Variable Selection via Grace-AKO with Applications in Genomics

Peixin Tian, Yiqian Hu, Zhonghua Liu*, Yan Dora Zhang*

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Hong Kong SAR, China.

Abstract

Motivation: Variable selection is a common statistical approach to detecting genes associated with clinical outcomes of scientific interest. There are hundreds of thousands of genes in genomic studies, while only a limited number of individual samples are available. Therefore, it is important to develop a method to identify genes associated with outcomes of interest that can control finite-sample false discovery rate (FDR) in high-dimensional data settings.

Results: This article proposes a novel method named Grace-AKO for graph-constrained estimation (Grace), which incorporates aggregation of multiple knockoffs (AKO) with the network-constrained penalty. Grace-AKO can control FDR in finite sample settings and improve stability. Simulation studies show that Grace-AKO has better performance in FDR control in finite sample settings than the original Grace model. We apply Grace-AKO to the prostate cancer data in the The Cancer Genome Atlas (TCGA) program by incorporating prostate-specific antigen (PSA) pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) as the prior information. Grace-AKO finally identifies 47 candidate genes associated with PSA level, and more than 75% of the detected genes can be validated.

Availability and implementation: We developed an R package for Grace-AKO available at: https://github.com/mxxptian/GraceAKO

Contact: doraz@hku.hk

Supplementary information: Supplementary data are available at Bioinformatics online.

*Contributed equally

1 Introduction

Identifying genes and pathways associated with human diseases is one primary goal in genomic studies. A major challenge is that genomic data is usually high-dimensional but with a limited sample size. Currently, there are rich public genomic data, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/), which provides gene-regulatory pathways including biological regulatory relationships between genes or gene products. These gene-regulatory pathways form a network that can be represented as a graphical structure, where the vertices are genes or gene products, and the edges are the gene-regulatory pathways. Inspired by the nature of genomic data, graph-constrained estimation (Grace) provides a new insight on variable selection for genomic data, and accounts for the graphical structures. The predictors in the network are gene-expression data linked by gene-regulatory pathways for a specific clinical outcome, which can also be called graph-structured covariates (Li and Li, 2010). These graphical structures, such as gene-regulatory pathways, can improve the sensitivity of detecting pathways (Rahnenführer et al., 2004). Furthermore, Li and Li (2008) developed a Grace model, network-constrained regularization procedure by encoding the graphical structures into a Laplacian matrix to incorporate this kind of prior biological information into regression analysis. Network-constrained regularization includes the Lasso (Tibshirani, 1996), and the elastic net regulation procedure (Zou and Hastie, 2005) as special cases. However, these conventional regression approaches only control false discovery rate (FDR) asymptotically with no guarantee in finite-sample FDR control.

The model-X knockoffs procedure was proposed to control FDR under a pre-specified FDR level in finite sample settings (Barber and Candès, 2015; Candes et al., 2018). However, owing to the randomness of model-X knockoffs drawings, the performance of model-X knockoffs is unstable. Multiple knockoffs allow for a more stable finite-sample FDR control and power improvement (Emery and Keich, 2019; Gimenez and Zou, 2019; Nguyen et al., 2020). Specifically, statistical aggregation is used to solve the instability by consolidating the model-X knockoffs inference. Aggregation of multiple knockoffs (AKO) was proposed by Nguyen et al. (2020), which rested on a reformulation of model-X knockoffs to introduce an intermediate feature statistic. It borrowed the idea of Meinshausen et al. (2009) to replicate model-X knockoff procedure multiple times, and then performed aggregation to generate new intermediate feature statistics. Hence,

it is more stable than the model-X knockoff procedure.

In this article, we propose a novel method, graph-constrained estimation with aggregation of multiple knockoffs (Grace-AKO). Specifically, the main procedure of Grace-AKO is as follows. First, we generate model-X knockoffs according to the correlation structure of the graph-structured covariates (Candes et al., 2018) and encode graphical structures into a Laplacian matrix (Chung, 1997). Second, we simultaneously fit the graph-structured covariates and their model-X knockoffs into Grace and compute multiple feature statistics: Lasso coefficient-differences (LCDs). Third, we repeat the steps above multiple times and convert the LCDs to new intermediate feature statistics, Aggregated Grace Coefficients (AGCs). Fourth, we conduct the Benjamini-Hochberg procedure on the multiple AGCs and select the candidate variables. The proposed Grace-AKO reduces the inherent instability of model-X knockoffs while taking the graphical structure into account for variable selection.

