr}})}\right] \tag{B.76}$$

$$\leq \mathbb{E}\left[\frac{V_{T,L}(0.5)}{\widehat{V}'_{T,L}(0.5)}\right] \tag{B.77}$$

$$\leq 1. \tag{B.78}$$

Inequality (B.76) follows from Doob's optional stopping theorem. Inequality (B.77) follows from the following two inequalities:

$$\text{(i) } V_{T,L}(0.5, \rho_{\text{thr}}) = \left|\{\text{null } j : \Phi^{\text{NN}}_{T,L}(j, \rho_{\text{thr}}) > 0.5\}\right| \tag{B.79}$$

$$\leq \left|\{\text{null } j : \Phi_{T,L}(j) > 0.5\}\right| = V_{T,L}(0.5), \tag{B.80}$$

$$\text{(ii) } \widehat{V}'_{T,L}(0.5, \rho_{\text{thr}}) = \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi^{\text{NN}}_{t,L}(q, \rho_{\text{thr}})}{L - (t-1)} \tag{B.81}$$

$$\geq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Phi_{t,L}(q)}{L - (t-1)} = \widehat{V}'_{T,L}(0.5). \tag{B.82}$$

where the inequalities in (i) and (ii) both follow from Equation (4.41) and the fact that $\Psi^{\text{NN}}_{T,L}(j, \rho_{\text{thr}}) \leq 1$. Finally, inequality (B.78) has already been proven to hold within the proof of Theorem 3 (FDR control - *T-Rex* selector) and, thus, it holds that $\text{FDR}(v, \rho_{\text{thr}}, T, L) \leq \alpha$. □

B.1.7 Technical Lemmas

Lemma 7. *Define* $\mathcal{V} \coloneqq \{\Phi^{\text{DA}}_{T,L}(j, \rho_{\text{thr}}(u_{\text{c}})) \geq 0.5,\ j = 1, \ldots, p\} \backslash \{1\}$. *Let*

$$\mathcal{F}_v \coloneqq \sigma\Big(\big\{R_{T,L}(u, \rho_{\text{thr}}(u_{\text{c}}))\big\}_{u \geq v}, \big\{V_{T,L}(u, \rho_{\text{thr}}(u_{\text{c}}))\big\}_{u \geq v}, \big\{\widehat{V}'_{T,L}(u, \rho_{\text{thr}}(u_{\text{c}}))\big\}_{u \geq v}\Big) \tag{B.83}$$

be a backward-filtration with respect to v. *Then, for all triples* $(T, L, \rho_{\text{thr}}(u_{\text{c}})) \in \{1, \ldots, L\} \times \mathbb{N}_+ \times [0, 1]$, $\{H_{T,L}(v, \rho_{\text{thr}}(u_{\text{c}}))\}_{v \in \mathcal{V}}$ *is a backward-running super-martingale with respect to* $\mathcal{F}_v$. *That is,*

$$\mathbb{E}\big[H_{T,L}(v - \epsilon^*_{T,L}(v, \rho_{\text{thr}}(u_{\text{c}}))) \mid \mathcal{F}_v\big] \geq H_{T,L}(v, \rho_{\text{thr}}(u_{\text{c}})), \tag{B.84}$$

where $v \in [0.5, 1)$ *and*

$$\epsilon^*_{T,L}(v, \rho_{\text{thr}}(u_c)) \coloneqq \inf\left\{\epsilon \in (0, v) : R_{T,L}(v - \epsilon, \rho_{\text{thr}}(u_c)) - R_{T,L}(v, \rho_{\text{thr}}(u_c)) = 1\right\} \tag{B.85}$$

with the convention that $\epsilon^*_{T,L}(v, \rho_{\text{thr}}(u_c)) = 0$ *if the infimum does not exist.*

Proof. The proof of Lemma 7 follows along the lines of the proof of Lemma 1 and by replacing $\Phi_{t,L}(j)$, $V_{T,L}(v)$, and $R_{T,L}(v)$ by their dependency-aware extensions $\Phi^{\text{DA}}_{t,L}(j, \rho_{\text{thr}}(u_c))$, $V_{T,L}(v, \rho_{\text{thr}}(u_c))$, and $R_{T,L}(v, \rho_{\text{thr}}(u_c))$, respectively. □

Lemma 8. *Let* $V^+_{T,L}(v, \rho_{\text{thr}}(u_c))$ *be as in* (B.39). *For all triples* $(T, L, v) \in \{1, \dots, L\} \times \mathbb{N}_+ \times [0.5, 1)$, $V^+_{T,L}(v, \rho_{\text{thr}}(u_c))$ *is monotonically increasing in* $\rho_{\text{thr}}(u_c)$*, i.e., for any* $u_c \in \{1, \dots, p-1\}$*, it holds that*

$$V^+_{T,L}(v, \rho_{\text{thr}}(u_c + 1)) \geq V^+_{T,L}(v, \rho_{\text{thr}}(u_c)). \tag{B.86}$$

Proof. Using the definition of $V^+_{T,L}(v, \rho_{\text{thr}}(u_c))$ in (B.39), we obtain

$$\begin{aligned}
&V^+_{T,L}(v, \rho_{\text{thr}}(u_c + 1)) && \text{(B.87)}\\
&\quad = \left|\left\{\text{null } j : \Psi^+_{T,L}(j, \rho_{\text{thr}}(u_c + 1)) \cdot \Phi_{T,L}(j) > v\right\}\right| && \text{(B.88)}\\
&\quad \geq \left|\left\{\text{null } j : \Psi^+_{T,L}(j, \rho_{\text{thr}}(u_c)) \cdot \Phi_{T,L}(j) > v\right\}\right| && \text{(B.89)}\\
&\quad = V^+_{T,L}(v, \rho_{\text{thr}}(u_c)). && \text{(B.90)}
\end{aligned}$$

The inequality in the third line follows from

$$\Psi^+_{T,L}(j, \rho_{\text{thr}}(u_c + 1)) \geq \Psi^+_{T,L}(j, \rho_{\text{thr}}(u_c)), \tag{B.91}$$

which is a consequence of the following two cases:

$$\begin{aligned}
&\text{(i) } \text{Gr}(j, \rho_{\text{thr}}(u_c)) = \varnothing : && \text{(B.92)}\\
&\qquad \Psi^+_{T,L}(j, \rho_{\text{thr}}(u_c + 1)) = 1 = \Psi^+_{T,L}(j, \rho_{\text{thr}}(u_c)), && \text{(B.93)}\\
&\text{(ii) } \text{Gr}(j, \rho_{\text{thr}}(u_c)) \neq \varnothing : && \text{(B.94)}\\
&\qquad \Psi^+_{T,L}(j, \rho_{\text{thr}}(u_c + 1)) && \text{(B.95)}
\end{aligned}$$

$$= \left[2 - \min_{j' \in \mathrm{Gr}(j, \rho_{\mathrm{thr}}(u_{\mathrm{c}}+1))} \left\{ \left| \Phi_{t,L}(j) - \Phi_{t,L}(j') \right| \right\} \right]^{-1} \tag{B.96}$$

$$\geq \left[2 - \min_{j' \in \mathrm{Gr}(j, \rho_{\mathrm{thr}}(u_{\mathrm{c}}))} \left\{ \left| \Phi_{t,L}(j) - \Phi_{t,L}(j') \right| \right\} \right]^{-1} \tag{B.97}$$

$$= \Psi^{+}_{T,L}(j, \rho_{\mathrm{thr}}(u_{\mathrm{c}})). \tag{B.98}$$

In (ii), the inequality in the third line follows from the fact that, for any $u_{\mathrm{c}} \in \{1, \ldots, p-1\}$, it holds that

$$\mathrm{Gr}(j, \rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \subseteq \mathrm{Gr}(j, \rho_{\mathrm{thr}}(u_{\mathrm{c}})), \quad j = 1, \ldots, p, \tag{B.99}$$

and, therefore,

$$\min_{j' \in \mathrm{Gr}(j, \rho_{\mathrm{thr}}(u_{\mathrm{c}}+1))} \left\{ \left| \Phi_{t,L}(j) - \Phi_{t,L}(j') \right| \right\} \tag{B.100}$$

$$\geq \min_{j' \in \mathrm{Gr}(j, \rho_{\mathrm{thr}}(u_{\mathrm{c}}))} \left\{ \left| \Phi_{t,L}(j) - \Phi_{t,L}(j') \right| \right\}. \tag{B.101}$$

□

Lemma 9. *Let* $\widehat{V}'^{+}_{T,L}(0.5, \rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ *be as in* (B.42)*. For all tuples* $(T, L) \in \{1, \ldots, L\} \times \mathbb{N}_{+}$, $\widehat{V}'^{+}_{T,L}(0.5, \rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ *is monotonically decreasing in* $\rho_{\mathrm{thr}}(u_{\mathrm{c}})$*, i.e., for any* $u_{\mathrm{c}} \in \{1, \ldots, p-1\}$*, it holds that*

$$\widehat{V}'^{+}_{T,L}(0.5, \rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \leq \widehat{V}'^{+}_{T,L}(0.5, \rho_{\mathrm{thr}}(u_{\mathrm{c}})). \tag{B.102}$$

Proof. Using the definition of $\widehat{V}'^{+}_{T,L}(0.5, \rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ in (B.42), we obtain

$$\widehat{V}'^{+}_{T,L}(0.5, \rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \tag{B.103}$$

$$= \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Psi^{+}_{t,L}(q, \rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \cdot \Phi_{t,L}(q)}{L - (t-1)} \tag{B.104}$$

$$\leq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Psi^{+}_{t,L}(q, \rho_{\mathrm{thr}}(u_{\mathrm{c}})) \cdot \Phi_{t,L}(q)}{L - (t-1)} \tag{B.105}$$

$$= \widehat{V}'^{+}_{T,L}(0.5, \rho_{\mathrm{thr}}(u_{\mathrm{c}})), \tag{B.106}$$

where the inequality in the third line follows from

$$\Psi^{+}_{t,L}(q, \rho_{\text{thr}}(u_{\text{c}} + 1)) \geq \Psi^{+}_{t,L}(q, \rho_{\text{thr}}(u_{\text{c}})), \tag{B.107}$$

which was shown to hold within the proof of Lemma 8. □

B.2 Additional Simulations

In this appendix, we consider two additional heavy-tailed simulation settings to complement the numerical experiments presented in Section 4.3. These settings differ from the simulation setting in Section 4.3 as follows:

1. Heavy-tailed predictor matrix $\boldsymbol{X}$: The p-dimensional samples of the predictor matrix (i.e., rows of $\boldsymbol{X}$) are sampled from a zero-mean multivariate heavy-tailed Student-t distribution with covariance matrix $\boldsymbol{\Sigma}$ (with $\rho = 0.8$) and 3 degrees of freedom,
2. Heavy-tailed noise vector $\boldsymbol{\epsilon}$: The noise vector $\boldsymbol{\epsilon}$ is sampled from a heavy-tailed Student-t distribution with 3 degrees of freedom.

