Signal Processing 234 (2025) 109990

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

High-dimensional false discovery rate control for dependent variables☆

Jasin Machkour [a,*,1], Michael Muma [a,2], Daniel P. Palomar [b,3]

[a] Robust Data Science Group at Technische Universität Darmstadt, Germany
[b] Convex Optimization Group, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
T-rex selector
False discovery rate (FDR) control
High-dimensional
Martingale theory
Survival analysis

ABSTRACT

Algorithms that ensure reproducible findings from large-scale, high-dimensional data are pivotal in numerous signal processing applications. In recent years, multivariate false discovery rate (FDR) controlling methods have emerged, providing guarantees even in high-dimensional settings where the number of variables surpasses the number of samples. However, these methods often fail to reliably control the FDR in the presence of highly dependent variable groups, a common characteristic in fields such as genomics and finance. To tackle this critical issue, we introduce a novel framework that accounts for general dependency structures. Our proposed dependency-aware T-Rex selector integrates hierarchical graphical models within the T-Rex framework to effectively harness the dependency structure among variables. Leveraging martingale theory, we prove that our variable penalization mechanism ensures FDR control. We further generalize the FDR-controlling framework by stating and proving a clear condition necessary for designing both graphical and non-graphical models that capture dependencies. Numerical experiments and a breast cancer survival analysis use-case demonstrate that the proposed method is the only one among the state-of-the-art benchmark methods that controls the FDR and reliably detects genes that have been previously identified to be related to breast cancer. An open-source implementation is available within the R package TRexSelector on CRAN.

1. Introduction

Reliably detecting as many as possible of the few relevant variables (i.e., features, biomarkers, or signals) in a large set of candidate variables given high-dimensional and noisy data while minimizing the number of false detections is required in many of today's signal processing applications, e.g. [1–11]. 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 [12–14]. 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 [15].

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. Such methods guarantee that the expected fraction of false discoveries among all discoveries (i.e., FDR) does not exceed a user-specified target level (e.g., 5, 10, 20%) while maximizing the number of selected variables. Other approaches that provide reliable detection of relevant variables are multiple testing methods that control the conservative family-wise error rate (FWER) (i.e., the probability of making one or more type I errors). However, the FWER-metric is not considered in this work, since it is a very conservative metric that usually leads to many missed discoveries in variable selection tasks. This weakness of the FWER-metric has lead to the development of the more liberal FDR-metric, which usually allows for more discoveries while controlling a user-defined target FDR [16].

Popular FDR-controlling methods for the low-dimensional setting, where the number of samples n is equal to or larger than the number of candidate variables p, are the Benjamini–Hochberg (*BH*) method [16],

☆ Extensive computations on the Lichtenberg High-Performance Computer of the Technische Universität Darmstadt were conducted for this research.
* Corresponding author.
E-mail addresses: jasin.machkour@tu-darmstadt.de (J. Machkour), michael.muma@tu-darmstadt.de (M. Muma), palomar@ust.hk (D.P. Palomar).
[1] J. Machkour is supported by the LOEWE initiative (Hesse, Germany) within the emergenCITY center.
[2] M. Muma is supported by the ERC Starting Grant ScReeningData (Project Number: 101042407).
[3] D.P. Palomar is supported by the Hong Kong GRF 16206123 research grant.

https://doi.org/10.1016/j.sigpro.2025.109990
Received 26 October 2024; Received in revised form 28 January 2025; Accepted 4 March 2025
Available online 15 March 2025

Benjamini-Yekutieli (*BY*) method [17], *fixed-X* knockoff method [18], and related approaches (e.g., [19,20]).

For the considered high-dimensional setting, where $p > n$, there exist *model-X* knockoff methods [21–24] and Terminating-Random Experiments (*T-Rex*) methods [25–31]. It has been shown that the computational complexity of the *T-Rex* selector [25] is linear in p, which makes it scalable to millions of variables in a reasonable computation time, while the *model-X* knockoff method [21] is practically infeasible in such large-scale settings (see Fig. 1 in [25]). 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 [32], genomics [33], and stock returns data [34].

In order to reduce the dependencies among the candidate variables, pruning approaches have been used [21,22,25]. In general, pruning methods cluster highly dependent 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 [35,36], 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.

Therefore, 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 hierarchical graphical model (i.e., binary tree) 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 [37], we prove that the proposed approach controls the FDR (Theorem 2).
(3) We 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 3).
(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 4).
(5) Numerical experiments and a real-world breast cancer survival analysis verify the theoretical results and demonstrate the practical usefulness of the proposed framework.

Organization: Section 2 briefly revisits the *T-Rex* selector and introduces the FDR control problem. Section 3 analyzes the relative occurrences of correlated variables for the original *T-Rex* framework (Theorem 1), describes the proposed *T-Rex+DA* methodology, and proves our main theoretical results (Theorems 2 and 3). Section 4 describes the implementation details and theoretical properties (Theorem 4) of the proposed *T-Rex+DA* calibration algorithm. Section 5 numerically verifies the theoretical FDR control results and compares the proposed approach against state-of-the-art methods. Section 6 presents the results of an FDR-controlled breast cancer survival analysis. Section 7 concludes the paper.

2. The T-Rex framework

In this section, the FDR and the true positive rate (TPR) are defined and the *T-Rex* framework is briefly revisited.

2.1. FDR and TPR

A general setting for sparse variable selection consists of $p = p_1 + p_0$ variables out of which p_1 variables are true active variables (i.e., variables associated with a response of interest $\boldsymbol{y}$) and p_0 variables are null (i.e., non-active) variables. The task of sparse variable selection is to determine an estimator $\widehat{\mathcal{A}}$ of the true active set $\mathcal{A} \subseteq \{1, \dots, p\}$ of cardinality $|\mathcal{A}| = p_1$. In this general setting, the false discovery rate (FDR) and the true positive rate (TPR) are defined by

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

and

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

respectively, where $\vee$ is the maximum operator, i.e., $|\widehat{\mathcal{A}}| \vee 1 := \max\{|\widehat{\mathcal{A}}|, 1\}$. In words,

(1) the FDR is the expectation of the false discovery proportion (FDP), i.e., the fraction of selected null variables among all selected variables, and
(2) the TPR is the expectation of the true positive proportion (TPP), i.e., the fraction of selected true active variables among all true active variables.

In practice, a tradeoff between the FDR and TPR must be managed. That is, the TPR is maximized while not exceeding a user-specified target FDR level $\alpha \in [0, 1]$, i.e., $\mathrm{FDR} \leq \alpha$.

2.2. The T-Rex selector

The *T-Rex* selector is an FDR-controlling variable selection framework. A schematic overview of the generic framework is provided in Fig. 1. A major characteristic of the *T-Rex* framework is that it conducts K supervised, independent, and early terminated random experiments based on a response vector $\boldsymbol{y} \in \mathbb{R}^n$ and the enlarged predictor matrices

$$\widetilde{\boldsymbol{X}}_k = \big[\boldsymbol{X} \ \overset{\circ}{\boldsymbol{X}}_k\big] \in \mathbb{R}^{n \times (p+L)}, \ k = 1, \dots, K,$$

where $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_p] \in \mathbb{R}^{n \times p}$ is the original predictor matrix containing the p predictor variables and $\overset{\circ}{\boldsymbol{X}}_k = [\overset{\circ}{\boldsymbol{x}}_1 \ \dots \ \overset{\circ}{\boldsymbol{x}}_L] \in \mathbb{R}^{n \times L}$ is a matrix containing L dummy variables. The dummy vectors $\overset{\circ}{\boldsymbol{x}}_l = [\overset{\circ}{x}_{1,l} \ \dots \ \overset{\circ}{x}_{n,l}]^\top$, $l = 1, \dots, L$, can be sampled independently from any univariate probability distribution with finite mean and variance (cf. Theorem 2 of [25]). The K early terminated random experiments are conducted with $\boldsymbol{y}$ and $\widetilde{\boldsymbol{X}}_k$, $k = 1, \dots, K$, as inputs to a forward variable selection method such as the *LARS* algorithm [38], *Lasso* [39], *elastic net* [40], or related methods (e.g., [8,9,41]). These forward selection methods include no more than one variable in each iteration.[4] All K random experiments are terminated independently of each other after T dummies have been included along their respective selection paths. This results in K candidate sets $C_{k,L}(T) \subseteq \{1, \dots, p\}$, $k = 1, \dots, K$, which contain the indices of all original variables in $\boldsymbol{X}$ that have been selected before the

[4] Since the *LARS* method is used throughout this paper, variables are only added but never removed along the solution paths. However, as also discussed in [25], the *T-Rex* selector, as well as the proposed *T-Rex+DA* selector, are compatible with forward selection methods, such as *Lasso* and *elastic net*, where previously selected variables may be removed at later steps. For such methods, the number of currently active dummies may decrease during the selection process. Nonetheless, this behavior does not affect the applicability of the *T-Rex* framework, because the solution paths are terminated as soon as T dummies are included for the first time. Therefore, the termination step of any forward selection process is always well-defined.