The remainder of this article is organized as follows. In Section 2, we describe the method of Grace-AKO under the linear regression framework. In Section 3, we assess the performance of Grace-AKO using simulation studies. In Section 4, we apply Grace-AKO to the prostate cancer data from the The Cancer Genome Atlas (TCGA) program by incorporating KEGG pathways as prior information. In Section 5, we briefly summarize our method.

2 Grace-AKO for Linear Models

We use the following linear regression to model the relationship between the outcome $\mathbf{y}$ and p predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p$:

$$\mathbf{y} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + \ldots + \mathbf{x}_p\hat{\beta}_p + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_n)^{\mathrm{T}}$ is the response variable, and the graph-structured covariates, $\mathbf{x}_i = (\mathrm{x}_{i1}, \mathrm{x}_{i2}, \cdots, \mathrm{x}_{ip})$, $i = 1, \cdots, n$ represents a vector of the predictors from genomic data (gene-expression data), and $\boldsymbol{\epsilon}$ is a vector of random errors. We further assume that the response is centred and the predictors are standardized,

$$\sum_{i=1}^{n} \mathrm{y}_i = 0, \quad \sum_{i=1}^{n} \mathrm{x}_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} \mathrm{x}_{ij}^2 = 1, \quad \text{for } j = 1, 2, \ldots, p.$$

The main goal is to estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ and select the predictors with nonzero coefficients. For some complex diseases, the regulatory relationships represented by graphical structures between genes or gene products shed light into underlying biological knowledge, where the covariates are the nodes of the graph, and the edges indicate functional relationships between two genes. Furthermore, in such a graphical structure, the genes are associated by edges with certain probabilities, in which the edge probability is treated as a weight in an undirected graph to form a weighted graph (Li and Li, 2010). Li and Li (2008) proposed a network-constrained regularization criterion:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2\boldsymbol{\beta}^T\boldsymbol{L}\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{L}$ is a non-negative Laplacian matrix of a weighted graph, and $\boldsymbol{X}$ is a $n \times p$ design matrix, and $||\cdot||_1$ indicates the L_1 norm, and λ_1 is the Lasso penalty, and λ_2 is the penalty for the Laplacian matrix $\boldsymbol{L}$. In equation (1), the L_1 norm deals with sparse matrices, and $\boldsymbol{\beta}^T\boldsymbol{L}\boldsymbol{\beta}$ induces a smooth solution of coefficients of the graph-structured covariates. Additionally, $\boldsymbol{L}$ depicts the graphical structure using set V including vertices corresponding to the predictors and a weighted edge $w(u, v)$, where u and v denote the predictors in the graphical structure, and the degree of vertex v is represented as $d_v = \sum_{u\sim v} w(u, v)$. In the genomic data, $w(u, v)$ quantifies the uncertainty of the edge between two vertices, such as the probability of an edge connecting two graph-structured covariates when the graphical structure is constructed from data. Motivated by Chung (1997), we apply the normalized $\boldsymbol{L}$ given in Li and Li (2008):

$$L(u, v) = \begin{cases} 1 - w(u, v)/d_u, & \text{if } u = v \text{ and } d_u \neq 0; \\ -w(u, v)/\sqrt{d_u d_v}, & \text{if } u \text{ and } v \text{ are adjacent;} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, we can rewrite the second penalty term $\boldsymbol{\beta}^T\boldsymbol{L}\boldsymbol{\beta}$ of equation (1) as follows (Li and Li, 2008):

$$\boldsymbol{\beta}^T\boldsymbol{L}\boldsymbol{\beta} = \sum_{u\sim v}(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}})^2 w(u, v).$$

While the network-constrained regularization procedure incorporates the graphical structures for variable selection, it does not guarantee false discovery rate (FDR) control in finite

sample settings. To address this issue, we introduce a kind of multiple knockoffs to control finite-sample FDR. The knockoff filter procedure was first proposed in Barber and Candès (2015), and Candes et al. (2018) further proposed the model-X knockoff filter to extend its application. Model-X knockoffs, $\tilde{\boldsymbol{X}}$ are generated from the original data and retain the same data structure as the originals $\boldsymbol{X}$, in which $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are pairwise exchangeable. We summarize the properties of model-X knockoffs as:

(1) Swapping the locations of related elements, $\mathbf{x}_j$ and $\tilde{\mathbf{x}_j}$, would not change the joint distribution of $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ conditional on $\mathbf{y}$.