The additional simulation results in Figures B.1, B.2, B.3, and B.4 verify the theoretical results in Section 4.2 and show that only the proposed *T-Rex+DA* selector reliably controls the FDR in these heavy-tailed settings.

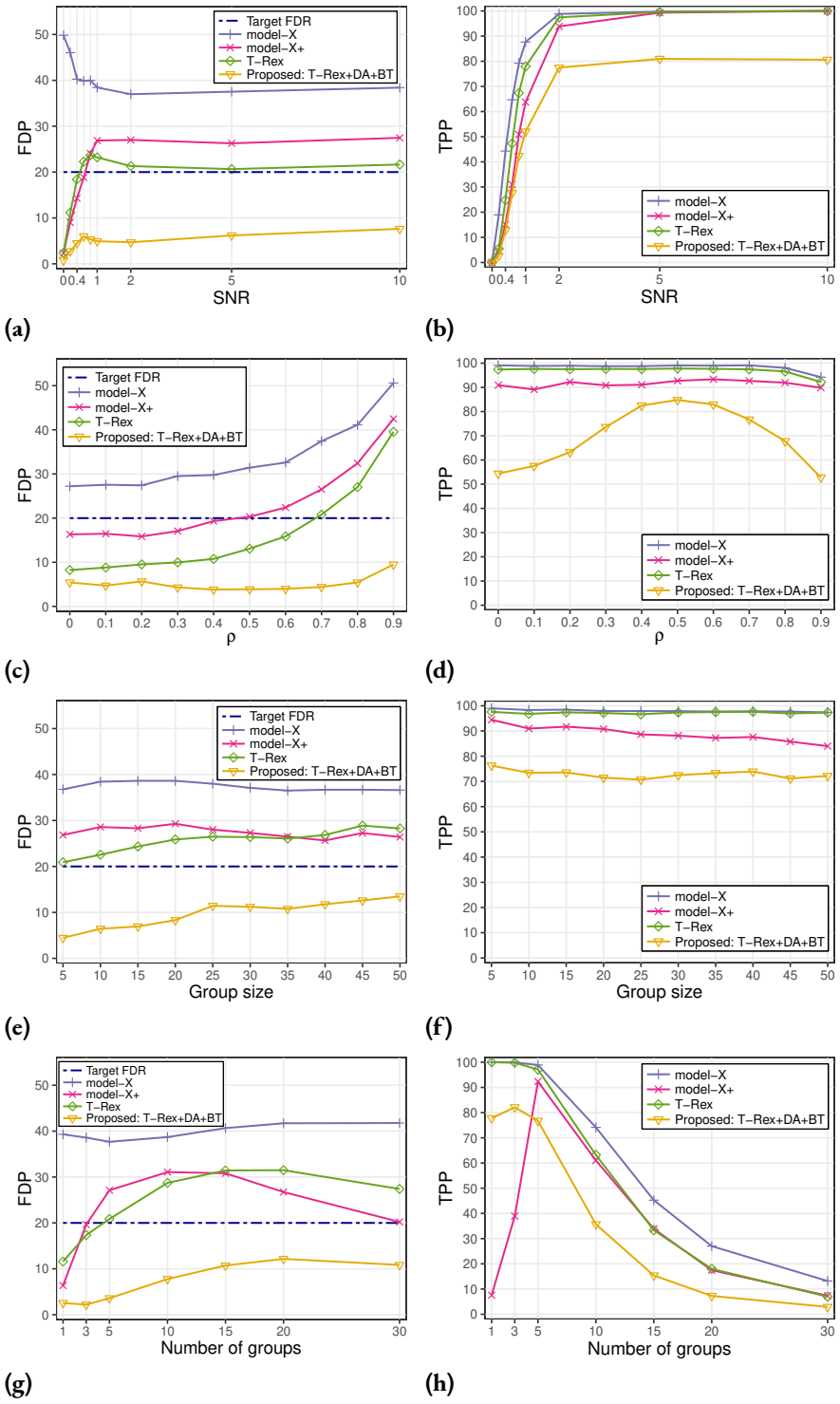


Figure B.1: Heavy-tailed predictor matrix $\boldsymbol{X}$.

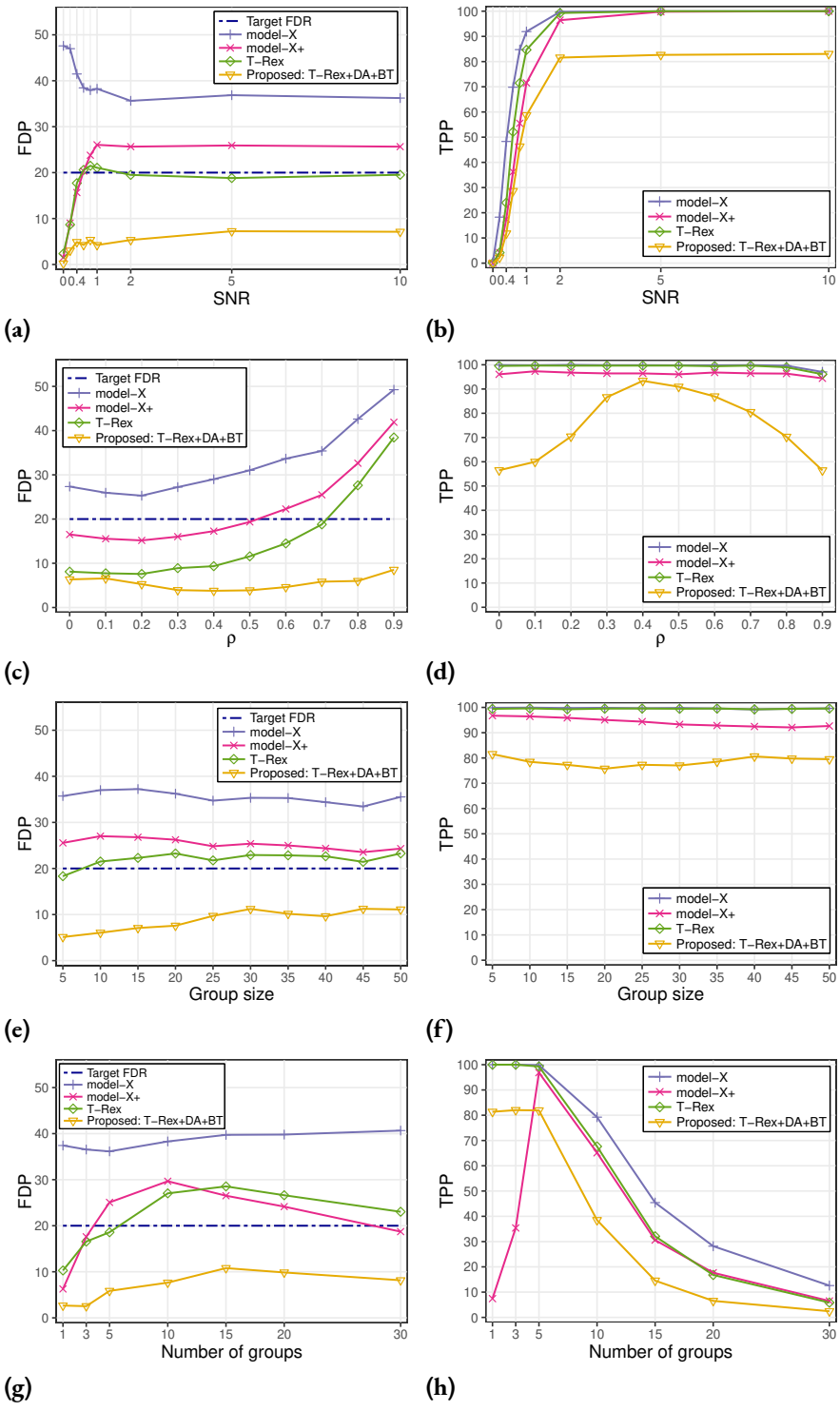
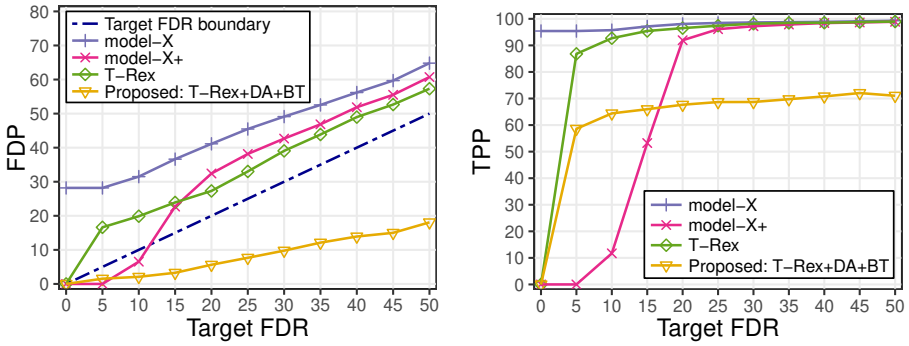
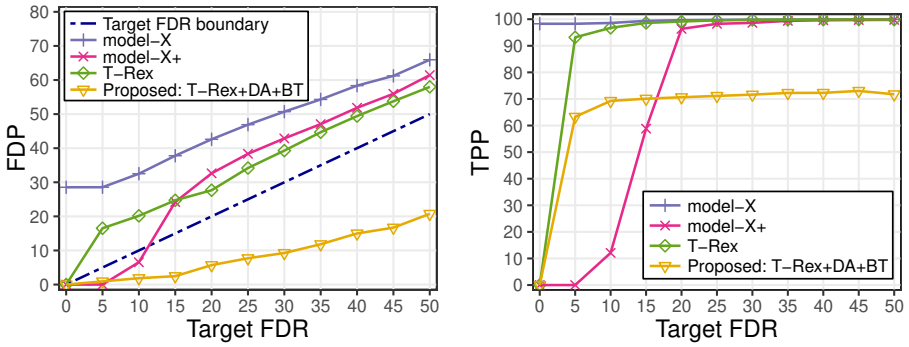


Figure B.2: Heavy-tailed noise vector ϵ.