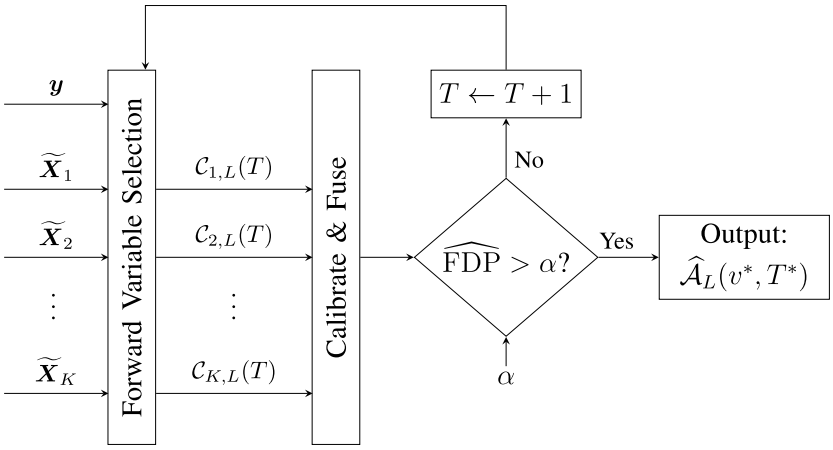

Fig. 1. Schematic overview of the *T-Rex* framework.

random experiments have been terminated. Using the candidate sets, relative occurrences are computed for each variable, i.e.,

$$\boldsymbol{\Phi}_{T,L}(j) := \begin{cases} \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_k(j,T,L), & T \geq 1 \\ 0, & T = 0, \end{cases} \tag{1}$$

where $\mathbb{1}_k(j,T,L)$ is an indicator function that takes the value one if the jth variable is contained in the candidate set $\mathcal{C}_{k,L}(T)$ and zero otherwise. Based on these relative occurrences, the final set of selected variables is given by

$$\widehat{\mathcal{A}}_L(v,T) := \{j : \boldsymbol{\Phi}_{T,L}(j) > v\}, \tag{2}$$

i.e., all variables whose relative occurrences exceed a voting threshold $v \in [0.5, 1)$ are selected. The extended calibration algorithm of the *T-Rex* selector [25] automatically determines the triple $(v^*, T^*, L) \in [0.5, 1) \times \{1, \dots, L\} \times \mathbb{N}_+$ such that the FDR is controlled at the user-defined target level $\alpha \in [0, 1]$, i.e.,

$$\text{FDR}(v^*, T^*, L) \leq \alpha.$$

The calibrated *T-Rex* parameters v and T are highlighted with a superscript "$*$" to emphasize that for any L, the parameters v^* and T^* are optimal in the sense that the FDR is controlled, while the number of selected variables is maximized (cf. Theorem 3 of [25]). For the choice of the number of dummies L, the extended calibration algorithm considers a tradeoff between the computer memory consumption for storing large dummy matrices and maximizing the number of selected variables. However, the FDR is always controlled for any choice of L (cf. Theorem 1 of [25]). The extended calibration algorithm achieves the optimal solution (v^*, T^*) by

(1) first terminating all random experiments after only one dummy has been included,
(2) computing a conservative estimator $\widehat{\text{FDP}}(v,T,L)$ that satisfies

$$\text{FDR}(v,T,L) = \mathbb{E}\big[\text{FDP}(v,T,L)\big] \leq \mathbb{E}\big[\widehat{\text{FDP}}(v,T,L)\big],$$

(3) iteratively increasing the number of included dummies T until $\widehat{\text{FDP}}(v = 1 - 1/K, T, L)$ (i.e., the FDP estimator at the effectively highest voting level $v = 1 - 1/K$) exceeds the target level α for the first time, and
(4) returning to the solution (v^*, T^*) of the preceding iteration T^* that satisfies the equation

$$v^* = \inf\big\{v \in [0.5, 1) : \widehat{\text{FDP}}(v, T^*, L) \leq \alpha\big\},$$

where v^* is the lowest feasible voting level such that $\widehat{\text{FDP}}$ does not exceed the target level α.

For details on the design and properties of the conservative FDP estimator $\widehat{\text{FDP}}$, we refer the interested reader to [25].

3. 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. Finally, 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.

3.1. 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* [39] 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

$$\hat{\boldsymbol{\beta}}_k(\lambda_k(T,L)) = \arg\min_{\boldsymbol{\beta}_k} \frac{1}{2}\left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}_k\boldsymbol{\beta}_k\right\|_2^2 + \lambda_k(T,L)\cdot\|\boldsymbol{\beta}_k\|_1, \tag{3}$$

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. The predictors in $\boldsymbol{X}$ are standardized and the response vector $\boldsymbol{y}$ is centered. 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.,

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0, \quad \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}^2 = 1,$$

$j = 1, \dots, p$. With these definitions in place we can formulate the following theorem:

Theorem 1 (*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*

$$\left|\boldsymbol{\Phi}_{T,L}(j) - \boldsymbol{\Phi}_{T,L}(j')\right| \leq \bar{\Lambda}\|\boldsymbol{y}\|_2 \cdot \sqrt{2(1-\rho_{j,j'})},$$

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

Proof. The proof is deferred to Appendix. □

3.2. Proposed: The dependency-aware T-Rex selector

From Theorem 1, 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 $\boldsymbol{\Phi}_{T,L}(j)$ by the dependency-aware relative occurrences $\boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}})$, $j=1,\dots,p$, which are defined as follows:

Definition 1 (*Dependency-Aware Relative Occurrences*)**.** The dependency-aware relative occurrence of variable $j\in\{1,\dots,p\}$ is defined by

$$\boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}}) := \Psi_{T,L}(j,\rho_{\mathrm{thr}})\cdot\boldsymbol{\Phi}_{T,L}(j), \tag{4}$$

where

$$\Psi_{T,L}(j,\rho_{\mathrm{thr}}) := \begin{cases} \dfrac{1}{2-\min\limits_{j'\in \mathrm{Gr}(j,\rho_{\mathrm{thr}})}\left\{\left|\boldsymbol{\Phi}_{T,L}(j)-\boldsymbol{\Phi}_{T,L}(j')\right|\right\}}, & \mathrm{Gr}(j,\rho_{\mathrm{thr}})\neq\emptyset \\ 1/2, & \mathrm{Gr}(j,\rho_{\mathrm{thr}})=\emptyset, \end{cases}$$

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

$$\mathrm{Gr}(j,\rho_{\mathrm{thr}})\subseteq\{1,\dots,p\}\backslash\{j\} \tag{5}$$

is the generic definition of the group of variables that are associated with variable j, and $\rho_{\mathrm{thr}}\in[0,1]$ is 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 $\mathrm{Gr}(j,\rho_{\mathrm{thr}})$. Note that, in Section 3.3, the generic definition of ρ_{thr} in Definition 1 will be specified as the cutoff of a dendrogram. That is, the value ρ_{thr} determines explicitly the number of the disjoint variable groups in the dendrogram and, therefore, also implicitly the size of the variable groups.

It is important to note that in the context of the *T-Rex* selector [25], when two variables exhibit very high correlation, even very high relative occurrences do not provide strong evidence for any of these variables being active. Consequently, to maintain FDR control, the dependency-aware *T-Rex* selector conservatively avoids selecting any variables from such highly correlated groups of variables with almost identical relative occurrences. This conservative approach ensures that spurious selections are minimized, thereby preserving the integrity of FDR control. Specifically, as reflected in Eq. (4), variables that fall within the same group and share identical (high or low) relative occurrences are assigned penalty terms of 0.5. This penalty effectively disqualifies these variables from selection, thereby meeting the stringent criteria necessary to uphold FDR control in the presence of highly correlated variables.

From Definition 1, 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 (2):

Corollary 1. *Let* $\widehat{\mathcal{A}}_L(v,T) := \{j : \boldsymbol{\Phi}_{T,L}(j) > v\}$ *and* $\widehat{\mathcal{A}}_L(v,\rho_{\mathrm{thr}},T) := \{j : \boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}}) > v\}$ *be the selected active sets of the ordinary T-Rex selector and the T-Rex+DA selector, respectively. Then, it holds that*

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

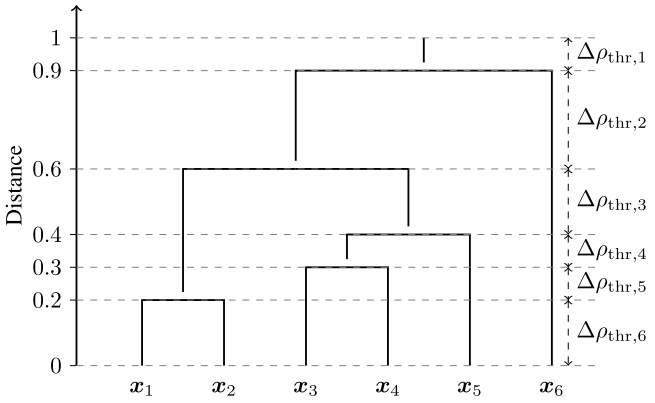

Fig. 2. Hierarchical graphical models: The dendrogram.

Proof. Using the definition of $\boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}})$ in (4), we obtain

$$\begin{aligned}\widehat{\mathcal{A}}_L(v,\rho_{\mathrm{thr}},T) &= \{j : \Psi_{T,L}(j,\rho_{\mathrm{thr}})\cdot\boldsymbol{\Phi}_{T,L}(j) > v\} \\ &\subseteq \{j : \boldsymbol{\Phi}_{T,L}(j) > v\} \\ &= \widehat{\mathcal{A}}_L(v,T),\end{aligned}$$

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

Loosely speaking, Corollary 1 indicates that the effect of replacing $\boldsymbol{\Phi}_{T,L}(j)$ by $\boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}})$ 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 $\mathrm{Gr}(j,\rho_{\mathrm{thr}})$ in (5) has to be specified. Therefore, this work develops a rigorous methodology for the design of $\mathrm{Gr}(j,\rho_{\mathrm{thr}})$ 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.

3.3. Clustering variables via hierarchical graphical models

In the following, we specify $\mathrm{Gr}(j,\rho_{\mathrm{thr}})$, $j=1,\dots,p$, using a hierarchical graphical model. That is, the variables $\boldsymbol{x}_1,\dots,\boldsymbol{x}_p$ 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. Fig. 2 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 denoted by $\Delta\rho_{\mathrm{thr},u}$, $u=1,\dots,p$. For example, cutting off the dendrogram in Fig. 2 at a distance of

$$\begin{aligned}1-\rho_{\mathrm{thr}}(u_{\mathrm{c}}=2) &:= 1-\sum_{u=1}^{u_{\mathrm{c}}=2}\Delta\rho_{\mathrm{thr},u} \\ &= 1-(0.1+0.3)=0.6,\end{aligned} \tag{6}$$

where $u_{\mathrm{c}}\in\{1,\dots,p\}$ 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 $\{\boldsymbol{x}_6\}$. Note that $\rho_{\mathrm{thr}}(u_{\mathrm{c}}) := \sum_{u=1}^{u_{\mathrm{c}}}\Delta\rho_{\mathrm{thr},u}$ is the sum of the discrete increments $\Delta\rho_{\mathrm{thr},u}$ of the dendrogram, as depicted in Fig. 2.