(2) Once the original covariates $\boldsymbol{X}$ are known, their model-X knockoffs, $\tilde{\boldsymbol{X}}$ provides no extra information on the response variable $\mathbf{y}$.

The knockoffs filter is a cheap method since it does not require strong assumptions about the design matrix $\boldsymbol{X}$. Nonetheless, owing to the random nature of model-X knockoffs sampling, multiple knockoffs approaches were developed to improve the stability, and aggregation of multiple knockoffs (AKO) is one of them (Nguyen et al., 2020). In this article, we propose a novel approach, namely Grace-AKO by introducing the aggregation idea of AKO to achieve finite-sample FDR control and stability improvement.

Grace-AKO takes the advantages of the network-constrained penalty and AKO to account for biological network information with finite-sample FDR control simultaneously. First, we generate model-X knockoffs, $\tilde{\boldsymbol{X}}$ from the original data matrix $\boldsymbol{X}$ (Candes et al., 2018). $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ is regarded as a new design matrix of the predictors. The graphical structure is represented by the Laplacian matrix $\boldsymbol{L}$.

Second, the response variable $\boldsymbol{y}$ and the new predictors $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ are fitted in equation (1). Motivated by Li and Li (2008), a natural solution of our model is equivalent to the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}\{\|\mathbf{y} - (\boldsymbol{X}, \tilde{\boldsymbol{X}})(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}})^T|_2\}, \tag{3}$$

subject to:

$$(1-\alpha)\sum_{j=1}^{p}\|\beta_j\|_1 + \alpha\sum_{u\sim v}(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}})^2 w(u,v) \leq t, \tag{4}$$

where $\alpha = \lambda_2/(\lambda_1+\lambda_2)$, and t is a constant value, and $(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}})^T$ is a vector of the new coefficients. In the function $(1-\alpha)\sum_{j=1}^{p}\|\beta_j\|_1 + \alpha\sum_{u\sim v}(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}})^2 w(u,v)$, $\|\cdot\|_1$ deals with sparse data matrix same as its function in the network-constrained penalty, and the second term of it penalizes the weighted sum of the squares of the difference of coefficients between the graph-structured covariates, which is scaled by the degree of the associated vertices in the network.

We use the Lasso coefficient-difference (LCD) as the feature statistic proposed by Candes et al. (2018):

$$W_j = |\hat{\beta}_j| - |\hat{\beta}_{j+p}|, \tag{5}$$

where $\hat{\beta}_j$, and $\hat{\beta}_{j+p}$ are the estimated coefficients of $\mathbf{x}_j$ and $\tilde{\mathbf{x}_j}$, respectively. A high positive value of W_j indicates that the distribution of $\mathbf{y}$ depends upon $\mathbf{x}_j$. Moreover, due to the symmetric distribution of W_j under the null, W_j equally takes on positive and negative values (Candes et al., 2018).

Third, following Nguyen et al. (2020), Grace-AKO converts the feature statistic W_j to a new intermediate feature statistic q_j:

$$q_j = \begin{cases} \frac{1+\#\{k:W_k \le -W_j\}}{p}, & W_j > 0; \\ 1, & W_j < 0. \end{cases} \tag{6}$$

We then repeat the aforementioned steps B times, including generating knockoffs and calculating the intermediate feature statistic q_j. Furthermore, we propose a new feature statistic, Aggregated Grace Coefficient (AGC) inspired by the quantile aggregation procedure in Meinshausen et al. (2009):

$$\bar{q}_j = \min\{1, \frac{Q_\gamma(\{q_j^{(b)} : b = \{1, 2, \cdots, B\})}{\gamma}\}, \tag{7}$$

where γ is the quantile point, and $Q(\cdot)$ denotes the quantile function, and B is the pre-specified replication times.

Fourth, we apply the multiple testing procedure, the Benjamini-Hochberg procedure (BH) to the AGCs, $\bar{\boldsymbol{q}} = (\bar{q}_1, \bar{q}_2, \cdots, \bar{q}_p)$. A data-dependent threshold is established by

$$t_{BH} = \max\{k : \bar{q}_{(k)} \le \frac{k}{\alpha p}\},$$

where α is the user-specified nominal FDR level. We finally choose the candidate variables satisfying the following requirement:

$$\hat{S} = \{j \in [p] : \bar{q}_{(j)} \leq \bar{q}_{(\hat{t}_{BH})}\}.$$

In this article, we measure the performance using the modified false discovery rate (mFDR):

$$\text{mFDR} = \mathbb{E}\Big[\frac{|\{j \in \hat{S} \cap S_0\}|}{|\hat{S}| + 1/\alpha}\Big],$$

where $S_0 = \{j \in [p] : \beta_j = 0\}$ which includes null variables. The implementation details are available in Algorithm 1.