(a) (b)

Figure B.3: Heavy-tailed predictor matrix $\boldsymbol{X}$.



(a) (b)

Figure B.4: Heavy-tailed noise vector $\boldsymbol{\epsilon}$.

C

Appendix to Chapter 5

This appendix to Chapter 5 contains the proofs of Theorem 14 (Appendix C.1), Theorem 15 (Appendix C.2), and Theorem 16 (Appendix C.3).

C.1 Proof of Theorem 14

Proof. First, note that, without loss of generality, $\mathcal{G}_m$ in (5.1) can be reduced to $\mathcal{G}_m = \{g \in \{1,\dots,p\} : \rho_{g,m} = 1\}, m = 1,\dots,p$, since x_g or x_m can be replaced by $-x_g$ or $-x_m$, respectively. The variable selection process in all K random experiments is not affected by such a replacement, because only the sign of the associated coefficient estimate is flipped.

Second, note that the relative occurrences within the selected active set in (3.3) are defined by

$$\Phi_{T,L}(j) := \begin{cases} \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_k(j,T,L), & T \geq 1 \\ 0, & T = 0 \end{cases}, \tag{C.1}$$

where the indicator function in the first case is given by

$$\mathbb{1}_k(j,T,L) = \begin{cases} 1, & j \in \mathcal{C}_{k,L}(T) \\ 0, & \text{otherwise} \end{cases}, \tag{C.2}$$

i.e., it is one if the jth variable is included in the candidate set of the kth random experiment and zero otherwise (see Definition 5).

Third, note that Lemma 2 (a) in [ZH05] states that for any strictly convex penalty function $f(\boldsymbol{\beta})$ in

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda f(\boldsymbol{\beta}), \tag{C.3}$$

it holds that if $\boldsymbol{x}_g = \boldsymbol{x}_m$, then $\hat{\beta}_g = \hat{\beta}_m$, $g, m \in \{1, \ldots, p\}$, for all $\lambda > 0$. Since the *elastic net* penalty in (2.20) is a strictly convex function of $\boldsymbol{\beta}$, we conclude that $\rho_{g,m} = 1$ implies $\hat{\beta}_g = \hat{\beta}_m$, $g, m \in \{1, \ldots, p\}$, for all $\lambda > 0$. From $\hat{\beta}_{g,k} = \hat{\beta}_{m,k}$ for all $\lambda > 0$, where $\hat{\beta}_{g,k}$ and $\hat{\beta}_{m,k}$ are the coefficient estimates of variables $\boldsymbol{x}_g$ and $\boldsymbol{x}_m$ corresponding to the kth random experiment, it follows that

$$\mathbb{1}_k(g, T, L) = \mathbb{1}_k(m, T, L) \tag{C.4}$$

for all $k \in \{1, \ldots, K\}$ and all tuples $(T, L) \in \{1, \ldots, L\} \times \mathbb{N}_+$. Consequently, $\rho_{j,m} = 1$ implies $\Phi_{T,L}(j) = \Phi_{T,L}(m)$ for all $j \in \mathcal{G}_m$. Thus, for all triples $(v, T, L) \in [0.5, 1) \times \{1, \ldots, L\} \times \mathbb{N}_+$ and for all $j, m \in \{1, \ldots, p\}$, the following two statements hold:

(a1) If $\rho_{j,m} = 1$ and $\Phi_{T,L}(j) > v$, then $\Phi_{T,L}(m) > v$.

(b1) If $\rho_{j,m} = 1$ and $\Phi_{T,L}(j) \leq v$, then $\Phi_{T,L}(m) \leq v$.

Using the definition of $\widehat{\mathcal{A}}_L(v, T)$ in (3.3), Statements (a1) and (b1) can be translated into the following equivalent statements that hold for all triples $(v, T, L) \in [0.5, 1) \times \{1, \ldots, L\} \times \mathbb{N}_+$ and for all $j, m \in \{1, \ldots, p\}$:

(a2) If $j \in \mathcal{G}_m$ and $j \in \widehat{\mathcal{A}}_L(v, T)$, then $G_m \subseteq \widehat{\mathcal{A}}_L(v, T)$.

(b2) If $j \in \mathcal{G}_m$ and $j \notin \widehat{\mathcal{A}}_L(v, T)$, then $G_m \cap \widehat{\mathcal{A}}_L(v, T) = \varnothing$.

Statements (a2) and (b2) are equivalent to Statements (i) and (ii) in the theorem. □

C.2 Proof of Theorem 15

Proof. First, we rewrite (5.8) as follows:

$$\mathcal{L}_{\text{IEN}} = \left(\boldsymbol{y}'^{\top}\boldsymbol{y}' - 2\boldsymbol{\beta}^{\top}\boldsymbol{X}'^{\top}\boldsymbol{y}' + \boldsymbol{\beta}^{\top}\boldsymbol{X}'^{\top}\boldsymbol{X}'\boldsymbol{\beta}\right) + \lambda_1\|\boldsymbol{\beta}\|_1. \tag{C.5}$$

Second, note that $\boldsymbol{X}'^\top \boldsymbol{X}' = \boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \sum_{m=1}^{M} \frac{\mathbf{1}_m \mathbf{1}_m^\top}{p_m}$. Then, plugging (5.7) into (C.5) yields

$$\begin{aligned}
\mathcal{L}_{\text{IEN}} = \Bigg[& \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} && \text{(C.6)} \\
& + \boldsymbol{\beta}^\top \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \sum_{m=1}^{M} \frac{\mathbf{1}_m \mathbf{1}_m^\top}{p_m} \right) \boldsymbol{\beta} \Bigg] + \lambda_1 \|\boldsymbol{\beta}\|_1 && \text{(C.7)} \\
= \; & \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta} && \text{(C.8)} \\
& + \lambda_2 \sum_{m=1}^{M} \frac{\boldsymbol{\beta}^\top \mathbf{1}_m \mathbf{1}_m^\top \boldsymbol{\beta}}{p_m} + \lambda_1 \|\boldsymbol{\beta}\|_1 && \text{(C.9)} \\
= \; & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{m=1}^{M} \frac{(\mathbf{1}_m^\top \boldsymbol{\beta})^2}{p_m}. && \text{(C.10)}
\end{aligned}$$

□

C.3 Proof of Theorem 16

Proof. Define $\hat{\boldsymbol{r}} := \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Taking the first derivative of (5.5) and setting it equal to zero, we obtain

$$\begin{aligned}
& \frac{\partial \mathcal{L}_{\text{IEN}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} && \text{(C.11)} \\
& = -2\boldsymbol{X}^\top \hat{\boldsymbol{r}} + \lambda_1 \frac{\partial \|\boldsymbol{\beta}\|_1}{\partial \boldsymbol{\beta}} \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} + 2\lambda_2 \sum_{m=1}^{M} \frac{\mathbf{1}_m \mathbf{1}_m^\top \hat{\boldsymbol{\beta}}}{p_m} \stackrel{!}{=} \mathbf{0}. && \text{(C.12)}
\end{aligned}$$

The jth and j'th equation of the system of equations in (C.12) are:

$$\begin{aligned}
& -2\boldsymbol{x}_j^\top \hat{\boldsymbol{r}} + \lambda_1 \operatorname{sign}(\hat{\beta}_j) + 2\lambda_2 \frac{\mathbf{1}_1^\top \hat{\boldsymbol{\beta}}}{p_1} = 0, && \text{(C.13)} \\
& -2\boldsymbol{x}_{j'}^\top \hat{\boldsymbol{r}} + \lambda_1 \operatorname{sign}(\hat{\beta}_{j'}) + 2\lambda_2 \frac{\mathbf{1}_2^\top \hat{\boldsymbol{\beta}}}{p_2} = 0. && \text{(C.14)}
\end{aligned}$$

Subtracting (C.14) from (C.13) and noting that $\operatorname{sign}(\hat{\beta}_j) = \operatorname{sign}(\hat{\beta}_{j'})$:

$$2(\boldsymbol{x}_{j'}^\top - \boldsymbol{x}_j^\top)\hat{\boldsymbol{r}} + 2\lambda_2 \left(\frac{\mathbf{1}_1^\top \hat{\boldsymbol{\beta}}}{p_1} - \frac{\mathbf{1}_2^\top \hat{\boldsymbol{\beta}}}{p_2} \right) = 0. \qquad \text{(C.15)}$$