With this generic description of hierarchical graphical models in place, we can specify the generic definition of the variable groups in (5) in a recursive fashion:

Definition 2 (*Hierarchical Group Design*)**.** The jth variable group following a hierarchical graphical model (i.e., binary tree/dendrogram) is defined by

$$\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) := \big\{j' \in \{1,\dots,p\}\backslash\{j\} \,:\, \mathrm{dist}_{u_{\mathrm{c}}-1}(j,j') \in [1-\rho_{\mathrm{thr}}(u_{\mathrm{c}}), 1-\rho_{\mathrm{thr}}(u_{\mathrm{c}}-1)]\big\}, \tag{7}$$

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

In this recursive definition of the variable groups, we consider $\rho_{\mathrm{thr}}(u_{\mathrm{c}})$ to be a variable that can be optimized and, therefore, include it in $\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ as a second argument. Note that the hierarchical clustering as depicted in Fig. 2 follows a bottom-up approach and the group design in Definition 2 is defined top-down. As we move down the dendrogram, the variable groups become more numerous but smaller, resulting in less stringent penalty factors when the variable groups are non-empty. That is, our approach combines bottom-up clustering with a top-down loosening of the penalty factors.

Remark 1. The following three distance measures are frequently used in hierarchical graphical models [42]:

1. Single linkage:

$$\mathrm{dist}_{u_{\mathrm{c}}}(g,h) := \min_{\substack{g'\in \mathrm{Gr}(g,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\\ h'\in \mathrm{Gr}(h,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}} 1-|\rho_{g',h'}|,$$

2. Complete linkage:

$$\mathrm{dist}_{u_{\mathrm{c}}}(g,h) := \max_{\substack{g'\in \mathrm{Gr}(g,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\\ h'\in \mathrm{Gr}(h,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}} 1-|\rho_{g',h'}|,$$

3. Average linkage:

$$\mathrm{dist}_{u_{\mathrm{c}}}(g,h) := \frac{\sum_{g'\in \mathrm{Gr}(g,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}\sum_{h'\in \mathrm{Gr}(h,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}\big(1-|\rho_{g',h'}|\big)}{|\mathrm{Gr}(g,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))|\cdot|\mathrm{Gr}(h,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))|}.$$

Remark 2. Note that, for all $u_{\mathrm{c}} \in \{1,\dots,p\}$, it holds that

$$\mathrm{Gr}(j_1,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) \cap \mathrm{Gr}(j_2,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) = \begin{cases} \emptyset, & \nexists\, j\in\{1,\dots,p\} \,:\, j_1,j_2 \in \mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\\ \mathrm{Gr}(j_1,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) \cup \mathrm{Gr}(j_2,\rho_{\mathrm{thr}}(u_{\mathrm{c}})), & \text{otherwise,}\end{cases}$$

i.e., any arbitrary pair of groups among the obtained p groups $\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$, $j=1,\dots,p$, 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.

3.4. Preliminaries for the FDR control theorem

The preliminaries and technical tools presented in this section draw upon the established methodologies and derivations detailed in [25]. By leveraging these foundational techniques, we aim to address the intricacies introduced by dependent variables, thereby extending the applicability of these tools to the proposed dependency-aware T-Rex selector.

Based on the recursive definition of the variable groups in (7), we can formulate the FDR, TPR, and the conservative FDP estimator $\widehat{\mathrm{FDP}}$. For this purpose, let

$$\begin{aligned} V_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) &:= \big|\widehat{\mathcal{A}}^0(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\big| \\ &:= \big|\{\text{null } j \,:\, \boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) > v\}\big|, \\ S_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) &:= \big|\widehat{\mathcal{A}}^1(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\big| \\ &:= \big|\{\text{active } j \,:\, \boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) > v\}\big|, \\ R_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) &:= \big|\widehat{\mathcal{A}}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\big| \\ &:= \big|\{j \,:\, \boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) > v\}\big|, \end{aligned} \tag{8}$$

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 $\widehat{\mathcal{A}}^0(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$, $\widehat{\mathcal{A}}^1(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$, and $\widehat{\mathcal{A}}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ are shortcuts (i.e., L and T are dropped) of the expressions $\widehat{\mathcal{A}}^0_L(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T)$, $\widehat{\mathcal{A}}^1_L(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T)$, and $\widehat{\mathcal{A}}_L(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T)$, respectively.

Definition 3 (*Dependency-Aware FDP and FDR*)**.** The dependency-aware FDR is defined as the expectation of the dependency-aware FDP, i.e.,

$$\begin{aligned} \mathrm{FDR}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T,L) &:= \mathbb{E}\big[\mathrm{FDP}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T,L)\big] \\ &:= \mathbb{E}\left[\frac{V_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}{R_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\vee 1}\right]. \end{aligned}$$

Definition 4 (*Dependency-Aware TPP and TPR*)**.** The dependency-aware TPR is defined as the expectation of the dependency-aware TPP, i.e.,

$$\begin{aligned} \mathrm{TPR}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T,L) &:= \mathbb{E}\big[\mathrm{TPP}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T,L)\big] \\ &:= \mathbb{E}\left[\frac{S_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}{p_1\vee 1}\right]. \end{aligned}$$

From Definition 3, 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 $V_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$, since the total number of selected variables $R_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ is observable. For this purpose, we plug the dependency-aware relative occurrences from Definition 1 and the group design from Definition 2 into the ordinary *T-Rex* estimator of $V_{T,L}(v)$ in [25], which yields the dependency-aware estimator of the number of selected null variables

$$\widehat{V}_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) := \sum_{j\in\widehat{\mathcal{A}}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}\Big(1-\boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\Big) + \widehat{V}'_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})), \tag{9}$$

where

$$\widehat{V}'_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) := \sum_{t=1}^{T}\frac{p-\sum_{q=1}^{p}\boldsymbol{\Phi}^{\mathrm{DA}}_{t,L}(q,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}{L-(t-1)}\cdot\frac{\sum_{j\in\widehat{\mathcal{A}}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}\Delta\boldsymbol{\Phi}^{\mathrm{DA}}_{t,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}{\sum_{j\in\widehat{\mathcal{A}}(0.5,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}\Delta\boldsymbol{\Phi}^{\mathrm{DA}}_{t,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))} \tag{10}$$

and $\Delta\boldsymbol{\Phi}^{\mathrm{DA}}_{t,L}(j,\rho_{\mathrm{thr}}) := \boldsymbol{\Phi}^{\mathrm{DA}}_{t,L}(j,\rho_{\mathrm{thr}}) - \boldsymbol{\Phi}^{\mathrm{DA}}_{t-1,L}(j,\rho_{\mathrm{thr}})$ is the increase in the dependency-aware relative occurrence from step $t-1$ to t. The expressions in (9) and (10) are derived along the lines of the ordinary estimator of $V_{T,L}(v)$ in [25] except that the ordinary relative occurrences have been replaced by the proposed dependency-aware relative occurrences in Definition 1. Thus, for more details on the derivation of (9) and (10), we refer the interested reader to [25].

Finally, the conservative estimator of the FDP is defined as follows:

Definition 5 (*Dependency-Aware FDP estimator*)**.** The dependency-aware FDP estimator is defined by

$$\widehat{\mathrm{FDP}}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}),T,L) := \frac{\widehat{V}_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}{R_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\vee 1}.$$

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

$$\max_{\upsilon,\rho_{\text{thr}}(u_c),T} R_{T,L}(\upsilon,\rho_{\text{thr}}(u_c)) \quad \text{subject to} \quad \widehat{\text{FDP}}(\upsilon,\rho_{\text{thr}}(u_c),T,L)\le\alpha. \tag{11}$$

In Section 3.5, we prove that satisfying the condition of the optimization problem in Eq. (11)ields FDR control and in Section 4, we propose an efficient algorithm to solve Eq. (11)nd prove that it yields an optimal solution.

3.5. Dependency-aware FDR control

In this section, using martingale theory [37] we state and prove that controlling $\widehat{\text{FDP}}(\upsilon,\rho_{\text{thr}}(u_c),T,L)$ at the target level α (i.e., the condition in the optimization problem in Eq. (11) guarantees FDR control. The required technical Lemmas 1, 2, and 3 are deferred to Appendix.

Theorem 2 (*Dependency-Aware FDR control*)**.** *For all quadruples* $(T,L,\rho_{\text{thr}}(u_c),\upsilon)\in\{1,\dots,L\}\times\mathbb{N}_+\times[0,1]\times[0.5,1)$ *that satisfy the equation*

$$\upsilon=\inf\left\{\nu\in[0.5,1):\widehat{\text{FDP}}(\nu,\rho_{\text{thr}}(u_c),T,L)\le\alpha\right\}, \tag{12}$$

and as $K\to\infty$*, the T-Rex+DA selector with* $\text{Gr}(j,\rho_{\text{thr}}(u_c))$ *from Definition 2 controls the FDR at any fixed target level* $\alpha\in[0,1]$*, i.e.,*

$$\text{FDR}(\upsilon,\rho_{\text{thr}}(u_c),T,L)\le\alpha.$$

Proof. Rewriting the expression for the FDP in Definition 3, we obtain

$$\begin{aligned}\text{FDP}(\upsilon,\rho_{\text{thr}}(u_c),T,L)&=\frac{V_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}{R_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))\vee 1}\\&=\frac{\widehat{V}_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}{R_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))\vee 1}\cdot\frac{V_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}{\widehat{V}_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}\\&\le\alpha\cdot\frac{V_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}{\widehat{V}_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}\\&\le\alpha\cdot\frac{V_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}{\widehat{V}'_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))}\\&=:\alpha\cdot H_{T,L}(\upsilon,\rho_{\text{thr}}(u_c)),\end{aligned}$$

where the inequality in the third line follows from the condition in (12) that all considered quadruples $(T,L,\rho_{\text{thr}}(u_c),\upsilon)$ must satisfy. Taking the expectation of the FDP, we obtain

$$\text{FDR}(\upsilon,\rho_{\text{thr}}(u_c),T,L)\le\alpha\cdot\mathbb{E}\left[H_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))\right]$$

and, thus, it remains to prove that $\mathbb{E}[H_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))]\le 1$. Since $\{H_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))\}_{\upsilon\in\mathcal{V}}$ is a backward-running super-martingale, as stated in Lemma 1, and (12) is a stopping time that is adapted to the filtration $\mathcal{F}_\upsilon$ in Lemma 1 (i.e., υ is $\mathcal{F}_\upsilon$-measurable), we can apply Doob's optional stopping theorem to obtain an upper bound for $\mathbb{E}[H_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))]$, i.e.,

$$\mathbb{E}\left[H_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))\right]\le\mathbb{E}\left[H_{T,L}(0.5,\rho_{\text{thr}}(u_c))\right]. \tag{13}$$

