Deep-gKnock: nonlinear group-feature selection with deep neural network

Guangyu Zhu
University of Rhode Island

Tingting Zhao
Northeastern University

Abstract

Feature selection is central to contemporary high-dimensional data analysis. Grouping structure among features arises naturally in various scientific problems. Many methods have been proposed to incorporate the grouping structure information into feature selection. However, these methods are normally restricted to a linear regression setting. To relax the linear constraint, we combine Deep Neural Networks (DNNs) with the recent Knockoffs technique, which has been successful in an individual feature selection context. We propose Deep group-feature selection using Knockoffs (Deep-gKnock) as a methodology for model interpretation and dimension reduction. Deep-gKnock performs *model-free* group-feature selection by controlling group-wise False Discovery Rate (gFDR). Our method improves the interpretability and reproducibility of DNNs. Experimental results on both high-dimensional synthetic and real data demonstrate that our method achieves superior power and accurate gFDR control compared with state-of-the-art methods.

1 Introduction

Feature selection for high-dimensional data is of fundamental importance for different applications across various scientific disciplines (Tang and Liu, 2014; Li et al., 2018). Grouping structure among features arises naturally in many statistical modeling problems. Common examples range from multilevel categorical features in a regression model to genetic markers from the same gene in genetic association studies. Incorporating the grouping structure information into feature selection can take advantage of the scientifically meaningful prior knowledge, increase the feature selection accuracy and improve the interpretability of the feature selection results (Huang et al., 2012).

In this paper, we focus on group-feature selection as an approach for model interpretation and dimension reduction in both linear and nonlinear settings. Our method can achieve stable feature selection results in a high-dimensional setting when $p > n$, which is usually a challenging problem for existing methods, where p is the number of features and n is the number of samples.

Group-feature selection has been studied from different perspectives. The group-Lasso, a generalization of the Lasso (Tibshirani, 1996), has been proposed as a mainstream approach to conduct group-wise feature selection (Yuan and Lin, 2006). To relax the linear constraint, Meier et al. (2008) extended the group-Lasso from linear regression to logistic regression. To speed up the computation for group-Lasso, Yang and Zou (2015) have further developed a more computationally tractable and efficient algorithm.

However, researchers have found that the feature selection results by Lasso and group-Lasso are sensitive to the choices of tuning parameters (Tibshirani, 1996; Su et al., 2016). In practice, the tuning parameter is often chosen by cross-validation (CV). But it has been reported that in high-dimensional settings the widely adopted CV typically tends to select a large number of false features (Bogdan et al., 2015). In order to ensure the selected features are correct and replicable, several methods have been proposed to preform feature selection while controlling the false discovery rate (FDR) which is the expected fraction of false selections among all selections.

Among them, Sorted L-One Penalized Estimation (SLOPE) (Bogdan et al., 2015) and Knockoffs (Barber et al., 2015; Candes et al., 2018) are state-of-the-art methods and have received the most attention. SLOPE was proposed to control the FDR in the classical multiple linear regression setting. SLOPE is defined to be the solution to a penalized objective function:

$$\arg\min_b \left\{ \frac{1}{2}\|y - Xb\|^2 + J_\lambda(b) \right\},$$

where $J_\lambda(b) = \sum_{i=1} \lambda_i |b|_{(i)}$, with $b \in \mathbb{R}^p$, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, and $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$ is the vector of sorted absolute values of coordinates of b. Brzyski et al. (2018) extended the SLOPE method as group-SLOPE to perform group-feature selection but it is only applicable to linear regression.

The notion of Knockoffs was first introduced by Barber et al. (2015) and later improved as model-X Knockoffs by Candes et al. (2018). The Knockoffs variables serve as negative controls and help identify the truly important features by comparing the feature importance between the original features and their Knockoffs counterpart. Originally, it is constrained to homoscedastic linear models with $n \geq p$ (Barber et al., 2015) and later extended to a group-sparse linear regression setting by Dai and Barber (2016).

In the current directions of SLOPE and Knockoffs, Group-SLOPE (Brzyski et al., 2018) and group-Knockoffs (Dai and Barber, 2016) are the only solution to group-feature selection. However, they suffer from the following limitations: (1) group-Knockoffs can only handle linear regression and are restricted to the $n > p$ setting; (2) group-SLOPE can only deal with linear regression and can not achieve robust feature selection results in a high-dimensional setting when $p > n$; (3) group-SLOPE does not provide an end-to-end group-wise feature selection and requires groups of features to be orthogonal to each other.