Algorithm 1: Algorithm of Grace-AKO

Data: $(\boldsymbol{X}, \boldsymbol{Y}, \gamma, q)$: $\boldsymbol{X}$-original data, $\boldsymbol{Y}$-response variable, B-loop times, γ-pre-specified quantiles, q-pre-specified FDR level.
Result: Variable selection set: $\hat{S}$.
while $b \leq B$ **do**
  Draw knockoffs variables $\tilde{\boldsymbol{X}}$;
  Fit $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ into Grace model;
  Calculate W_j based on (5); Calculate new statistics q_j based on (6).
Do aggregation by quantile function (7) with a matrix of q_j and γ to get AGCs;
Use BH procedure to compute threshold: $t_{BH} = max\{k : \bar{q}_{(k)} \leq \frac{k}{\alpha p}\}$;
Variable selection set: $\hat{S} = \{j \in [p] : \bar{q}_{(j)} \leq \bar{q}_{(\hat{t}_{BH})}\}$.

As an illustration in Candes et al. (2018), the primary objective of the knockoff filter procedure was to build an as permissive as possible data-dependent threshold. The threshold can provide a controllable estimation of FDR to ensure model-X knockoffs property. Nguyen et al. (2020) further established an independence or positive correlation between the intermediate feature statistics and the original feature statistics in Candes et al. (2018) to replace the originals. Additionally, AKO still provided a data-dependent threshold based on the specification in Candes et al. (2018).

3 Simulation Studies

In this section, we perform a wide range of simulation studies to evaluate our proposed method, Grace-AKO, and compare it to the network-constrained regularization, abbreviated as Grace

(Li and Li, 2008). We suppose that there are 10 transcription factors (TFs) and each regulates 10 genes. The graphical structure includes p unconnected regulatory modules with p TFs in total and edges linked each of the TFs and 10 genes regulated. Here we denote the first 44 genes related to the response variable $\mathbf{y}$. We assume that the data are simulated from the following settings:

(1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where
$$\boldsymbol{\beta} = (5, \overbrace{\frac{5}{\sqrt{10}}, \cdots, \frac{5}{\sqrt{10}}}^{10}, -5, \overbrace{\frac{-5}{\sqrt{10}}, \cdots, \frac{-5}{\sqrt{10}}}^{10}, 3, \overbrace{\frac{3}{\sqrt{10}}, \cdots, \frac{3}{\sqrt{10}}}^{10}, -3, \overbrace{\frac{-3}{\sqrt{10}}, \cdots, \frac{-3}{\sqrt{10}}}^{10}, 0, \cdots 0),$$
and $\boldsymbol{\epsilon}$ is generated from $N(0, \boldsymbol{\sigma}^2)$;

(2) The noise level is denoted as $\sigma^2 = \sum_u \beta_u^2/4$;

(3) For each expression level TF, X is drawn from normal distribution: $\boldsymbol{X}_g \sim N(0, 1)$, and conditional on the TF, the expression levels of genes which regulated to the specific TF are drawn from a conditional normal distribution with correlation of 0.7.

We set $n = (100, 200, 300)$ and $p = (10, 20, 40, 60)$ respectively, to simulate 12 settings in total. As the empirical research in Nguyen et al. (2020) demonstrated, performance was consistent and robust once the iteration B approached 25 times. In the meanwhile, FDR could almost be controlled under the pre-specified level. Therefore, we fix $B = 25$, and $\gamma = 0.1$ in all scenarios. Additionally, we specify the FDR control level α at 0.1. According to Li and Li (2008), Grace performed better than the Lasso (Tibshirani, 1996) and the elastic net (Zou and Hastie, 2005) in term of the combination of $\|\cdot\|_1$ and $\|\cdot\|_2$. We thus focus on the comparison of Grace-AKO and Grace in our simulation studies. We assess the regression performance using the average values of mFDR, standard errors, and the number of selected variables.

To tune the parameters, we apply 10-fold cross-validation. The values of tuning parameter $\lambda_{\boldsymbol{L}}$ are specified from the range of 0.1 to 2.0 with a step size of 0.1, which ensures that the $\boldsymbol{L}$ matrix is non-negative. Moreover, the values of λ_1 are drawn from the range of 110 to 200 with a step size of 5. The validation set of λ_2 ranges from 1 to 10 with a step size of 1. Each simulation setting is repeated for 30 times.