Defining

$$\begin{aligned}\Psi^+_{t,L}(j,\rho_{\text{thr}}(u_c))&:=\begin{cases}\dfrac{1}{2-\min\limits_{j'\in\text{Gr}(j,\rho_{\text{thr}}(u_c))}\left\{\left|\Phi_{T,L}(j)-\Phi_{T,L}(j')\right|\right\}}, & \text{Gr}(j,\rho_{\text{thr}}(u_c))\neq\varnothing\\ 1, & \text{Gr}(j,\rho_{\text{thr}}(u_c))=\varnothing\end{cases}\\&\ge\begin{cases}\dfrac{1}{2-\min\limits_{j'\in\text{Gr}(j,\rho_{\text{thr}}(u_c))}\left\{\left|\Phi_{T,L}(j)-\Phi_{T,L}(j')\right|\right\}}, & \text{Gr}(j,\rho_{\text{thr}}(u_c))\neq\varnothing\\ 1/2, & \text{Gr}(j,\rho_{\text{thr}}(u_c))=\varnothing\end{cases}\\&=\Psi_{t,L}(j,\rho_{\text{thr}}(u_c)),\end{aligned}$$

$(t,L)\in\{1,\dots,T\}\times\mathbb{N}_+$, $H_{T,L}(0.5,\rho_{\text{thr}}(u_c))$ can be upper bounded as follows:

$$\begin{aligned}H_{T,L}(0.5,\rho_{\text{thr}}(u_c))&=\frac{V_{T,L}(0.5,\rho_{\text{thr}}(u_c))}{\widehat{V}'_{T,L}(0.5,\rho_{\text{thr}}(u_c))}\\&\le\frac{V^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))}{\widehat{V}'^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))}\\&=:H^+_{T,L}(0.5,\rho_{\text{thr}}(u_c)).\end{aligned} \tag{14}$$

The inequality in the second line follows from

$$\begin{aligned}&\text{(i) } V^+_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))\\&\quad:=|\{\text{null } j:\Psi^+_{T,L}(j,\rho_{\text{thr}}(u_c))\cdot\Phi_{T,L}(j)>\upsilon\}|\\&\quad\ge|\{\text{null } j:\Psi_{T,L}(j,\rho_{\text{thr}}(u_c))\cdot\Phi_{T,L}(j)>\upsilon\}|\\&\quad=V_{T,L}(\upsilon,\rho_{\text{thr}}(u_c)),\end{aligned} \tag{15}$$

$$\begin{aligned}&\text{(ii) } \widehat{V}'^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))\\&\quad:=\sum_{t=1}^{T}\frac{p-\sum_{q=1}^{p}\Psi^+_{t,L}(q,\rho_{\text{thr}}(u_c))\cdot\Phi_{t,L}(q)}{L-(t-1)}\\&\quad\le\sum_{t=1}^{T}\frac{p-\sum_{q=1}^{p}\Psi_{t,L}(q,\rho_{\text{thr}}(u_c))\cdot\Phi_{t,L}(q)}{L-(t-1)}\\&\quad=\widehat{V}'_{T,L}(0.5,\rho_{\text{thr}}(u_c)).\end{aligned} \tag{16}$$

Next, we show that $H^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))$ is monotonically increasing in $\rho_{\text{thr}}(u_c)$, i.e.,

$$\begin{aligned}H^+_{T,L}(0.5,\rho_{\text{thr}}(u_c+1))&=\frac{V^+_{T,L}(0.5,\rho_{\text{thr}}(u_c+1))}{\widehat{V}'^+_{T,L}(0.5,\rho_{\text{thr}}(u_c+1))}\\&\ge\frac{V^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))}{\widehat{V}'^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))}\\&=H^+_{T,L}(0.5,\rho_{\text{thr}}(u_c)),\end{aligned} \tag{17}$$

where the inequality in the second line follows from Lemmas 2 and 3. Combining these preliminaries and noting that $\Psi^+_{T,L}(j,\rho_{\text{thr}}(u_c))=1$ yields

$$\begin{aligned}V_{T,L}(0.5)&:=|\{\text{null } j:\Phi_{T,L}(j)>0.5\}|,\\\widehat{V}'_{T,L}(0.5)&:=\sum_{t=1}^{T}\frac{p-\sum_{q=1}^{p}\Phi_{t,L}(q)}{L-(t-1)},\\H_{T,L}(0.5)&:=\frac{V_{T,L}(0.5)}{\widehat{V}'_{T,L}(0.5)},\end{aligned} \tag{18}$$

i.e., the counterparts of the dependency-aware expressions $V^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))$, $V'^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))$, and $H^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))$, respectively, we finally obtain

$$\begin{aligned}\mathbb{E}\left[H_{T,L}(\upsilon,\rho_{\text{thr}}(u_c))\right]&\le\mathbb{E}\left[H_{T,L}(0.5,\rho_{\text{thr}}(u_c))\right]\\&\le\mathbb{E}\left[H^+_{T,L}(0.5,\rho_{\text{thr}}(u_c))\right]\\&\le\mathbb{E}\left[H^+_{T,L}(0.5,\rho_{\text{thr}}(p))\right]\\&=\mathbb{E}\left[H^+_{T,L}(0.5,1)\right]\\&=\mathbb{E}\left[H_{T,L}(0.5)\right]\\&\le 1,\end{aligned} \tag{19}$$

where the first inequality is the same as in (13), the second and third inequalities follow from (14) and (17), respectively, and the equation in the fourth line follows from (6) (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 $j\in\{1,\dots,p\}$, and, therefore, it holds that $\Psi^+_{t,L}(j,1)=1$ for all $j\in\{1,\dots,p\}$. Thus, $H^+_{T,L}(0.5,1)$ boils down to its ordinary counterpart $H_{T,L}(0.5)$ in (18), i.e., $H^+_{T,L}(0.5,1)=H_{T,L}(0.5)$. The proof of Inequality (19) requires $K\to\infty$. It is omitted because it is exactly as in the proof of Theorem 1 (FDR control) in [25]. □

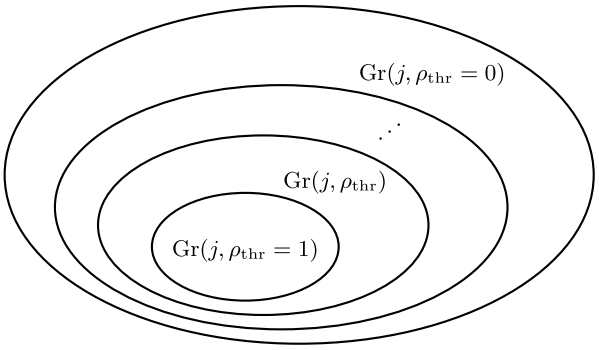

Fig. 3. Illustration of the group design principle for dependency-aware FDR control in Theorem 3.

Remark 3. The condition $K \to \infty$ in Theorem 2 is required to use the law of large numbers in the proof of the FDR control property (for details, see [25]). This condition ensures that the average behavior of the random variable, which follows the negative hypergeometric distribution and models the number of included null variables before terminating the random experiments, stabilizes. It has already been shown that this convergence to the mean of the negative hypergeometric distribution is very fast, i.e., the mean stabilization has already been verified for $K = 20$ in Appendices G and J of [25]. We have added further verifications in Figure 11, deferred to Appendix C in the supplementary materials, for $K = 5, 20, 300$, demonstrating that there is no substantial improvement beyond $K = 20$ random experiments.

3.6. General group design principle

The *T-Rex+DA* selector is not only suitable for the considered binary tree graphs or dendrograms but also for various other dependency models. In fact, a closer look at Lemmas 2 and 3, which are essential for the proof of Theorem 2 (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 2 and 3:

Theorem 3 (*Group Design Principle*)**.** *Consider the generic definition of the variable groups in Definition 1, 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*

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

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

Proof. The FDR control property in Theorem 2 holds if Lemmas 2 and 3 hold. Lemmas 2 and 3 hold for any definition of $\mathrm{Gr}(j, \rho_{\mathrm{thr}})$ that satisfies the group design principle in Theorem 3. □

Loosely speaking, Theorem 3 states that the cardinalities of any variable group j must be monotonically decreasing in ρ_{thr} and follow the subset structure illustrated in Fig. 3. Thus, any dependency model (e.g., graph models, time series models, equicorrelated models, etc.) that follows the design principle in Theorem 3 can be incorporated into the *T-Rex+DA* selector. This property makes the *T-Rex+DA* selector a versatile method that can cope with various dependency models.

4. Optimal dependency-aware T-Rex algorithm

In this section, we propose an efficient calibration algorithm and prove that it yields an optimal solution of Eq. (11) That is, it optimally calibrates the parameters υ, $\rho_{\mathrm{thr}}(u_{\mathrm{c}})$, 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

Algorithm 1 Extended *T-Rex+DA* Calibration.