Using (C.15), we obtain

$$\frac{1}{\|\boldsymbol{y}\|_2}\left|\frac{\mathbf{1}_1^\top\hat{\boldsymbol{\beta}}}{p_1}-\frac{\mathbf{1}_2^\top\hat{\boldsymbol{\beta}}}{p_2}\right| = \frac{1}{\lambda_2\|\boldsymbol{y}\|_2}\left|(\boldsymbol{x}_j^\top-\boldsymbol{x}_{j'}^\top)\hat{\boldsymbol{r}}\right| \tag{C.16}$$

$$\leq \frac{1}{\lambda_2}\|\boldsymbol{x}_j-\boldsymbol{x}_{j'}\|_2\frac{\|\hat{\boldsymbol{r}}\|_2}{\|\boldsymbol{y}\|_2} = \frac{1}{\lambda_2}\sqrt{2(1-\rho_{j,j'})}\frac{\|\hat{\boldsymbol{r}}\|_2}{\|\boldsymbol{y}\|_2} \tag{C.17}$$

$$\leq \frac{1}{\lambda_2}\sqrt{2(1-\rho_{j,j'})}, \tag{C.18}$$

where the inequality and the equation in the second line follow from the Cauchy-Schwarz inequality and from $\|\boldsymbol{x}_j-\boldsymbol{x}_{j'}\|_2 = \sqrt{\|\boldsymbol{x}_j\|_2^2+\|\boldsymbol{x}_{j'}\|_2^2-2\boldsymbol{x}_j^\top\boldsymbol{x}_{j'}} = \sqrt{1+1-2\rho_{j,j'}}$, respectively. The last line follows from the fact that $\hat{\boldsymbol{\beta}}$ is the minimizer of (5.5) and, therefore, $\mathcal{L}(\hat{\boldsymbol{\beta}}) \leq \mathcal{L}(\mathbf{0}) \Leftrightarrow \|\hat{\boldsymbol{r}}\|_2^2+\lambda_1\|\hat{\boldsymbol{\beta}}\|_1+\lambda_2\sum_{m=1}^M\frac{(\mathbf{1}_m^\top\hat{\boldsymbol{\beta}})^2}{p_m} \leq \|\boldsymbol{y}\|_2^2 \Rightarrow \|\hat{\boldsymbol{r}}\|_2 \leq \|\boldsymbol{y}\|_2$. Since the inequality in (C.18) holds for all $j \in \mathcal{G}_1$ and $j' \in \mathcal{G}_2$, the smallest upper bound is given by the largest $\rho_{j,j'}$, i.e.,

$$\frac{1}{\|\boldsymbol{y}\|_2}\left|\frac{\mathbf{1}_1^\top\hat{\boldsymbol{\beta}}}{p_1}-\frac{\mathbf{1}_2^\top\hat{\boldsymbol{\beta}}}{p_2}\right| \leq \frac{1}{\lambda_2}\sqrt{2\left(1-\max_{j\in\mathcal{G}_1,j'\in\mathcal{G}_2}\{\rho_{j,j'}\}\right)}. \tag{C.19}$$

□

List of Acronyms & Abbreviations

a.s.	Almost sure
AI	Artificial intelligence
AIC	Akaike information criterion
asdp	Approximate semidefinite program
BH	Benjamini-Hochberg
BIC	Bayesian information criterion
BY	Benjamini-Yekutieli
CDF	Cumulative distribution function
CPU	Central processing unit
CRAN	Comprehensive R Archive Network
CRT	Conditional randomization test
CV	Cross-validation
CVE	Cross-validation error
dB	Decibel
DOA	Direction-of-arrival
EN	Elastic net
ETF	Exchange traded fund
EU	European Union
EV	Explained variance
FDP	False discovery proportion
FDR	False discovery rate
FWER	Family-wise error rate
GWAS	Genome-wide association study
HIV-1	Human immunodeficiency virus type 1
i.i.d.	Independent and identically distributed
IC	Information criterion

IEN	Informed elastic net
LARS	Least angle regression
Lasso	Least absolute shrinkage and selection operator
MSTE	Mean squared wealth tracking error
NHG	Negative hypergeometric distribution
NN	Nearest neighbors
NP	Non-deterministic polynomial time
OLS	Ordinary least squares
PC	Principal component
PCA	Principal component analysis
PEV	Percentage of explained variance
PFER	Per family error rate
PI	Protease inhibitor
PMF	Probability mass function
Q-Q	Quantile-quantile
QP	Quadratic program
RAM	Random-access memory
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing
S&P 500	Standard & Poor's 500
Screen-T-Rex	Screening-T-Rex
SNP	Single nucleotide polymorphism
SNR	Signal-to-noise ratio
SPCA	Sparse PCA
SSD	Solid-state drive
SSR	Sum of squared residuals
SVD	Singular value decomposition
T-LARS	Terminating-LARS
T-Rex	Terminating-random experiments
T-Rex+DA	Dependency-aware T-Rex
T-Rex+DA+AR1	T-Rex+DA for autoregressive dependency models

T-Rex+DA+BT	T-Rex+DA for binary tree dependency models
T-Rex+DA+NN	T-Rex+DA for nearest neighbors dependency models
T-Rex+GVS	Grouped variable selection T-Rex
TCGA	The Cancer Genome Atlas
TPP	True positive proportion
TPR	True positive rate
TSM	Treatment-selected mutation

List of Operators & Symbols

The following list contains the most frequently used mathematical operators and symbols in this dissertation. All other expressions are defined on first use.

Operators

$\lvert a \rvert$	Absolute value of a scalar a
$\lvert \mathcal{A} \rvert$	Cardinality of a set $\mathcal{A}$
$\lVert \boldsymbol{a} \rVert_1$	ℓ_1-norm of a vector $\boldsymbol{a}$
$\lVert \boldsymbol{a} \rVert_2$	ℓ_2-norm of a vector $\boldsymbol{a}$
$\mathbb{E}[A]$	Expected value of a random variable A
$\mathrm{Var}[A]$	Variance of a random variable A
$\boldsymbol{a}^\top$	Transpose of a vector $\boldsymbol{a}$
$\boldsymbol{A}^\top$	Transpose of a matrix $\boldsymbol{A}$
$\mathrm{rank}(\boldsymbol{A})$	Rank of a matrix $\boldsymbol{A}$
$\mathrm{tr}(\boldsymbol{A})$	Trace of a matrix $\boldsymbol{A}$
$\mathcal{A} \cap \mathcal{B}$	Intersection of the sets $\mathcal{A}$ and $\mathcal{B}$
$\mathcal{A} \cup \mathcal{B}$	Union of the sets $\mathcal{A}$ and $\mathcal{B}$
$a \vee b$	Maximum of two scalars a and b
$\mathrm{sign}(a)$	Sign of a scalar a
$\bigcup_{z=1}^{Z} \mathcal{A}_z$	Union of the sets $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_Z$
$\max_{v,T} f(v,T)$ s.t. $g(v,T) \leq \alpha$	Optimization problem that returns the maximum value of the function $f(v,T)$ with respect to the parameters v and T and subject to the constraint $g(v,T) \leq \alpha$
$\arg\min_{\boldsymbol{\beta}} h(\boldsymbol{\beta})$	Argument of the minimum of the function $h(\boldsymbol{\beta})$

Symbols

n	Number of data points

p_0	Number of null variables
p_1	Number of true active variables
p	Total number of variables
$\boldsymbol{y}$	Response vector
$\boldsymbol{X}$	Original predictor matrix
$\mathring{\boldsymbol{X}}_k$	Dummy predictor matrix of the kth random experiment of the *T-Rex* selector
$\widetilde{\boldsymbol{X}}_k$	Enlarged predictor matrix of the kth random experiment of the *T-Rex* selector (contains original predictors and dummies)
$\boldsymbol{x}_j$	jth original predictor
$\mathring{\boldsymbol{x}}_{k,l}$	lth dummy predictor within the kth random experiment of the *T-Rex* selector
$\boldsymbol{\beta}$	Coefficient vector
$\boldsymbol{w}$	Asset weight vector
$\boldsymbol{\epsilon}$	Noise vector
$a \in \mathbb{N}$	a is an element of the natural numbers
$a \in \mathbb{Z}$	a is an integer
$a \in \mathbb{R}$	a is a real valued scalar
$\boldsymbol{a} \in \mathbb{R}^n$	$\boldsymbol{a}$ is a real valued n-dimensional vector
$\boldsymbol{A} \in \mathbb{R}^{n \times p}$	$\boldsymbol{A}$ is a real valued $(n \times p)$-dimensional matrix
λ_1	Sparsity tuning parameter of the *Lasso* and the *elastic net*
λ_2	Ridge tuning parameter of the *elastic net*
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{N}(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\varnothing$	Empty set
K	Number of random experiments
L	Number of dummies
T	Number of included dummies after which the forward variable selection process in each random experiment is terminated
T^*	Optimal value of T as determined by the calibration process of the *T-Rex* selector
v	Voting level
v^*	Optimal value of v as determined by the calibration process of the *T-Rex* selector
α	Target FDR level

$\widehat{\alpha}$	FDR estimator
$\mathbb{1}_k(j, T, L)$	Indicator function of the kth random experiment of the *T-Rex* selector
$\mathrm{NHG}(p_0 + L, p_0, T)$	Negative hypergeometric distribution with $p_0 + L$ total elements, p_0 success elements, and T failures
$A \overset{d}{\leq} B$	Random variable A is stochastically dominated by random variable B
$\mathcal{Z}$	Index set of null variables
$\mathcal{A}$	Index set of active variables
$\widehat{\mathcal{A}}_L(v, T)$	Estimator of the active set, i.e., index set of the selected variables
$\widehat{\mathcal{A}}(v)$	Short hand for $\widehat{\mathcal{A}}_L(v, T)$
$\widehat{\mathcal{A}}^0(v)$	Index set of the selected null variables
$\widehat{\mathcal{A}}^1(v)$	Index set of the selected active variables
$\mathcal{C}_{k,L}(T)$	Candidate set of the kth random experiment, i.e., index set of the included variables in the kth random experiment
$V_{T,L}(v)$	Number of selected null variables
$\widehat{V}_{T,L}(v)$	Estimator of $V_{T,L}(v)$
$S_{T,L}(v)$	Number of selected true active variables
$R_{T,L}(v)$	Total number of selected variables
$\Phi_{T,L}(j)$	Relative occurrence of variable j
$\Delta\Phi_{t,L}(j)$	Change in relative occurrence from step $t-1$ to t for variable j
$\Phi'_{T,L}(j)$	Deflated relative occurrence of variable j
$\mathrm{Gr}(j, \rho_{\text{thr}})$	Group of variables that are associated with variable j
$\mathrm{SW}(j, \rho_{\text{thr}})$	Sliding window function
$\Psi_{T,L}(j, \rho_{\text{thr}})$	Dependency-aware penalty function with hierarchical graphical model group design
$\Psi^{\mathrm{NN}}_{T,L}(j, \rho_{\text{thr}})$	Dependency-aware penalty function with nearest neighbors group design
$\Psi^{\mathrm{AR}}_{T,L}(j, \rho_{\text{thr}})$	Dependency-aware penalty function with autoregressive group design
$\Phi^{\mathrm{DA}}_{T,L}(j, \rho_{\text{thr}})$	Dependency-aware relative occurrences with hierarchical graphical model group design
$\Phi^{\mathrm{NN}}_{T,L}(j, \rho_{\text{thr}})$	Dependency-aware relative occurrences with nearest neighbors group design
$\Phi^{\mathrm{AR}}_{T,L}(j, \rho_{\text{thr}})$	Dependency-aware relative occurrences with autoregressive group design