As Table 1 shows, the mFDR values of Grace-AKO are controlled at the pre-specified level, and Grace cannot control mFDR in finite sample settings. Furthermore, as the data dimension p increases with a fixed sample size n, the mFDR for Grace increases which can even reach at

0.67. However, Grace-AKO can always control the mFDR under 0.1. Moreover, the standard errors for Grac-AKO is alway smaller than Grace's, which indicates that Grace-AKO can also improve the stability. In Figure 1, we depict a scatter plot for the findings when $n = 300$. The upper figure's black line denotes the expected mFDR control level. Grace, in particular, has a higher mFDR than Grace-AKO, but Grace cannot control mFDR in finite sample settings.

Table 1: The mean and standard errors (in brackets) of modified false discovery rate (mFDR) over 30 replications for Grace-AKO v.s. Grace model with a pre-specified mFDR level at 0.1.

p	$n = 100$		$n = 200$		$n = 300$	
	Grace-AKO	**Grace**	**Grace-AKO**	**Grace**	**Grace-AKO**	**Grace**
10	0.01 (0.018)	0.15 (0.032)	0.03 (0.032)	0.22 (0.049)	0.03 (0.025)	0.26 (0.053)
20	0.02 (0.027)	0.26 (0.045)	0.02 (0.023)	0.37 (0.045)	0.03 (0.040)	0.43 (0.010)
40	0.03 (0.030)	0.24 (0.062)	0.04 (0.030)	0.48 (0.040)	0.06 (0.047)	0.61 (0.030)
60	0.04 (0.026)	0.40 (0.053)	0.03 (0.026)	0.55 (0.032)	0.06 (0.034)	0.67 (0.018)

As shown in Table 2, the number of selected variables varies significantly between these two models. The number of true variables is fixed at 44. When n and p increase, the number of variables chosen by Grace increases. Notably, when $n = 300$ and $p = 60$, Grace identifies 138 candidate variables when the number of true variables is 44. It demonstrates that Grace could identify many false positive variables in finite sample settings. Grace-AKO has relatively more stable performance under all the data settings, which can correctly capture the true candidate genes. This is the reason that in Figure 1, the power of Grace inflates falsely. However, given the uncontrolled mFDR, it's not meaningful to discuss the power performance. Moreover, Grace-AKO can control mFDR with decent power in finite sample settings.

Table 2: The average numbers of variables selected for Grace-AKO v.s. Grace over 30 replications with a pre-specified mFDR level at 0.1. The number of true variables is 44.

p	$n = 100$		$n = 200$		$n = 300$	
	Grace-AKO	**Grace**	**Grace-AKO**	**Grace**	**Grace-AKO**	**Grace**
10	24	45	33	56	37	62
20	25	52	32	71	37	83
40	24	59	34	90	39	118
60	24	63	32	101	39	138

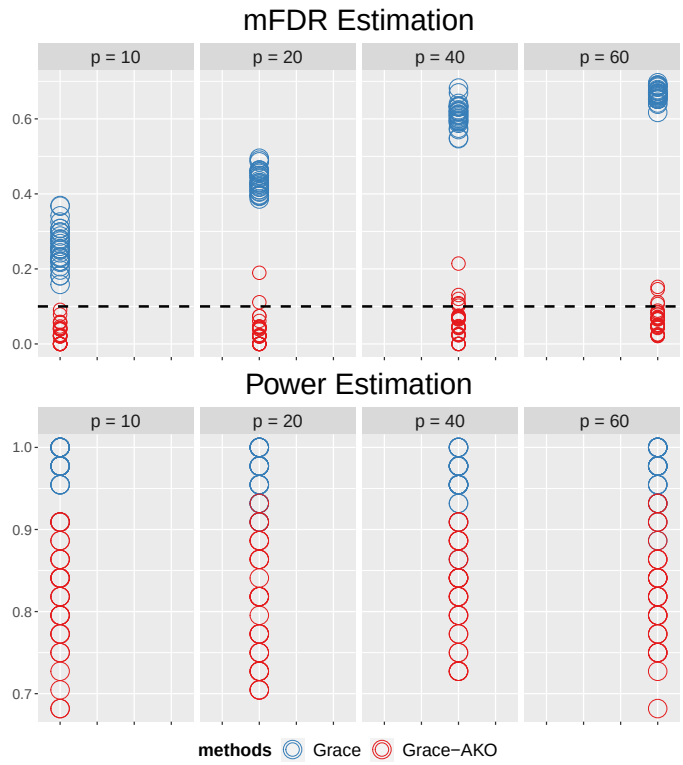

Figure 1: Figures of mFDR and Power for Grace-AKO v.s Grace model when $n = 300$. The upper figure plots mFDR and lower figure plots power. The red circle represents our new method, Grace-AKO and the blue one represents Grace. Each setting has been replicated for 30 times, and each dot represents one trail. The black dashed line is the pre-specified mFDR level which equals to 0.1.