1. **Input**: $\alpha \in [0, 1]$, $\boldsymbol{X}$, $\boldsymbol{y}$, K, $\tilde{\upsilon}$, $\tilde{\rho}_{\mathrm{thr}}$, $L_{\max}$, $T_{\max}$.
2. **Set** $L = p$, $T = 1$.
3. **While** $\widehat{\mathrm{FDP}}(\upsilon = \tilde{\upsilon}, \rho_{\mathrm{thr}}(u_{\mathrm{c}}) = \tilde{\rho}_{\mathrm{thr}}, T, L) > \alpha$ and $L \le L_{\max}$ **do**:

   Set $L \leftarrow L + p$.
4. **Set** $\Delta\upsilon = \frac{1}{K}$, $\widehat{\mathrm{FDP}}(\upsilon = 1 - \Delta\upsilon, \rho_{\mathrm{thr}}(u_{\mathrm{c}}) = \tilde{\rho}_{\mathrm{thr}}, T, L) = 0$.
5. **While** $\widehat{\mathrm{FDP}}(\upsilon = 1 - \Delta\upsilon, \rho_{\mathrm{thr}}(u_{\mathrm{c}}) = \tilde{\rho}_{\mathrm{thr}}, T, L) \le \alpha$ and $T \le T_{\max}$ **do**:

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

   5.1.1. **For** $u_{\mathrm{c}} = 1, \dots, p$ **do**:

   i. **Compute** $\widehat{\mathrm{FDP}}(\upsilon, \rho_{\mathrm{thr}}(u_{\mathrm{c}}), T, L)$ as in Definition 5.
   ii. **If** $\widehat{\mathrm{FDP}}(\upsilon, \rho_{\mathrm{thr}}(u_{\mathrm{c}}), T, L) \le \alpha$

   Compute $\widehat{\mathcal{A}}_L(\upsilon, \rho_{\mathrm{thr}}(u_{\mathrm{c}}), T)$ as in (20).

   Else

   Set $\widehat{\mathcal{A}}_L(\upsilon, \rho_{\mathrm{thr}}(u_{\mathrm{c}}), T) = \emptyset$.

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

$$\max_{\upsilon', \rho_{\mathrm{thr}}(u'_{\mathrm{c}}), T'} \left| \widehat{\mathcal{A}}_L(\upsilon', \rho_{\mathrm{thr}}(u'_{\mathrm{c}}), T') \right|$$

   subject to $T' \in \{1, \dots, T - 1\}$

$$u'_{\mathrm{c}} \in \{1, \dots, p\}$$

$$\upsilon' \in \{0.5, 0.5 + \Delta\upsilon, 0.5 + 2\Delta\upsilon, \dots, 1 - \Delta\upsilon\}$$

   and let $(\upsilon^*, \rho_{\mathrm{thr}}(u^*_{\mathrm{c}}), T^*)$ be a solution.
7. **Output**: $(\upsilon^*, \rho_{\mathrm{thr}}(u^*_{\mathrm{c}}), T^*, L)$ and $\widehat{\mathcal{A}}_L(\upsilon^*, \rho_{\mathrm{thr}}(u^*_{\mathrm{c}}), T^*)$.

1. An open source implementation of the proposed calibration algorithm for the *T-Rex+DA* selector is available within the R packages 'TRexSelector' [43] and 'tlars' [44] on CRAN.

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 paper, the hyperparameters are chosen based on the suggestions in [25] as follows: $\tilde{\upsilon} = 0.75$, $\tilde{\rho}_{\mathrm{thr}} = \rho_{\mathrm{thr}}(\lfloor 0.75 \cdot p \rfloor)$, $L_{\max} = 10p$, $T_{\max} = \lceil n/2 \rceil$, where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote rounding towards the nearest integer and the nearest higher integer, respectively. Moreover, it was shown for various applications and in extensive simulations that there are no improvements for $K > 20$ random experiments [25–27] 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 $\widehat{\mathrm{FDP}}(\upsilon = \tilde{\upsilon}, \rho_{\mathrm{thr}}(u_{\mathrm{c}}) = \tilde{\rho}_{\mathrm{thr}}, T = 1, L)$ falls below the target FDR level α at a reference point $(\upsilon, \rho_{\mathrm{thr}}(u_{\mathrm{c}}), T) = (\tilde{\upsilon}, \tilde{\rho}_{\mathrm{thr}}, 1)$ 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 $\widehat{\text{FDP}}(\upsilon = 1-\Delta\upsilon, \rho_{\text{thr}}(u_c) = \tilde{\rho}'_{\text{thr}}, T, L) < \widehat{\text{FDP}}(\upsilon = 1-\Delta\upsilon, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T, L)$, the algorithm proceeds by increasing the number of included dummies T at a reference point $(\upsilon, \rho_{\text{thr}}(u_c)) = (1-\Delta\upsilon, \tilde{\rho}_{\text{thr}})$, where $\Delta\upsilon = 1/K$, until $\widehat{\text{FDP}}(\upsilon = 1-\Delta\upsilon, \rho_{\text{thr}}(u_c) = \tilde{\rho}_{\text{thr}}, T, L)$ 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 υ and $\rho_{\text{thr}}(u_c)$ that maximize the number of selected variables by solving the optimization problem in 6. Finally, the obtained solution $(\upsilon^*, \rho_{\text{thr}}(u_c^*), T^*, L)$ yields the FDR-controlled set of selected variables

$$\widehat{\mathcal{A}}_L(\upsilon^*, \rho_{\text{thr}}(u_c^*), T^*) = \left\{ j : \Phi^{\text{DA}}_{T^*,L}(j, \rho_{\text{thr}}(u_c^*)) > \upsilon^* \right\}. \tag{20}$$

Supposing that the dendrogram for the hierarchical group design in Definition 2 is pre-computed, the computational complexity of the *T-Rex+DA* selector in Algorithm 1 stems from the computation of K early terminated random experiments with expected complexity $\mathcal{O}(np)$, which is the same complexity as that of the ordinary *T-Rex* selector in [25]. The computational complexity of the dendrogram depends on the algorithm and linkage type that is used to compute it (see [42] for an overview).

In the following, we state and prove that Algorithm 1 yields an optimal solution of Eq. (11)

Theorem 4 (*Optimal Dependency-Aware Calibration*)**.** *Suppose that L, as obtained by Algorithm 1, is fixed. Then, any triple $(\upsilon^*, \rho_{\text{thr}}(u_c^*), T^*)$ of a quadruple $(\upsilon^*, \rho_{\text{thr}}(u_c^*), T^*, L)$, as obtained by Algorithm 1, is an optimal solution of Eq. (11)*

Proof. From Eq. (8), it follows that for all quadruples $(\upsilon, \rho_{\text{thr}}(u_c), T, L)$ that satisfy $\widehat{\text{FDP}}(\upsilon, \rho_{\text{thr}}(u_c), T, L) \leq \alpha$, the objective functions in Step 6 of Algorithm 1 and in the optimization problem in Eq. (11)re equal, i.e., $|\widehat{\mathcal{A}}_L(\upsilon', \rho_{\text{thr}}(u'_c), T')| = R_{T',L}(\upsilon', \rho_{\text{thr}}(u'_c))$. Therefore, and since all attainable values of u_c are considered in Step 6 of Algorithm 1, it suffices to show that for fixed $\rho_{\text{thr}}(u_c)$ and L, the set of feasible tuples (υ, T) of Eq. (11)s a subset of or equal to the set of feasible tuples obtained by Algorithm 1. Since, ceteris paribus, $\widehat{\text{FDP}}(\upsilon, \rho_{\text{thr}}(u_c), T, L)$ is monotonically decreasing in υ, and monotonically increasing in T [25], for $T = T_{\text{fin}}$, $\widehat{\text{FDP}}(\upsilon, \rho_{\text{thr}}(u_c), T, L)$ attains its minimum value at $(\upsilon, T) = (1-\Delta\upsilon, T_{\text{fin}})$, where $T_{\text{fin}} \in \{1, \dots, L\}$ is implicitly defined through the inequalities

$$\widehat{\text{FDP}}(1-\Delta\upsilon, \rho_{\text{thr}}(u_c), T_{\text{fin}}, L) \leq \alpha$$

and

$$\widehat{\text{FDP}}(1-\Delta\upsilon, \rho_{\text{thr}}(u_c), T_{\text{fin}}+1, L) > \alpha.$$

Thus, the feasible set of the optimization problem in Eq. (11)s given by

$$\begin{aligned}\left\{(\upsilon, T) : \widehat{\text{FDP}}(\upsilon, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\right\} = \{(\upsilon, T) : \; & \upsilon \in [0.5, 1-\Delta\upsilon], \\ & T \in \{1, \dots, T_{\text{fin}}\}, \\ & \widehat{\text{FDP}}(\upsilon, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\Big\}.\end{aligned} \tag{21}$$

Since $\Delta\upsilon = 1/K$, the upper endpoint of the interval $[0.5, 1-\Delta\upsilon]$ asymptotically (i.e., $K \to \infty$) coincides with the supremum of the interval $[0.5, 1)$. That is, the set in (21) contains all feasible solutions of Eq. (11) However, since the υ-grid in Algorithm 1 is, as in [25], adapted to K, all values of $R_{T,L}(\upsilon, \rho_{\text{thr}}(u_c))$ that are attained by off-grid solutions can also be attained by on-grid solutions. Thus, instead of (21) only the following fully discrete feasible set of Eq. (11)eeds to be considered:

$$\begin{aligned}\Big\{(\upsilon, T) : \; & \upsilon \in \{0.5, 0.5+\Delta\upsilon, 0.5+2\Delta\upsilon, \dots, 1-\Delta\upsilon\}, \\ & T \in \{1, \dots, T_{\text{fin}}\}, \\ & \widehat{\text{FDP}}(\upsilon, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\Big\}.\end{aligned} \tag{22}$$

Since the "while"-loop in Step 5 of Algorithm 1 is terminated when $T = T_{\text{fin}} + 1$, the feasible set of the optimization problem in Step 6 of Algorithm 1 is given by

$$\begin{aligned}\Big\{(\upsilon, T) : \; & \upsilon \in \{0.5, 0.5+\Delta\upsilon, 0.5+2\Delta\upsilon, \dots, 1-\Delta\upsilon\}, \\ & T \in \{1, \dots, T_{\text{fin}}\} \\ & \widehat{\text{FDP}}(\upsilon, \rho_{\text{thr}}(u_c), T, L) \leq \alpha\Big\},\end{aligned}$$

which is equal to (22). □

Remark 4. Note that in contrast to Theorem 2, where $K \to \infty$ is required to prove that the FDR (i.e., the expected value of the FDP) does not exceed the user-specified target FDR level α, Theorem 4 does not require $K \to \infty$ because the condition $\widehat{\text{FDP}} \leq \alpha$ in Eq. (11)nly states that the conservative FDP estimator (and not its expected value) must not exceed α. Therefore, $K \to \infty$ is not a necessary condition to prove the optimality of Algorithm 1.