List of Figures

Figure 2.1 Illustration of the (a) *Lasso*, (b) ridge regression, and (c) *elastic net* optimization problems for two variables with associated coefficients β_1 and β_2. 12

Figure 2.2 Illustration of the *LARS* algorithm as in [Efr+04] for two predictors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. 15

Figure 3.1 The sequential computation time of the *T-Rex* selector is multiple orders of magnitude lower than that of the *model-X* knockoff method [Can+18]. 32

Figure 3.2 Simplified overview of the *T-Rex* selector framework. 33

Figure 3.3 The enlarged predictor matrices $\widetilde{\boldsymbol{X}}_k, k = 1, \ldots, K$, replace the original predictor matrix $\boldsymbol{X}$ in each random experiment within the *T-Rex* selector framework. 34

Figure 3.4 Ingredient 1 - sampling dummies from the univariate standard normal distribution. 35

Figure 3.5 Ingredient 2 - early terminating the solution paths of the random experiments. 36

Figure 3.6 Ingredient 3 - fusing the candidate sets based on their relative occurrences and a voting level $v \in [0.5, 1)$. 37

Figure 3.7 The *T-Rex* selector controls the FDR for all values of v and T while achieving a high power, even at low values of T. 53

Figure 3.8 The *model-X* knockoff method fails to control the FDR. Among the FDR-controlling methods, the *T-Rex* selector outperforms the *model-X* knockoff+ method in terms of power. 54

Figure 3.9 Average FDP and TPP in the case of dependent predictors. 55

Figure 3.10 The proposed ordinary and the confidence-based *Screen-T-Rex* selector both control the FDR at the self-estimated levels while achieving a reasonably high TPR. 61

Figure 4.1 Hierarchical graphical models: The dendrogram. 70

Figure 4.2 Illustration of the group design principle for dependency-aware FDR control in Theorem 10. 74

Figure 4.3 Only the proposed *T-Rex+DA* selector with a binary tree group model (*T-Rex+DA+BT*) reliably controls the FDR in all settings

while achieving a reasonably high TPR. 77
Figure 4.4 The proposed *T-Rex+DA* selector reliably controls the FDR in all settings while achieving a reasonably high TPR in harsh high correlation settings. 78
Figure 4.5 The numerical experiments verify that the proposed *T-Rex+DA+AR*1 selector for the autoregressive dependency model controls the FDR for all autocorrelation values ρ, while the original *T-Rex* selector loses its FDR control property for $\rho > 0.5$. 87

Figure 5.1 Solution paths of the (a) *EN* [ZH05] and (b) the proposed *IEN*. . 99
Figure 5.2 Relative computation times of one random experiment with $L = p$ and $T = 1$ of the *T-Rex*, *T-Rex+GVS (EN)*, and the proposed *T-Rex+GVS (IEN)*. 100
Figure 5.3 For the first PC, the proposed *T-Rex* PCA methods empirically control the FDR at a level of 10% while achieving an optimal TPR of 100% even at low SNRs. 107
Figure 5.4 Cumulative percentage of explained variance (PEV). 108

Figure 6.1 The heatmap visualizes the correlation matrix of the first 150 SNPs that were generated using the software HAPGEN2 [SMD11]. 112
Figure 6.2 The proposed *T-Rex* selector is the only method that has an average FDP below the target FDR level and that has a non-zero power. . . 114
Figure 6.3 The proposed *T-Rex+GVS* methods have the highest TPR values (i.e., average TPP), while their FDR values (i.e., average FDP) stay below the target level of 20%. 115
Figure 6.4 Number of selected mutations that are reported (green) and not reported (orange) in TSM lists for HIV-1 PI-type drugs. 117
Figure 6.5 Number of selected genes in the TCGA breast cancer survival analysis study. 119

Figure 7.1 Exemplary daily returns of the sorted 438 stocks. 125
Figure 7.2 The proposed *T-Rex+DA+NN* selector closely follows the S&P 500 index using the fewest number of stocks in almost all quarters. . 126
Figure 7.3 The proposed *T-Rex+DA+AR*1 selector shows (a) the best wealth tracking performance and (b) the lowest mean squared wealth tracking error (MSTE) in all quarters while (c) requiring only few stocks at a target FDR of 1% and $\rho_{\text{thr}} = 0.07$. 128
Figure 7.4 Correlation matrices of the 28 most influential stocks (according to their index weights) in the S&P 500 index. 129

Figure 8.1 The plot shows the accumulated monthly downloads of the developed open source R software packages TRexSelector [Mac+24c] and

tlars [Mac+24b]. 133

Figure A.1 Exemplary numerical verification of Corollary 4 and A-I. 161
Figure A.2 Exemplary numerical verification of A-II. 162
Figure A.3 Exemplary numerical verification of A-III. 163
Figure A.4 Exemplary numerical verification of Corollary 4 and A-I for HAPGEN2 genomics data. 164
Figure A.5 Exemplary numerical verification of A-II for HAPGEN2 genomics data. 165
Figure A.6 Exemplary numerical verification of A-III for HAPGEN2 genomics data. 166
Figure A.7 The *fixed-X* knockoff method fails to control the FDR. In terms of power, the proposed *T-Rex* selector outperforms the *fixed-X* knockoff method, the *fixed-X* knockoff+ method, and the *BY* method and shows a comparable performance to the *BH* method. 167
Figure A.8 Illustration of Theorem 4 (Dummy generation). 168
Figure A.9 Average FDP and TPP in the case of non-Gaussian predictors in $\boldsymbol{X}$. 169
Figure A.10 Average FDP and TPP in the case of non-Gaussian noise. 170

Figure B.1 Heavy-tailed predictor matrix $\boldsymbol{X}$. 184
Figure B.2 Heavy-tailed noise vector $\boldsymbol{\epsilon}$. 185
Figure B.3 Heavy-tailed predictor matrix $\boldsymbol{X}$. 186
Figure B.4 Heavy-tailed noise vector $\boldsymbol{\epsilon}$. 186

List of Tables

6.1 The proposed *T-Rex* selector is the only method whose average FDP lies below the target FDR level of 10% while achieving a non-zero power. 114
6.2 For all methods, the average achieved FDP is lower than the average estimated/target FDR, i.e., all methods control the FDR. 115
6.3 Results for the HIV-1 PI-type drugs. 117

7.1 The proposed *T-Rex+DA+NN* selector achieves a lower average (over all quarters) mean squared wealth tracking error (MSTE) using a smaller portfolio. 126
7.2 MSTE and number of selected stocks averaged over all testing periods. . . . 128

A.1 Overview of frequently used expressions. 140

References

[ABG08] O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: A process point of view*. Springer Science & Business Media, 2008 (cited on page 64).

[Abd+07] H. Abdi et al. "Bonferroni and Šidák corrections for multiple comparisons". In: *Encyclopedia of measurement and statistics* 3.01 (2007), p. 2007 (cited on page 4).

[Abr+06] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. "Adapting to unknown sparsity by controlling the false discovery rate". In: *Ann. Statist.* 34.2 (2006), pp. 584–653 (cited on page 3).

[Aka98] H. Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. Ed. by E. Parzen, K. Tanabe, and G. Kitagawa. New York, NY: Springer New York, 1998, pp. 199–213 (cited on page 12).

[All74] D. M. Allen. "The relationship between variable selection and data agumentation and a method for prediction". In: *Technometrics* 16.1 (1974), pp. 125–127 (cited on page 3).

[AL10] M. Avellaneda and J.-H. Lee. "Statistical arbitrage in the US equities market". In: *Quant. Finance* 10.7 (2010), pp. 761–782 (cited on page 129).

[Bak16a] M. Baker. "1,500 scientists lift the lid on reproducibility". In: *Nature* 533.7604 (2016), pp. 452–454 (cited on page 2).

[Bak16b] M. Baker. "Reproducibility crisis". In: *Nature* 533.26 (2016), pp. 353–66 (cited on page 80).

[Bal06] D. J. Balding. "A tutorial on statistical methods for population association studies". In: *Nat. Rev. Genet.* 7.10 (2006), pp. 781–791 (cited on pages 2, 64, 79).

[BC15] R. F. Barber and E. J. Candès. "Controlling the false discovery rate via knockoffs". In: *Ann. Statist.* 43.5 (2015), pp. 2055–2085 (cited on pages 2, 4, 19, 21, 23, 26, 41, 102, 116, 157, 158).

[BC19] R. F. Barber and E. J. Candès. "A knockoff filter for high-dimensional selective inference". In: *Ann. Statist.* 47.5 (2019), pp. 2504–2537 (cited on page 5).

[BFP17] K. Benidis, Y. Feng, and D. P. Palomar. "Sparse portfolios for high-dimensional financial index tracking". In: *IEEE Trans. Signal Process.* 66.1 (2017), pp. 155–170 (cited on pages 1, 122, 123, 125, 126).