4 Application to the TCGA Prostate Cancer Data

To demonstrate the usefulness of Grace-AKO, we apply it to a gene-expression data of prostate-specific antigen (PSA) level from the The Cancer Genome Atlas (TCGA) program. The TCGA program is a landmark cancer genomics program with over 20,000 primary cancers and standard samples spanning 33 cancer types (Koboldt et al., 2012). Additionally, the Kyoto Encyclopedia of Genes and Genomes (KEGG), as a public database, has enabled a profound insight into the graphical structures of genes and various clinical outcomes (Kanehisa and Goto, 2000). It contributes to understanding various aspects of biological systems and pathway information. Prostate cancer is the most common malignancy in mid-aged males and the second leading malignancy (Stangelberger et al., 2008). Wang et al. (2018) indicated that metastatic prostate

cancer remained incurable even in patients who finished intensive multimodal therapy. It is an urgent challenge to propose a novel approach for disease management via identifying prognostic determinant genes.

We first remove the samples with missing measurements and then apply PSA pathways (Kim et al., 2018) to construct the normalized Laplacian matrix for Grace-AKO. We finally obtain the data with sample size $n = 339$ and dimension $p = 5947$. We denote the PSA level as our response. To reduce computational demands, we first perform variable screening via correlation learning following Fan and Lv (2008). The final sample size and dimension are $n = 339$ and $p = 600$, respectively, consisting of the data structure in simulation studies. We then standardize the explanatory variables and response variables. The parameter λ_1 ranges from 1 to 40 length out as 4, and λ_2 is among 1, 2, 3, 4, and 5. We denote the target FDR level at 0.2 and $B = 25$ for the AKO procedure. The quantile point is $\gamma = 0.3$. For a fair comparison, we both conduct 10-fold cross-validation to select the tuning parameters. Furthermore, we only consider variables with p-values of less than 1 when performing the Benjamini-Hochberg procedure. We show the details in Figure 2.

There are 90 genes selected by Grace in total, and 47 genes selected by Grace-AKO. Grace might lose some accuracy when the coefficients' signs are different (Li and Li, 2010). After checking the intersection of variable selection sets, we find that the genes selected by Grace-AKO are a subset of Grace's. However, due to the inflated FDR in the simulation studies, the results of Grace may identify many false discoveries. Additionally, previous studies confirmed 35 out of 47 candidate genes for Grace-AKO as true associated genes. Some details of them are listed in Table 3. For example, IL9 was recently assigned as an essential gene in tumor immunity, and AK6 was already regarded as a biomarker in prostate cancer treatments (Liu et al., 2015; Wan et al., 2020). HLA-DRB5 was highly related to MHC-II genes, which were the related genes of prostate cancer, and so was the CHRNB2 (Axelrod et al., 2019). Moreover, a high concentration of IFNA2 was related to advanced prostate carcinoma (Erb et al., 2013). PLK1 was over-expressed in many cancers, including prostate cancer, and scientists found that translation of the PP2A-PLK1 SDL interaction caused the expression of PLK1 and PP2A, which were commonly regarded as a biomarker in the cancer cells (Cunningham et al., 2016). Melloy (2019) indicated the percentage of the case with alteration of ANAPC7 achieved beyond