5. 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 [21], *model-X* knockoff+ [21], and *T-Rex* selector [25].

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

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

That is, each block mimics a dependency structure that is often present in biomedical data (e.g., gene expression [32] and genomics data [33]) 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 $\boldsymbol{\beta} = [\beta_1 \dots \beta_p]^\top \in \mathbb{R}^p$ 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.

The results in Figs. 4 and 5 are averaged over 955 Monte Carlo replications.[5] 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

[5] 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.

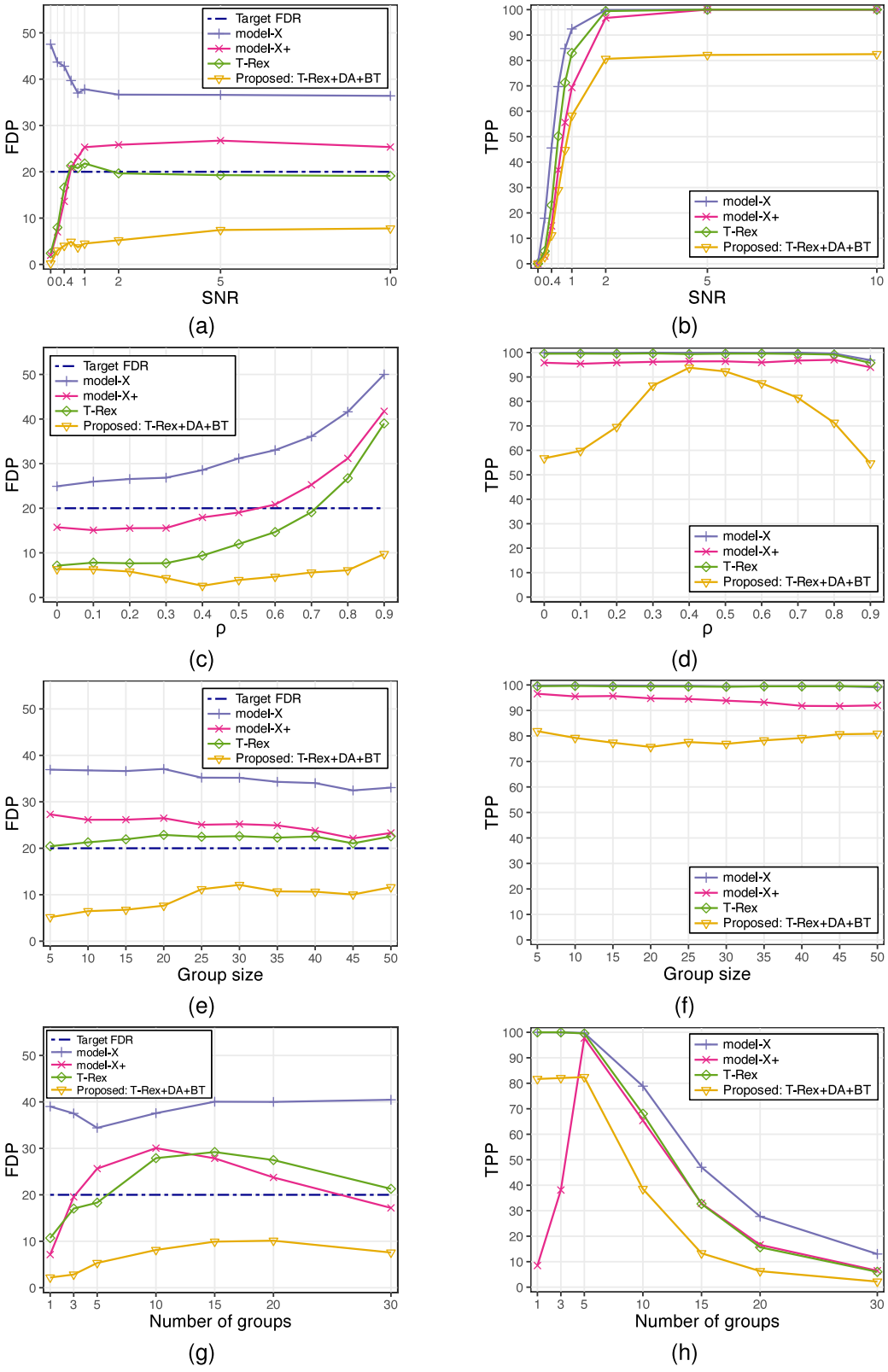

Fig. 4. Only the proposed *T-Rex+DA* selector with a binary tree group model (*T-Rex+DA+BT*) reliably controls the FDR in all settings while achieving a reasonably high TPR. In Figure (c), we see that with increasing correlations among the variables in a group, the benchmark methods exhibit an alarming increase in FDR.

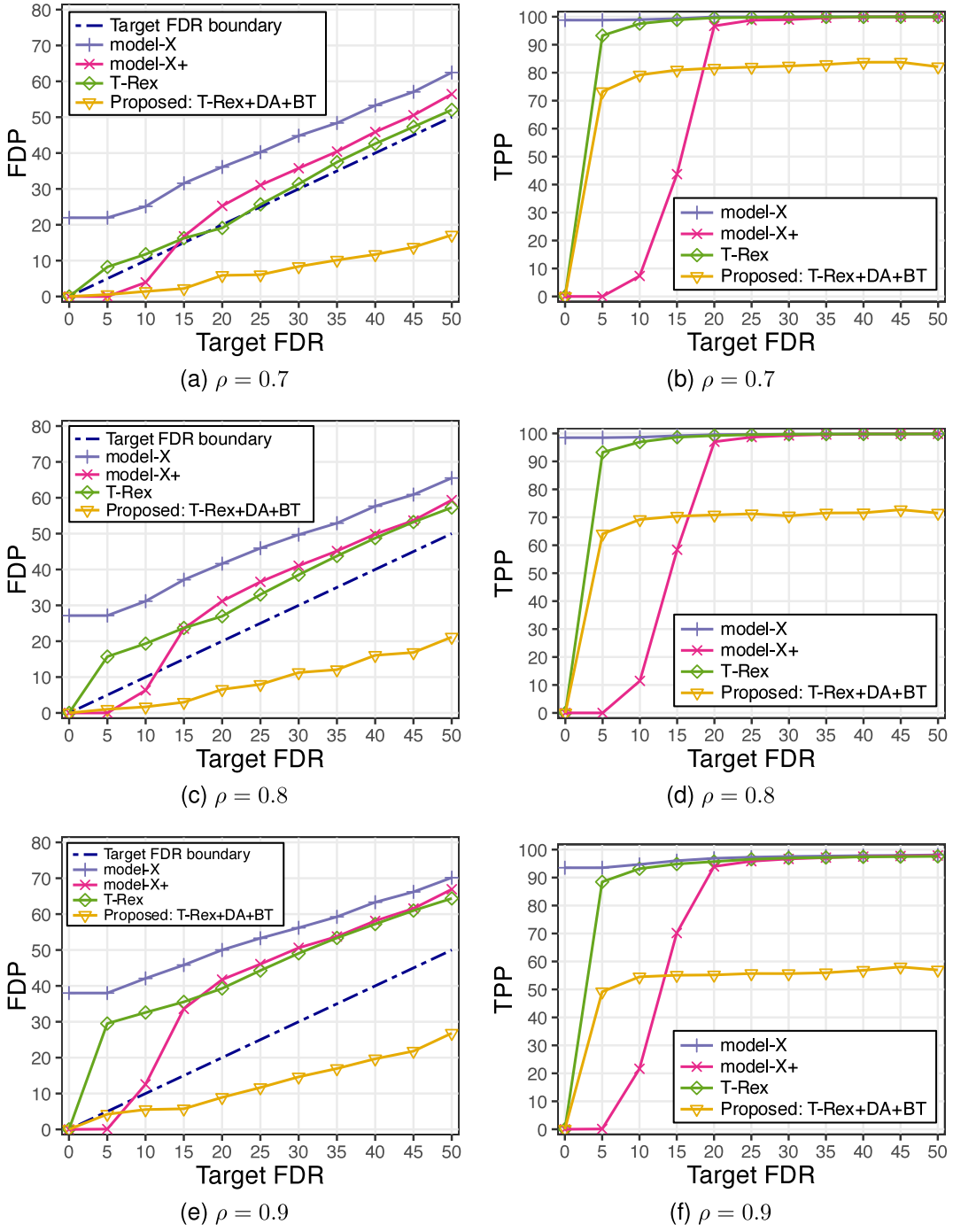

Fig. 5. The proposed *T-Rex+DA* selector reliably controls the FDR in all settings while achieving a reasonably high TPR in harsh high correlation settings. We observe that with increasing correlations among the variables in a group, the benchmark methods do not control the FDR for almost any choice of target FDR.

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 the unfortunately still ongoing reproducibility crisis in many scientific fields [45].

As discussed in Section 3.2, the conservative nature of the proposed dependency-aware *T-Rex* selector results in a smaller set of selected variables as compared to the original *T-Rex* selector [25], particularly in the presence of highly correlated groups. This conservative behavior directly impacts the observed FDR in the simulations, which may appear far from the desired level because the method prioritizes maintaining control over spurious discoveries. In such cases, variables from highly correlated groups are penalized, effectively reducing false discoveries and maintaining the FDR control guarantees. That is, the observed FDR deviation is a result of the stringent penalty mechanism employed to preserve control in challenging high-dependency scenarios.