[BP19] K. Benidis and D. P. Palomar. *sparseIndexTracking: Design of portfolio of stocks to track an index*. R package version 0.1.1. 2019. URL: https://CRAN.R-project.org/package=sparseIndexTracking (cited on page 126).

[Ben+16] K. Benidis, Y. Sun, P. Babu, and D. P. Palomar. "Orthogonal sparse PCA and covariance estimation via procrustes reformulation". In: *IEEE Trans. Signal Process.* 64.23 (2016), pp. 6211–6226 (cited on page 101).

[BH95] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 57.1 (1995), pp. 289–300 (cited on pages 4, 19, 20, 102, 157, 158).

[BY01] Y. Benjamini and D. Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *Ann. Statist.* 29.4 (2001), pp. 1165–1188 (cited on pages 4, 19, 20, 102, 158).

[Bes+21] V. Bessonneau, R. R. Gerona, J. Trowbridge, R. Grashow, T. Lin, H. Buren, R. Morello-Frosch, and R. A. Rudel. "Gaussian graphical modeling of the serum exposome and metabolome reveals interactions between environmental chemicals and endogenous metabolites". In: *Sci. Rep.* 11.1 (2021), p. 7607 (cited on page 136).

[Bhu+19] N. Bhushan, F. Mohnert, D. Sloot, L. Jans, C. Albers, and L. Steg. "Using a Gaussian graphical model to explore relationships between items and variables in environmental psychology research". In: *Front. Psychol.* 10 (2019), p. 453193 (cited on page 136).

[Bre+23] H. Brehier, A. Breloy, M. N. El Korso, and S. Kumar. "Robust and Globally Sparse PCA via Majorization-Minimization and Variable Splitting". In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. 2023, pp. 1–5 (cited on page 101).

[Bre+21] A. Breloy, S. Kumar, Y. Sun, and D. P. Palomar. "Majorization-minimization on the Stiefel manifold with application to robust sparse PCA". in: *IEEE Trans. Signal Process.* 69 (2021), pp. 1507–1520 (cited on page 101).

[BJ79] J. Buckley and I. James. "Linear regression with censored data". In: *Biometrika* 66.3 (1979), pp. 429–436 (cited on page 119).

[BV11] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media, 2011 (cited on pages 1, 9, 11).

[Bun+19] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Res.* 47.D1 (2019), pp. D1005–D1012 (cited on pages 1, 65, 110, 134).

[Can+18] E. J. Candès, Y. Fan, L. Janson, and J. Lv. "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 80.3 (2018), pp. 551–577 (cited on pages 2, 4, 19, 22, 23, 32, 64, 76, 86, 102, 110, 113, 126, 156, 157).

[Cha+07] S. J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter, G. Thomas, J. N. Hirschhorn, G. Abecasis, D. Altshuler, J. E. Bailey-Wilson, et al. "Replicating genotype–phenotype associations". In: *Nature* 447.7145 (2007), pp. 655–660 (cited on page 2).

[CZP20] J. Chen, W. Zhang, and H. V. Poor. "A false discovery rate oriented approach to parallel sequential change detection problems". In: *IEEE Trans. Signal Process.* 68 (2020), pp. 1823–1836 (cited on page 1).

[CSY18] Z. Chen, F. Sohrabi, and W. Yu. "Sparse activity detection for massive connectivity". In: *IEEE Trans. Signal Process.* 66.7 (2018), pp. 1890–1904 (cited on page 1).

[Chu+07] P.-J. Chung, J. F. Bohme, C. F. Mecklenbrauker, and A. O. Hero. "Detection of the number of signals using the Benjamini-Hochberg procedure". In: *IEEE Trans. Signal Process.* 55.6 (2007), pp. 2497–2508 (cited on page 1).

[Col+16] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, et al. "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data". In: *Nucleic Acids Res.* 44.8 (2016), e71–e71 (cited on page 118).

[CDT19] E. Commission, C. Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019 (cited on page 136).

[Cox75] D. R. Cox. "A note on data-splitting for the evaluation of significance levels". In: *Biometrika* 62.2 (1975), pp. 441–444 (cited on page 5).

[DH97] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997 (cited on page 58).

[DS12] P. Di Lorenzo and A. H. Sayed. "Sparse distributed learning based on diffusion adaptation". In: *IEEE Trans. Signal Process.* 61.6 (2012), pp. 1419–1433 (cited on page 1).

[Dzi+20] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin. "Sensitivity and specificity of information criteria". In: *Brief. Bioinform.* 21.2 (2020), pp. 553–565 (cited on page 3).

[Efr+04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. "Least angle regression". In: *Ann. Statist.* 32.2 (2004), pp. 407–499 (cited on pages 3, 11, 14–16, 32, 85, 141, 155).

[ET94] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC Press, 1994 (cited on page 58).

[FF92] E. F. Fama and K. R. French. "The cross-section of expected stock returns". In: *the Journal of Finance* 47.2 (1992), pp. 427–465 (cited on page 129).

[FL06] J. Fan and R. Li. "Statistical challenges with high dimensionality: Feature selection in knowledge discovery". In: *arXiv preprint math/0602133* (2006) (cited on page 1).

[FST14] W. Fithian, D. Sun, and J. Taylor. "Optimal inference after model selection". In: *arXiv preprint, arXiv:1410.2597* (2014) (cited on page 5).

[Fri+07] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. "Pathwise coordinate optimization". In: *Ann. Appl. Stat.* 1.2 (2007), pp. 302–332 (cited on pages 3, 16, 17).

[FHT10] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *J. Stat. Softw.* 33.1 (2010), pp. 1–22 (cited on pages 36, 156).

[GC18] M. D. Gallagher and A. S. Chen-Plotkin. "The post-GWAS era: From association to function". In: *Am. J. Hum. Genet.* 102.5 (2018), pp. 717–730 (cited on pages 2, 110).

[GBS09] Y. Gavrilov, Y. Benjamini, and S. K. Sarkar. "An adaptive step-down procedure with proven FDR control under independence". In: *Ann. Statist.* 37.2 (2009), pp. 619–629 (cited on page 157).

[Gra+20] J. Graff, M. J. Ringuette, T. Singh, and F. D. Lagor. "Reduced-order modeling for dynamic mode decomposition without an arbitrary sparsity parameter". In: *AIAA J.* 58.9 (2020), pp. 3919–3931 (cited on pages 135, 136).

[HLZ15] C. R. Harvey, Y. Liu, and H. Zhu. ". . . and the Cross-Section of Expected Returns". In: *Rev. Financ. Stud.* 29.1 (2015), pp. 5–68 (cited on pages 1, 2).

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009 (cited on pages 3, 11–13, 17).

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015 (cited on page 11).

[Hir73] A. Hirotugu. “Information theory and an extension of the maximum likelihood principle”. In: *Second International Symposium on Information Theory*. 1973, pp. 267–281 (cited on page 3).

[Hoc88] Y. Hochberg. “A sharper Bonferroni procedure for multiple tests of significance”. In: *Biometrika* 75.4 (1988), pp. 800–802 (cited on page 4).

[HK70] A. E. Hoerl and R. W. Kennard. “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1 (1970), pp. 55–67 (cited on pages 17, 18).

[Hog+08] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding. “Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies”. In: *PLOS Genet.* 4.7 (2008), e1000130 (cited on page 4).

[Hol79] S. Holm. “A simple sequentially rejective multiple test procedure”. In: *Scand. J. Stat.* (1979), pp. 65–70 (cited on page 4).

[13] “How Science Goes Wrong”. In: *The Economist* (Oct. 21, 2013). URL: https://economist.com/leaders/2013/10/21/how-science-goes-wrong (visited on 06/26/2024) (cited on page 2).

[Hu+15] Z. Hu, G. Pan, Y. Wang, and Z. Wu. “Sparse principal component analysis via rotation and truncation”. In: *IEEE Trans. Neural Netw. Learn. Syst.* 27.4 (2015), pp. 875–890 (cited on page 101).

[Huf18] J. E. Huffman. “Examining the current standards for genetic discovery and replication in the era of mega-biobanks”. In: *Nat. Commun.* 9.1 (2018), pp. 1–4 (cited on pages 1, 2).

[Ioa05] J. P. Ioannidis. “Why most published research findings are false”. In: *PLOS Med.* 2.8 (2005), e124 (cited on page 2).

[Iqb+16] K. Iqbal, B. Buijsse, J. Wirth, M. B. Schulze, A. Floegel, and H. Boeing. “Gaussian graphical models identify networks of dietary intake in a German adult population”. In: *J. Nutr.* 146.3 (2016), pp. 646–652 (cited on page 136).

[JV02] R. Jansen and R. Van Dijk. “Optimal benchmark tracking with small portfolios”. In: *J. Portf. Manag.* 28.2 (2002), p. 33 (cited on page 122).

[Jol03] I. T. Jolliffe. “Principal component analysis”. In: *Technometrics* 45.3 (2003), p. 276 (cited on page 100).

[JC16] I. T. Jolliffe and J. Cadima. “Principal component analysis: a review and recent developments”. In: *Philos. Trans. R. Soc. A* 374.2065 (2016), p. 20150202 (cited on page 100).

[KA17] H. F. M. Kamel and H. S. A. B. Al-Amodi. “Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine”. In:

Genomics Proteomics Bioinformatics 15.4 (2017), pp. 220–235 (cited on pages 64, 118).

[Kle13] A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013 (cited on page 23).

[KMM24] T. Koka, J. Machkour, and M. Muma. "False Discovery Rate Control for Gaussian Graphical Models via Neighborhood Screening". In: *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*. 2024, pp. 2482–2486 (cited on pages 134, 136).

[Lac06] M. Lacroix. "Significance, detection and markers of disseminated breast cancer cells". In: *Endocr.-Relat. Cancer* 13.4 (2006), pp. 1033–1067 (cited on page 120).

[Lee+16] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. "Exact post-selection inference, with application to the lasso". In: *Ann. Statist.* 44.3 (2016), pp. 907–927 (cited on page 5).