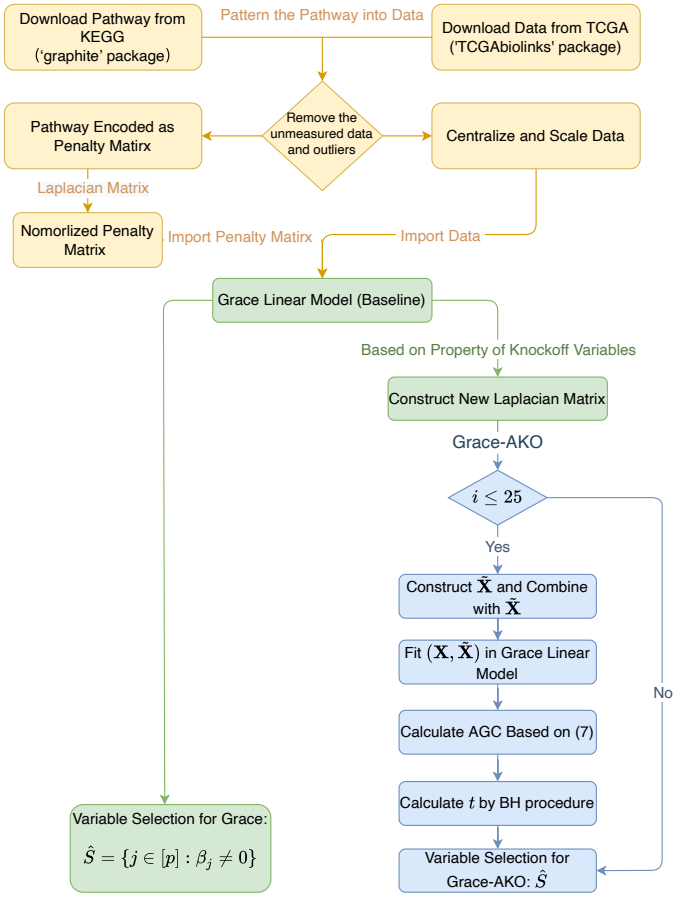

Figure 2: Flowchart of Grace-AKO. These two procedures are differentiated by colors, in which the green one is Grace and the blue one is Grace-AKO. First, we encode the Laplacian matrix account for the network structure and sample model-X knockoffs. Second, we compute the feature statistics LCDs. Third, we aggregate the LCDs to get AGCs. Fourth, we select the variables whose AGCs satisfy the requirements.

5.21 percentage points. The results demonstrated that the APC/C had a profound effect on cancer survival. MYC in 2010 had already been confirmed to be affected by the loss of the tumor in Koh et al. (2010).

Furthermore, there were also some subnetworks in our findings. H2BC9, H2AC14, and

H2AC7 consisted of a small subnetwork. The IFNA7, IL9, IL5RA, IFNA2, PDGFB, and KIT consisted of a subnetwork that contained pathways. Except for KIT, the rest of these genes were treated as potential genes for prostate cancer treatment. There was also a pathway between CAMK4 and PRKACF. Dadwal et al. (2018) indicated that CAMKK2 was a crucial androgen receptor target for prostate cancer tumor growth. MYC was linked with RXRA. In Ray et al. (2020), RXRA, which was discovered as a novel target of miR-191, was conserved in a cell line derived from radio recurrent prostate cancer.

Table 3: Genes Selected by Grace-AKO in the application of Prostate Cancer and PSA Pathways.

Genes Name	Genes Expression
ANAPC7	Ubiquitous expression in testis (RPKM 11.7), esophagus (RPKM 9.0) and 25 other tissues
AK6	Ubiquitous expression in testis (RPKM 18.6), adrenal (RPKM 17.2) and 25 other tissues
CHRNB2	Biased expression in brain (RPKM 9.5), adrenal (RPKM 0.7) and 1 other tissue
COL4A5	Broad expression in endometrium (RPKM 21.7), skin (RPKM 10.8) and 20 other tissues
PRLR	Biased expression in placenta (RPKM 18.9), endometrium (RPKM 11.6) and 8 other tissues
CAMK4	Broad expression in brain (RPKM 7.0), lymph node (RPKM 3.7) and 21 other tissues
CHRNB2	Biased expression in brain (RPKM 9.5), adrenal (RPKM 0.7) and 1 other tissue
PPP2R2D	Ubiquitous expression in pancreas (RPKM 6.3), ovary (RPKM 5.9) and 25 other tissues

Note: The rest of the identified genes can be found in the supplemental materials.

5 Discussion

This article introduces Grace-AKO to perform the variable selection by incorporating the network-constrained penalty (Li and Li, 2008) and AKO (Nguyen et al., 2020). In contrast to the conventional variable selection process, the network-constrain penalty applies a normalized Laplacian matrix to encode the graphical structures between the potential genes or the gene products. It applies the L_1 penalty to selected variables and the L_2 penalty to degree-scaled differences of coefficients concerning the graphical structure. However, Grace has a poor performance in finite-sample FDR control. Our proposed Grace-AKO solves this problem by conducting variable selection via multiple knockoff variables. In finite sample settings, Grace-AKO can control the FDR under a user-specific level. Moreover, Grace-AKO addresses the instability of model-X knockoffs by introducing an aggregation process and proposes a new feature statistics AGC. We assess the performance under a wide range of simulation settings, and the simulation studies demonstrate that Grace-AKO has better performance in finite-sample

FDR control.