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,

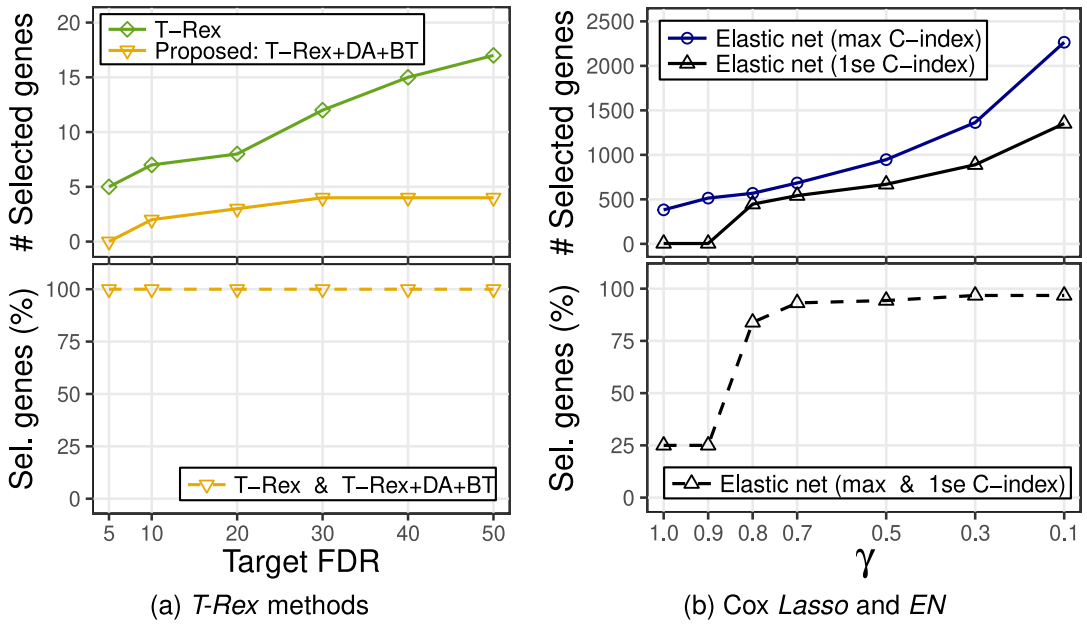

Fig. 6. Number of selected genes in the TCGA breast cancer survival analysis study.

(2) the noise vector ϵ is sampled from a heavy-tailed Student-t distribution with 3 degrees of freedom,

are deferred to Appendix C in the supplementary materials. 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.

6. Breast cancer survival analysis

To illustrate the practical usefulness of the proposed *T-Rex+DA* selector in large-scale high-dimensional settings, we demonstrate the performance of the proposed method via a breast cancer survival analysis using gene expression and survival time data from the open source resource The Cancer Genome Atlas (TCGA) [12,46]. 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.

6.1. 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 [47], $p = 19{,}405$ 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 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 [48].

6.2. Methods and results

As benchmark methods, we consider the Cox proportional hazards *Lasso* and *elastic net* [49], which are specifically designed for censored survival data, and the ordinary *T-Rex* selector [25]. 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 [49], 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, e.g., Fig. 1 in [25]), it is practically infeasible in this large-scale high-dimensional setting. Therefore, we cannot consider it in this survival analysis.

Fig. 6 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 proposed dependency-aware *T-Rex* selector. In accordance with Corollary 1, all genes that were selected by the proposed method 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 [50–52].

7. Conclusion

The dependency-aware *T-Rex* 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. A real world TCGA breast cancer survival analysis showed 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 other fields that require large-scale high-dimensional variable selection with FDR-control guarantees. In fact, the group design principle was already successfully applied as a guiding principle in adapting the proposed *T-Rex+DA* selector for FDR-controlled sparse financial index tracking [34,53].

CRediT authorship contribution statement

Jasin Machkour: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michael Muma:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Daniel P. Palomar:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Proofs and technical lemmas

A.1. Proof of Theorem 1

Proof. Note that for any $\hat{\beta}_{j,k} \neq 0$, the indicator function in (1) can be written as follows:

$$\mathbb{1}_k(j,T,L) = \left|\operatorname{sign}\left(\hat{\beta}_{j,k}\right)\right|, \quad j = 1,\dots,p. \tag{23}$$

Thus, we can rewrite the left-hand side of the inequality in Theorem 1 as follows:

$$\begin{aligned}
&\frac{\left|\boldsymbol{\Phi}_{T,L}(j) - \boldsymbol{\Phi}_{T,L}(j')\right|}{\|\boldsymbol{y}\|_2} \\
&= \frac{1}{\|\boldsymbol{y}\|_2 \cdot K}\left|\sum_{k=1}^{K}\left(\mathbb{1}_k(j,T,L) - \mathbb{1}_k(j',T,L)\right)\right| \\
&\leq \frac{1}{\|\boldsymbol{y}\|_2 \cdot K}\sum_{k=1}^{K}\left|\mathbb{1}_k(j,T,L) - \mathbb{1}_k(j',T,L)\right| \\
&= \frac{1}{\|\boldsymbol{y}\|_2 \cdot K}\sum_{k=1}^{K}\left|\left|\operatorname{sign}\left(\hat{\beta}_{j,k}\right)\right| - \left|\operatorname{sign}\left(\hat{\beta}_{j',k}\right)\right|\right| \\
&\leq \sqrt{2(1-\rho_{j,j'})}\cdot\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\lambda_k(T,L)}\cdot\frac{\|\hat{\boldsymbol{r}}_k\|_2}{\|\boldsymbol{y}\|_2} \\
&\leq \sqrt{2(1-\rho_{j,j'})}\cdot\bar{\Lambda},
\end{aligned}$$

where $\bar{\Lambda} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\lambda_k(T,L)}$. The equation in the second line follows from the definition of the relative occurrences in (1), the inequality in the third line is a consequence of the triangle inequality, the equation in the fourth line follows from (23), and the inequality in the fifth line follows from Lemma 4, which is deferred to Appendix B in the supplementary materials. The inequality in the last line holds since $\hat{\boldsymbol{\beta}}_k$ is by definition the minimizer of (25) and, therefore,

$$\mathcal{L}\left(\boldsymbol{\beta}_k = \hat{\boldsymbol{\beta}}_k, \lambda_k(T,L)\right) \leq \mathcal{L}\left(\boldsymbol{\beta}_k = \boldsymbol{0}, \lambda_k(T,L)\right)$$

and, equivalently,

$$\frac{1}{2}\|\hat{\boldsymbol{r}}_k\|_2^2 + \lambda_k(T,L)\|\hat{\boldsymbol{\beta}}_k\|_1 \leq \frac{1}{2}\|\boldsymbol{y}\|_2^2,$$

which yields $\|\hat{\boldsymbol{r}}_k\|_2 \leq \|\boldsymbol{y}\|_2$. Finally, we obtain

$$\left|\boldsymbol{\Phi}_{T,L}(j) - \boldsymbol{\Phi}_{T,L}(j')\right| \leq \bar{\Lambda}\|\boldsymbol{y}\|_2 \cdot \sqrt{2(1-\rho_{j,j'})}.$$

□

A.2. Technical lemmas

Lemma 1. *Define* $\mathcal{V} := \{\boldsymbol{\Phi}^{\mathrm{DA}}_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) \geq 0.5,\ j = 1,\dots,p\}\backslash\{1\}$. *Let*

$$\mathcal{F}_v := \sigma\left(\left\{V_{T,L}(u,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))_{u\geq v}\right\}, \left\{\hat{V}'_{T,L}(u,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))_{u\geq v}\right\}\right)$$

be a backward-filtration with respect to v. *Then, for all triples* $(T,L,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) \in \{1,\dots,L\}\times\mathbb{N}_+\times[0,1]$, $\{H_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\}_{v\in\mathcal{V}}$ *is a backward-running super-martingale with respect to* $\mathcal{F}_v$. *That is,*

$$\mathbb{E}\left[H_{T,L}(v-\epsilon^*_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})))\mid\mathcal{F}_v\right] \geq H_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})),$$

where $v \in [0.5,1)$ *and*

$$\epsilon^*_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) := \inf\left\{\epsilon\in(0,v) :\ R_{T,L}(v-\epsilon,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) - R_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) = 1\right\}$$

with the convention that $\epsilon^*_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) = 0$ *if the infimum does not exist.*

Proof. The proof of Lemma 1 follows along the lines of the proof of Lemma 5 in [25] and by replacing the expressions of $\boldsymbol{\Phi}_{t,L}(j)$, $V_{T,L}(v)$, and $R_{T,L}(v)$ by the expressions of their dependency-aware extensions $\boldsymbol{\Phi}^{\mathrm{DA}}_{t,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$, $V_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$, and $R_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$, respectively. □

Lemma 2. *Let* $V^+_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ *be as in* (15). *For all triples* $(T,L,v) \in \{1,\dots,L\}\times\mathbb{N}_+\times[0.5,1)$, $V^+_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ *is monotonically increasing in* $\rho_{\mathrm{thr}}(u_{\mathrm{c}})$, *i.e., for any* $u_{\mathrm{c}} \in \{1,\dots,p-1\}$, *it holds that*

$$V^+_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \geq V^+_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})).$$

Proof. Using the definition of $V^+_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))$ in (15), we obtain

$$\begin{aligned}
&V^+_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \\
&\quad= \left|\left\{\text{null } j :\ \Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1))\cdot\boldsymbol{\Phi}_{T,L}(j) > v\right\}\right| \\
&\quad\geq \left|\left\{\text{null } j :\ \Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))\cdot\boldsymbol{\Phi}_{T,L}(j) > v\right\}\right| \\
&\quad= V^+_{T,L}(v,\rho_{\mathrm{thr}}(u_{\mathrm{c}})).
\end{aligned}$$

The inequality in the third line follows from

$$\Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \geq \Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})),$$

which is a consequence of the following two cases:

(i) $\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) = \varnothing$:

$$\Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) = 1 = \Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})),$$