[Loc+14] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. "A significance test for the lasso". In: *Ann. Statist.* 42.2 (2014), pp. 413–468 (cited on page 5).

[LHA14] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol.* 15.12 (2014), pp. 1–21 (cited on page 118).

[Lu+18] Y. Lu, Y. Fan, J. Lv, and W. Stafford Noble. "DeepPINK: reproducible feature selection in deep neural networks". In: *Adv. Neural Inf. Process. Syst.* 31 (2018) (cited on page 4).

[Mac+17] J. Machkour, B. Alt, M. Muma, and A. M. Zoubir. "The outlier-corrected-data-adaptive Lasso: A new robust estimator for the independent contamination model". In: *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*. 2017, pp. 1649–1653 (cited on page 1).

[Mac+24a] J. Machkour, A. Breloy, M. Muma, D. P. Palomar, and F. Pascal. "Sparse PCA with False Discovery Rate Controlled Variable Selection". In: *Proc. IEEE 49th Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. 2024, pp. 9716–9720 (cited on pages 90, 121, 132).

[Mac+20] J. Machkour, M. Muma, B. Alt, and A. M. Zoubir. "A robust adaptive Lasso estimator for the independent contamination model". In: *Signal Process.* 174 (2020), p. 107608 (cited on page 1).

[MMPewa] J. Machkour, M. Muma, and D. P. Palomar. "The Terminating-Random Experiments Selector: Fast High-Dimensional Variable Selection with False Discovery Rate Control". In: *arXiv preprint arXiv:2110.06048* (under review) (cited on pages 30, 64, 75, 109, 131).

[MMP22] J. Machkour, M. Muma, and D. P. Palomar. "False Discovery Rate Control for Grouped Variable Selection in High-Dimensional Linear Models Using the T-Knock Filter". In: *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*. 2022, pp. 892–896 (cited on pages 75, 90, 109, 132).

[MMP23a] J. Machkour, M. Muma, and D. P. Palomar. "False Discovery Rate Control for Fast Screening of Large-Scale Genomics Biobanks". In: *Proc. 22nd IEEE Statist. Signal Process. Workshop (SSP)*. 2023, pp. 666–670 (cited on pages 30, 75, 109, 131).

[MMP23b] J. Machkour, M. Muma, and D. P. Palomar. "The Informed Elastic Net for Fast Grouped Variable Selection and FDR Control in Genomics Research". In: *Proc. IEEE 9th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*. 2023, pp. 466–470 (cited on pages 90, 109, 132).

[MMP24] J. Machkour, M. Muma, and D. P. Palomar. "FDR-Controlled Sparse Index Tracking with Autoregressive Stock Dependency Models". In: *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*. 2024, pp. 2662–2666 (cited on pages 64, 121, 132).

[MMPewb] J. Machkour, M. Muma, and D. P. Palomar. "High-Dimensional False Discovery Rate Control for Dependent Variables". In: *arXiv preprint arXiv:2401.15796* (under review) (cited on pages 64, 109, 132).

[MPMew] J. Machkour, D. P. Palomar, and M. Muma. "FDR-Controlled Portfolio Optimization for Sparse Financial Index Tracking". In: *arXiv preprint arXiv:2401.15139* (under review) (cited on pages 64, 121, 132).

[Mac+24b] J. Machkour, S. Tien, D. P. Palomar, and M. Muma. *tlars: The T-LARS Algorithm: Early-Terminated Forward Variable Selection*. R package version 1.0.1. 2024. URL: https://CRAN.R-project.org/package=tlars (cited on pages 30, 36, 64, 90, 103, 109, 121, 133, 164).

[Mac+24c] J. Machkour, S. Tien, D. P. Palomar, and M. Muma. *TRexSelector: T-Rex Selector: High-Dimensional Variable Selection & FDR Control*. R package version 1.0.0. 2024. URL: https://CRAN.R-project.org/package=TRexSelector (cited on pages 30, 50, 64, 74, 90, 109, 121, 133, 164).

[MO07] D. Maringer and O. Oyewumi. "Index tracking with constrained portfolios". In: *Intell. Syst. Account. Finance Manag.* 15.1-2 (2007), pp. 57–71 (cited on page 122).

[Mei13] N. Meinshausen. "Sign-constrained least squares estimation for high-dimensional regression". In: *Electron. J. Statist.* 7 (2013), pp. 1607–1631 (cited on page 123).

[MB10] N. Meinshausen and P. Bühlmann. "Stability selection". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 72.4 (2010), pp. 417–473 (cited on page 5).

[MMB09] N. Meinshausen, L. Meier, and P. Bühlmann. "P-values for high-dimensional regression". In: *J. Amer. Statist. Assoc.* 104.488 (2009), pp. 1671–1681 (cited on page 5).

[Mil84] A. J. Miller. "Selection of subsets of regression variables". In: *J. R. Stat. Soc. Ser. A. Gen.* 147.3 (1984), pp. 389–410 (cited on page 45).

[Mil02] A. J. Miller. *Subset selection in regression*. CRC Press, 2002 (cited on page 45).

[Mul21] A. Mullard. "Half of top cancer studies fail high-profile reproducibility effort". In: *Nature* 600.7889 (2021), pp. 368–369 (cited on page 2).

[MC12] F. Murtagh and P. Contreras. "Algorithms for hierarchical clustering: an overview". In: *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* 2.1 (2012), pp. 86–97 (cited on pages 69, 92, 93, 95, 124).

[Nat95] B. K. Natarajan. "Sparse approximate solutions to linear systems". In: *SIAM J. Comput.* 24.2 (1995), pp. 227–234 (cited on pages 3, 11).

[Nor+21] A. Norbury, S. H. Liu, J. J. Campaña-Montes, L. Romero-Medrano, M. L. Barrigón, E. Smith, P. O. A. J. R.-B. L. S.-V. M. Villalba Hospital Madrid Alcón-Durán Ana Di Stasio Ezequiel García-Vega Juan Manuel Martín-Calvo Ana López-Gómez, M. B.-G. S. M. C.-J. R. F.-C. A. H.-M. E. V.-O. S. Infanta Elena Hospital, et al. "Social media and smartphone app use predicts maintenance of physical activity during Covid-19 enforced isolation in psychiatric outpatients". In: *Mol. Psychiatry* 26.8 (2021), pp. 3920–3930 (cited on page 136).

[OPT00a] M. R. Osborne, B. Presnell, and B. A. Turlach. "A new approach to variable selection in least squares problems". In: *IMA J. Numer. Anal.* 20.3 (2000), pp. 389–403 (cited on page 16).

[OPT00b] M. R. Osborne, B. Presnell, and B. A. Turlach. "On the lasso and its dual". In: *J. Comput. Graph. Statist.* 9.2 (2000), pp. 319–337 (cited on page 16).

[ŐLG21] Á. Ősz, A. Lánczky, and B. Győrffy. "Survival analysis in breast cancer using proteomic data from four independent datasets". In: *Sci. Rep.* 11.1 (2021), p. 16787 (cited on page 64).

[Pal24] D. P. Palomar. *Portfolio Optimization: Theory and Application*. To appear in Cambridge Univ. Press, 2024 (cited on pages 1, 122).

[Pri07] J.-L. Prigent. *Portfolio optimization and performance analysis*. CRC Press, 2007 (cited on page 122).

[Rei+01] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, et al. "Linkage disequilibrium in the human genome". In: *Nature* 411.6834 (2001), pp. 199–204 (cited on pages 90, 95, 111, 158).

[RB24] Z. Ren and R. F. Barber. "Derandomised knockoffs: leveraging e-values for false discovery rate control". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 86.1 (2024), pp. 122–154 (cited on page 4).

[RWC21] Z. Ren, Y. Wei, and E. Candès. "Derandomizing knockoffs". In: *J. Amer. Statist. Assoc.* (2021), pp. 1–11 (cited on page 4).

[Rhe+05] S.-Y. Rhee, W. J. Fessel, A. R. Zolopa, L. Hurley, T. Liu, J. Taylor, D. P. Nguyen, S. Slome, D. Klein, M. Horberg, et al. "HIV-1 Protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance". In: *The Journal of infectious diseases* 192.3 (2005), pp. 456–465 (cited on page 116).

[Rhe+06] S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. "Genotypic predictors of human immunodeficiency virus type 1 drug resistance". In: *Proceedings of the National Academy of Sciences* 103.46 (2006), pp. 17355–17360 (cited on page 116).

[RM11] D. Ruppert and D. S. Matteson. *Statistics and data analysis for financial engineering*. Vol. 13. Springer, 2011 (cited on page 129).

[SS99] H. Sackrowitz and E. Samuel-Cahn. "P values as random variables—expected P values". In: *Am. Stat.* 53.4 (1999), pp. 326–331 (cited on page 158).

[SMM23] F. Scheidt, J. Machkour, and M. Muma. "Solving FDR-Controlled Sparse Regression Problems with Five Million Variables on a Laptop". In: *Proc. IEEE 9th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*. 2023, pp. 116–120 (cited on pages 134, 135).

[Sch+24] C. A. Schroth, C. Eckrich, I. Kakouche, S. Fabian, O. von Stryk, A. M. Zoubir, and M. Muma. "Emergency response person localization and vital sign estimation using a semi-autonomous robot mounted SFCW radar". In: *IEEE Trans. Biomed. Eng.* (2024) (cited on pages 134, 135).

[Sch78] G. Schwarz. "Estimating the dimension of a model". In: *Ann. Statist.* 6.2 (1978), pp. 461–464 (cited on pages 3, 12).

[Sco+13] A. Scozzari, F. Tardella, S. Paterlini, and T. Krink. "Exact and heuristic approaches for the index tracking problem with UCITS constraints". In: *Ann. Oper. Res.* 205 (2013), pp. 235–250 (cited on page 122).

[SDC03] M. R. Segal, K. D. Dahlquist, and B. R. Conklin. "Regression approaches for microarray data analysis". In: *J. Comput. Biol.* 10.6 (2003), pp. 961–980 (cited on pages 2, 64, 79).