References

Axelrod, M. L., Cook, R. S., Johnson, D. B., and Balko, J. M. (2019). Biological consequences of MHC-II expression by tumor cells in cancer. *Clin Cancer Res*, 25(8):2392–2402.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Chung, F. R. K. (1997). *Spectral Graph Theory*. American Mathematical Society.

Cunningham, C. E., Li, S., Vizeacoumar, F. S., Bhanumathy, K. K., Lee, J. S., Parameswaran, S., Furber, L., Abuhussein, O., Paul, J. M., McDonald, M., Templeton, S. D., Shukla, H., El Zawily, A. M., Boyd, F., Alli, N., Mousseau, D. D., Geyer, R., Bonham, K., Anderson, D. H., Yan, J., Yu-Lee, L.-Y., Weaver, B. A., Uppalapati, M., Ruppin, E., Sablina, A., Freywald, A., and Vizeacoumar, F. J. (2016). Therapeutic relevance of the protein phosphatase 2a in cancer. *Oncotarget*, 7(38):61544–61561.

Dadwal, U. C., Chang, E. S., and Sankar, U. (2018). Androgen receptor-CaMKK2 axis in prostate cancer and bone microenvironment. *Front Endocrinol (Lausanne)*, 9:335.

Emery, K. and Keich, U. (2019). Controlling the FDR in variable selection via multiple knockoffs. *arXiv e-prints*, page arXiv:1911.09442.

Erb, H. H., Langlechner, R. V., Moser, P. L., Handle, F., Casneuf, T., Verstraeten, K., Schlick, B., Schäfer, G., Hall, B., Sasser, K., et al. (2013). Il6 sensitizes prostate cancer to the antiproliferative effect of ifnα2 through irf9. *Endocrine-related cancer*, 20(5):677.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 70(5):849–911.

Gimenez, J. R. and Zou, J. (2019). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 89:2184–2192.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

Kim, I., Choi, S., and Kim, S. (2018). BRCA-pathway: a structural integration and visualization system of TCGA breast cancer data on KEGG pathways. *BMC Bioinformatics*, 19(1):42.

Koboldt, D., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., Fulton, L., Dooling, D., Ding, L., Mardis, E., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

Koh, C. M., Bieberich, C. J., Dang, C. V., Nelson, W. G., Yegnasubramanian, S., and De Marzo, A. M. (2010). MYC and prostate cancer. *Genes Cancer*, 1(6):617–628.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–82.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4(3):1498–1516.

Liu, C., Zhang, L., Huang, Y., Lu, K., Tao, T., Chen, S., Zhang, X., Guan, H., Chen, M., and Xu, B. (2015). MicroRNA-328 directly targets p21-activated protein kinase 6 inhibiting prostate cancer proliferation and enhancing docetaxel sensitivity. *Mol Med Rep*, 12(5):7389–7395.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Melloy, P. (2019). The anaphase-promoting complex: a key mitotic regulator associated with somatic mutations occurring in cancer. *Genes, Chromosomes and Cancer*, 59(3):189–202.

Nguyen, T.-B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2020). Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR.

Rahnenführer, J., Domingues, F. S., Maydt, J., and Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical applications in genetics and molecular biology*, 3(1).

Ray, H., a. . J. R., Haughey, C., Hoey, C., Jeon, J., Murphy, R., Dura-Perez, L., McCabe, N., Downes, M., Jain, S., Boutros, P. C., Mills, I. G., and Liu, S. K. (2020). miR-191 promotes radiation resistance of prostate cancer through interaction with RXRA. *Cancer Letters*, 473:107–117.

Stangelberger, A., Waldert, M., and Djavan, B. (2008). Prostate cancer in elderly men. *Rev Urol*, 10(2):111–119.

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

Wan, J., Wu, Y., Ji, X., Huang, L., Cai, W., Su, Z., Wang, S., and Xu, H. (2020). IL-9 and IL-9-producing cells in tumor immunity. *Cell Commun Signal*, 18(1):50.

Wang, G., Zhao, D., Spring, D. J., and DePinho, R. A. (2018). Genetics and biology of prostate cancer. *Genes Dev*, 32(17-18):1105–1140.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.