(ii) $\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})) \neq \varnothing$:

$$\begin{aligned}
&\Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \\
&\quad= \left[2 - \min_{j'\in\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1))}\left\{\left|\boldsymbol{\Phi}_{t,L}(j) - \boldsymbol{\Phi}_{t,L}(j')\right|\right\}\right]^{-1} \\
&\quad\geq \left[2 - \min_{j'\in\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}\left\{\left|\boldsymbol{\Phi}_{t,L}(j) - \boldsymbol{\Phi}_{t,L}(j')\right|\right\}\right]^{-1} \\
&\quad= \Psi^+_{T,L}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})).
\end{aligned}$$

In (ii), the inequality in the third line follows from the fact that, for any $u_{\mathrm{c}} \in \{1,\dots,p-1\}$, it holds that

$$\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1)) \subseteq \mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}})), \quad j = 1,\dots,p,$$

and, therefore,

$$\min_{j'\in\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}+1))}\left\{\left|\boldsymbol{\Phi}_{t,L}(j) - \boldsymbol{\Phi}_{t,L}(j')\right|\right\} \geq \min_{j'\in\mathrm{Gr}(j,\rho_{\mathrm{thr}}(u_{\mathrm{c}}))}\left\{\left|\boldsymbol{\Phi}_{t,L}(j) - \boldsymbol{\Phi}_{t,L}(j')\right|\right\}.$$

□

Lemma 3. *Let $\widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c))$ be as in (16). For all tuples $(T, L) \in \{1, \dots, L\} \times \mathbb{N}_+$, $\widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c))$ is monotonically decreasing in $\rho_{\text{thr}}(u_c)$, i.e., for any $u_c \in \{1, \dots, p-1\}$, it holds that*

$$\widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c + 1)) \leq \widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c)).$$

Proof. Using Eq. (10), we obtain

$$\begin{aligned}
&\widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c + 1)) \\
&= \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Psi^{+}_{t,L}(q, \rho_{\text{thr}}(u_c + 1)) \cdot \Phi_{t,L}(q)}{L - (t-1)} \\
&\leq \sum_{t=1}^{T} \frac{p - \sum_{q=1}^{p} \Psi^{+}_{t,L}(q, \rho_{\text{thr}}(u_c)) \cdot \Phi_{t,L}(q)}{L - (t-1)} \\
&= \widehat{V}'^{+}_{T,L}(0.5, \rho_{\text{thr}}(u_c)),
\end{aligned}$$

where the inequality in the third line follows from

$$\Psi^{+}_{t,L}(q, \rho_{\text{thr}}(u_c + 1)) \geq \Psi^{+}_{t,L}(q, \rho_{\text{thr}}(u_c)),$$

which was shown to hold within the proof of Lemma 2. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.sigpro.2025.109990.

Data availability

The data/code used in this work is publicly available.

References

[1] P.-J. Chung, J.F. Bohme, C.F. Mecklenbrauker, A.O. Hero, Detection of the number of signals using the benjamini-hochberg procedure, IEEE Trans. Signal Process. 55 (6) (2007) 2497–2508.

[2] J. Chen, W. Zhang, H.V. Poor, A false discovery rate oriented approach to parallel sequential change detection problems, IEEE Trans. Signal Process. 68 (2020) 1823–1836.

[3] Z. Chen, F. Sohrabi, W. Yu, Sparse activity detection for massive connectivity, IEEE Trans. Signal Process. 66 (7) (2018) 1890–1904.

[4] Z. Tan, Y.C. Eldar, A. Nehorai, Direction of arrival estimation using co-prime arrays: A super resolution viewpoint, IEEE Trans. Signal Process. 62 (21) (2014) 5565–5576.

[5] K. Benidis, Y. Feng, D.P. Palomar, Sparse portfolios for high-dimensional financial index tracking, IEEE Trans. Signal Process. 66 (1) (2017) 155–170.

[6] A.M. Zoubir, V. Koivunen, Y. Chakhchoukh, M. Muma, Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts, IEEE Signal Process. Mag. 29 (4) (2012) 61–80.

[7] A.M. Zoubir, V. Koivunen, E. Ollila, M. Muma, Robust statistics for signal processing, Cambridge Univ. Press, 2018.

[8] J. Machkour, B. Alt, M. Muma, A.M. Zoubir, The outlier-corrected-data-adaptive lasso: A new robust estimator for the independent contamination model, in: 25th Eur. Signal Process. Conf., EUSIPCO, 2017, pp. 1649–1653.

[9] J. Machkour, M. Muma, B. Alt, A.M. Zoubir, A robust adaptive lasso estimator for the independent contamination model, Signal Process. 174 (2020) 107608.

[10] C. Yang, X. Shen, H. Ma, B. Chen, Y. Gu, H.C. So, Weakly convex regularized robust sparse recovery methods with theoretical guarantees, IEEE Trans. Signal Process. 67 (19) (2019) 5046–5061.

[11] K.-Q. Shen, C.-J. Ong, X.-P. Li, Z. Hui, E.P. Wilder-Smith, A feature selection method for multilevel mental fatigue EEG classification, IEEE. Trans. Biomed. Eng. 54 (7) (2007) 1231–1237.

[12] K. Tomczak, P. Czerwińska, M. Wiznerowicz, Review the cancer genome atlas (TCGA): an immeasurable source of knowledge, Contemp Oncol (Pozn) 2015 (1) (2015) 68–77.

[13] Á. Ősz, A. Lánczky, B. Győrffy, Survival analysis in breast cancer using proteomic data from four independent datasets, Sci. Rep. 11 (1) (2021) 16787.

[14] O. Aalen, O. Borgan, H. Gjessing, Survival and event history analysis: A process point of view, Springer Science & Business Media, 2008.

[15] H.F.M. Kamel, H.S.A.B. Al-Amodi, Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine, Genom. Proteom. Bioinform. 15 (4) (2017) 220–235.

[16] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B. Stat. Methodol. 57 (1) (1995) 289–300.

[17] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Ann. Statist. 29 (4) (2001) 1165–1188.

[18] R.F. Barber, E.J. Candès, Controlling the false discovery rate via knockoffs, Ann. Statist. 43 (5) (2015) 2055–2085.

[19] Y. Gavrilov, Y. Benjamini, S.K. Sarkar, An adaptive step-down procedure with proven FDR control under independence, Ann. Statist. 37 (2) (2009) 619–629.

[20] R.F. Barber, E.J. Candès, A knockoff filter for high-dimensional selective inference, Ann. Statist. 47 (5) (2019) 2504–2537.

[21] E.J. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection, J. R. Stat. Soc. Ser. B. Stat. Methodol. 80 (3) (2018) 551–577.

[22] M. Sesia, C. Sabatti, E.J. Candès, Gene hunting with hidden Markov model knockoffs, Biometrika 106 (1) (2019) 1–18.

[23] Y. Romano, M. Sesia, E.J. Candès, Deep knockoffs, J. Amer. Statist. Assoc. (2019) 1–12.

[24] Z. Ren, Y. Wei, E. Candès, Derandomizing knockoffs, J. Amer. Statist. Assoc. (2021) 1–11.

[25] J. Machkour, M. Muma, D.P. Palomar, The terminating-random experiments selector: Fast high-dimensional variable selection with false discovery rate control, Signal Process. 231 (2025) 109894.

[26] J. Machkour, M. Muma, D.P. Palomar, False discovery rate control for grouped variable selection in high-dimensional linear models using the t-knock filter, in: 30th Eur. Signal Process. Conf., EUSIPCO, 2022, pp. 892–896.

[27] J. Machkour, M. Muma, 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.

[28] F. Scheidt, J. Machkour, 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.

[29] J. Machkour, M. Muma, 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.

[30] T. Koka, J. Machkour, 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.

[31] J. Machkour, A. Breloy, M. Muma, D.P. Palomar, 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.

[32] M.R. Segal, K.D. Dahlquist, B.R. Conklin, Regression approaches for microarray data analysis, J. Comput. Biol. 10 (6) (2003) 961–980.

[33] D.J. Balding, A tutorial on statistical methods for population association studies, Nat. Rev. Genet. 7 (10) (2006) 781–791.

[34] J. Machkour, D.P. Palomar, M. Muma, FDR-controlled portfolio optimization for sparse financial index tracking, in: Proc. IEEE 50th Int. Conf. Acoust. Speech Signal Process., ICASSP, 2025.

[35] 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, Nucleic Acids Res. 47 (D1) (2019) D1005–D1012.

[36] 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, PLOS Med. 12 (3) (2015) e1001779.

[37] D. Williams, Probability with martingales, Cambridge Univ. Press, 1991.

[38] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2) (2004) 407–499.

[39] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B. Stat. Methodol. 58 (1) (1996) 267–288.

[40] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B. Stat. Methodol. 67 (2) (2005) 301–320.

[41] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (476) (2006) 1418–1429.

[42] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 2 (1) (2012) 86–97.

[43] J. Machkour, S. Tien, D.P. Palomar, M. Muma, TRexSelector: T-Rex selector: High-dimensional variable selection & FDR control, 2024, R package version 1.0.0. Available: https://CRAN.R-project.org/package=TRexSelector.

[44] J. Machkour, S. Tien, D.P. Palomar, M. Muma, tlars: The T-LARS algorithm: Early-terminated forward variable selection, 2024, R package version 1.0.1. Available: https://CRAN.R-project.org/package=tlars.

[45] M. Baker, Reproducibility crisis, Nature 533 (26) (2016) 353–366.

[46] 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, Nucleic Acids Res. 44 (8) (2016) e71.

[47] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 1–21.
[48] J. Buckley, I. James, Linear regression with censored data, Biometrika 66 (3) (1979) 429–436.
[49] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for cox's proportional hazards model via coordinate descent, J. Stat. Softw. 39 (5) (2011) 1.
[50] C. Zhou, M. Wang, J. Yang, H. Xiong, Y. Wang, J. Tang, Integral membrane protein 2A inhibits cell growth in human breast cancer via enhancing autophagy induction, Cell Commun. Signal. 17 (2019) 1–14.
[51] M. Lacroix, Significance, detection and markers of disseminated breast cancer cells, Endocr.- Relat. Cancer 13 (4) (2006) 1033–1067.
[52] 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, BioMed Res. Int. 2021 (2021).
[53] J. Machkour, M. Muma, 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.