[SSC19] M. Sesia, C. Sabatti, and E. J. Candès. "Gene hunting with hidden Markov model knockoffs". In: *Biometrika* 106.1 (2019), pp. 1–18 (cited on pages 64, 92, 111–113, 158).

[Sha95] J. P. Shaffer. "Multiple hypothesis testing". In: *Annu. Rev. Psychol.* 46.1 (1995), pp. 561–584 (cited on pages 3, 4).

[SS13] R. D. Shah and R. J. Samworth. "Variable selection with error control: another look at stability selection". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 75.1 (2013), pp. 55–80 (cited on page 5).

[Sim+11] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. "Regularization paths for Cox's proportional hazards model via coordinate descent". In: *J. Stat. Softw.* 39.5 (2011), p. 1 (cited on pages 11, 119).

[SL21] Z. Song and J. Li. "Variable selection with false discovery rate control in deep neural networks". In: *Nat. Mach. Intell.* 3.5 (2021), pp. 426–433 (cited on pages 4, 5).

[SPP18] C. Steffens, M. Pesavento, and M. E. Pfetsch. "A compact formulation for the $\ell_{2,1}$ mixed-norm minimization problem". In: *IEEE Trans. Signal Process.* 66.6 (2018), pp. 1483–1497 (cited on page 135).

[Sto74] M. Stone. "Cross-validatory choice and assessment of statistical predictions". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 36.2 (1974), pp. 111–133 (cited on page 3).

[Sto77] M. Stone. "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 39.1 (1977), pp. 44–47 (cited on page 3).

[Sto03] J. D. Storey. "The positive false discovery rate: a Bayesian interpretation and the q-value". In: *Ann. Statist.* 31.6 (2003), pp. 2013–2035 (cited on page 20).

[STS04] J. D. Storey, J. E. Taylor, and D. Siegmund. "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 66.1 (2004), pp. 187–205 (cited on pages 23, 26, 102, 157, 158).

[SBC17] W. Su, M. Bogdan, and E. J. Candès. "False discoveries occur early on the lasso path". In: *Ann. Statist.* 45.5 (2017), pp. 2133–2150 (cited on page 41).

[SMD11] Z. Su, J. Marchini, and P. Donnelly. "HAPGEN2: simulation of multiple disease SNPs". In: *Bioinformatics* 27.16 (2011), pp. 2304–2305 (cited on pages 110, 112, 157–159, 162).

[Sud+15] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. "UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLOS Med.* 12.3 (2015), e1001779 (cited on pages 1, 2, 65, 110, 131, 134, 135).

[Sur+22] P. Surendran, I. D. Stewart, V. P. Au Yeung, M. Pietzner, J. Raffler, M. A. Wörheide, C. Li, R. F. Smith, L. B. Wittemans, L. Bomba, et al. "Rare and common genetic determinants of metabolic individuality and their effects on human health". In: *Nat. Med.* 28.11 (2022), pp. 2321–2332 (cited on page 136).

[TEN14] Z. Tan, Y. C. Eldar, and A. Nehorai. "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint". In: *IEEE Trans. Signal Process.* 62.21 (2014), pp. 5565–5576 (cited on page 1).

[The10] The International HapMap 3 Consortium. "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467.7311 (2010), pp. 52–58 (cited on page 110).

[Tib96] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 58.1 (1996), pp. 267–288 (cited on pages 3, 11, 32, 85, 141, 155).

[Tib97] R. Tibshirani. "The lasso method for variable selection in the Cox model". In: *Stat. Med.* 16.4 (1997), pp. 385–395 (cited on page 11).

[Tib+05] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. "Sparsity and smoothness via the fused lasso". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 67.1 (2005), pp. 91–108 (cited on page 11).

[Tib+16] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. "Exact post-selection inference for sequential regression procedures". In: *J. Amer. Statist. Assoc.* 111.514 (2016), pp. 600–620 (cited on page 5).

[TA77] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Great Falls, MT, USA: Winston, 1977 (cited on pages 17, 18).

[TCW15] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". In: *Contemp Oncol (Pozn)* 2015.1 (2015), pp. 68–77 (cited on pages 2, 64, 118).

[Uff+21] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. "Genome-wide association studies". In: *Nat. Rev. Methods Primers* 1.1 (2021), p. 59 (cited on page 110).

[US08] M. O. Ulfarsson and V. Solo. "Sparse variable PCA using geodesic steepest descent". In: *IEEE Trans. Signal Process.* 56.12 (2008), pp. 5823–5832 (cited on page 101).

[Vis+17] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. "10 years of GWAS discovery: Biology, function, and translation". In: *Am. J. Hum. Genet.* 101.1 (2017), pp. 5–22 (cited on page 2).

[WJ20] W. Wang and L. Janson. "A power analysis of the conditional randomization test and knockoffs". In: *arXiv preprint arXiv:2010.02304* (2020) (cited on page 134).

[WR09] L. Wasserman and K. Roeder. "High dimensional variable selection". In: *Ann. Statist.* 37.5A (2009), pp. 2178–2201 (cited on page 5).

[WBC17] A. Weinstein, R. Barber, and E. J. Candès. "A power and prediction analysis for knockoffs with lasso statistics". In: *arXiv preprint, arXiv:1712.06465* (2017) (cited on page 134).

[Wei+23] A. Weinstein, W. J. Su, M. Bogdan, R. Foygel Barber, and E. J. Candès. "A power analysis for model-X knockoffs with ℓ p-regularized statistics". In: *Ann. Statist.* 51.3 (2023), pp. 1005–1029 (cited on page 134).

[Wil91] D. Williams. *Probability with martingales*. Cambridge Univ. Press, 1991 (cited on page 23).

[WT08] D. M. Witten and R. Tibshirani. "Testing significance of features by lassoed principal components". In: *Ann. Appl. Stat.* 2.3 (2008), pp. 986–1012 (cited on page 101).

[WBS07] Y. Wu, D. D. Boos, and L. A. Stefanski. "Controlling variable selection by the addition of pseudovariables". In: *J. Amer. Statist. Assoc.* 102.477 (2007), pp. 235–243 (cited on page 45).

[XLX16] F. Xu, Z. Lu, and Z. Xu. "An efficient optimization approach for a cardinality-constrained index tracking problem". In: *Optim. Methods Softw.* 31.2 (2016), pp. 258–271 (cited on page 122).

[Xu+21] Z. Xu, L. Xiang, R. Wang, Y. Xiong, H. Zhou, H. Gu, J. Wang, L. Peng, et al. "Bioinformatic analysis of immune significance of RYR2 mutation in breast cancer". In: *BioMed Res. Int.* 2021 (2021) (cited on page 120).

[Yan+19] C. Yang, X. Shen, H. Ma, B. Chen, Y. Gu, and H. C. So. "Weakly convex regularized robust sparse recovery methods with theoretical guarantees". In: *IEEE Trans. Signal Process.* 67.19 (2019), pp. 5046–5061 (cited on page 1).

[YL06] M. Yuan and Y. Lin. "Model selection and estimation in regression with grouped variables". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 68.1 (2006), pp. 49–67 (cited on pages 11, 96).

[Zer+14] T. Zerenner, P. Friederichs, K. Lehnertz, and A. Hense. "A Gaussian graphical model approach to climate networks". In: *Chaos* 24.2 (2014) (cited on page 136).

[Zho+19] C. Zhou, M. Wang, J. Yang, H. Xiong, Y. Wang, and J. Tang. "Integral membrane protein 2A inhibits cell growth in human breast cancer via enhanc-

ing autophagy induction". In: *Cell Commun. Signal.* 17 (2019), pp. 1–14 (cited on page 120).

[Zou06] H. Zou. "The adaptive lasso and its oracle properties". In: *J. Amer. Statist. Assoc.* 101.476 (2006), pp. 1418–1429 (cited on pages 3, 11, 32, 141).

[ZH05] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 67.2 (2005), pp. 301–320 (cited on pages 11, 17–19, 32, 91, 99, 102, 141, 188).

[ZHT06] H. Zou, T. Hastie, and R. Tibshirani. "Sparse principal component analysis". In: *J. Comput. Graph. Stat.* 15.2 (2006), pp. 265–286 (cited on pages 3, 101–105).

[ZHT07] H. Zou, T. Hastie, and R. Tibshirani. "On the "degrees of freedom" of the lasso". In: *Ann. Statist.* 35.5 (2007), pp. 2173–2192 (cited on page 13).

[ZX18] H. Zou and L. Xue. "A selective overview of sparse principal component analysis". In: *Proc. IEEE* 106.8 (2018), pp. 1311–1320 (cited on page 101).

[ZI04] A. M. Zoubir and D. R. Iskander. *Bootstrap techniques for signal processing*. Cambridge Univ. Press, 2004 (cited on page 58).

[Zou+12] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma. "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts". In: *IEEE Signal Process. Mag.* 29.4 (2012), pp. 61–80 (cited on page 1).

[Zou+18] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma. *Robust statistics for signal processing*. Cambridge Univ. Press, 2018 (cited on page 1).

Erklärungen laut Promotionsordnung

§8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass von mir zu keinem vorherigen Zeitpunkt bereits ein Promotionsversuch unternommen wurde. Andernfalls versichere ich, dass der promotionsführende Fachbereich über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs informiert ist.

§9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation, abgesehen von den in ihr ausdrücklich genannten Hilfsmitteln, selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde. Weiterhin versichere ich, dass die „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Technischen Universität Darmstadt“ sowie die „Leitlinien zum Umgang mit digitalen Forschungsdaten an der TU Darmstadt“ in den jeweils aktuellen Versionen bei der Verfassung der Dissertation beachtet wurden.

§9 Abs. 2 PromO

Ich versichere hiermit, dass die vorliegende Dissertation bisher noch nicht zu Prüfungszwecken gedient hat.

Darmstadt, 25.06.2024

Jasin Machkour