To resolve all the limitations, we propose Deep-gKnock (Deep group-feature selection using Knockoffs), which combines model-X Knockoffs and Deep Neural Networks (DNNs) to perform model-free group-feature selection in both linear and nonlinear settings while controlling the group-wise FDR (gFDR). DNNs are a natural choice to model complex nonlinear relationships and perform end-to-end deep representation learning (Kingma and Welling, 2013) for high-dimensional data. However, DNNs are often treated as black-box due to its lack of interpretability and reproducibility. Deep-gKnock constructs group Knockoffs features to perform group-feature selection for DNNs.

Figure 1 provides an overview for our Deep-gKnock procedure, which includes (1) generating Group Knockoffs features; (2) incorporating original features and Group Knockoffs features into a DNN architecture; (3) computing Knockoffs statistic; (4) filtering out the unimportant group-features using Knockoffs statistic. Experimental results demonstrate that our method achieves superior power and accurate FDR control compared with state-of-the-art methods.

To summarize, we make the following contributions: (1) end-to-end group-wise feature selection and deep representations for a $p > n$ setting; (2) flexible modeling framework using DNNs with enhanced interpretability and reproducibility; (3) comprehensive experimental results to characterize the performance of our approach with varying key parameters, model architecture changes and robustness to model misspecification; (4) superior performance in terms of power and controlled gFDR in both linear and nonlinear settings for high-dimensional synthetic and real-world genome-wide association studies in the $p \gg n$ regime. The results for our method are highly reproducible. Code for both our method and all benchmark methods in comparison will be made available on GitHub upon publication.

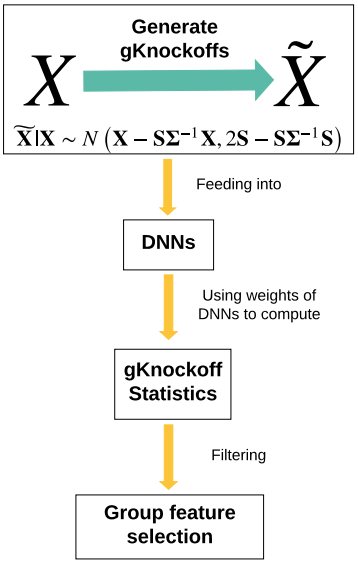

Figure 1: A graphical illustration of the four steps of Deep-gKnock. This figure is best viewed in color.

2 Background

2.1 Problem statement

In our problem, we have n independent and identically distributed (i.i.d) observations $\mathbf{X}_i, \mathbf{Y}_i$, where $\mathbf{X}_i \in \mathbb{R}^p$, $\mathbf{Y}_i \in \mathbb{R}^r$, $i = 1, \ldots, n$. We use $\mathbf{X}_i$ to denote the feature vector and $\mathbf{Y}_i$ to denote the response variable, which can be either discrete or continuous. Denote $\chi = \{1, 2, \ldots, p\}$. We assume there exists group structure within the p features, which can be partitioned into m groups with group sizes $p_1, \ldots, p_m$. The index of the features in the jth group is denoted as G_j, where $|G_j| = p_j$. It satisfies that $G_j \subset \chi$ for $j = 1, 2, \ldots, m$, $\cup_{j=1}^m G_j = \chi$ and $\cap_{j=1}^m G_j = \emptyset$. Assume that there exists a subset $\mathcal{S}_0 \subset \{1, \ldots, m\}$ such that conditional on the groups of features in $\mathcal{S}_0$, the response $\mathbf{Y}_i$ is independent of groups of features in the complement $\mathcal{S}_0^c$.

Denote $\hat{\mathcal{S}} \subset \{1, \ldots, m\}$ as the set of all the selected groups of features. Our goal is to ensure high True Positive Rate (TPR) defined as $\text{TPR} = \frac{|\hat{S} \cap S_0|}{|S_0|}$ while controlling the gFDR, which is the expected proportion of irrelevant groups among all groups of features selected and is defined as

$$\text{gFDR} = \text{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{S}_0^c|}{\max\{|\hat{\mathcal{S}}|, 1\}}\right].$$

2.2 Model-X Knockoffs framework review

The Knockoffs features are constructed as negative controls to help identify the truly important features by comparing the feature importance between the original and their Knockoffs counterpart. Model-X Knockoffs features are generated to perfectly mimic the arbitrary dependence structure among the original features but are conditionally independent of the response given the original features. However, model-X Knockoffs procedure (Candes et al., 2018) is only able to construct Knockoffs variables for individual feature selection. Our deep-gKnock procedure described in Section 3 extends model-X Knockoffs procedure to generate group Knockoffs features, which allows group structure among features.

For better understanding, we review the model-X Knockoffs method first. Model-X Knockoffs is designed for the individual feature selection and does not consider the grouping structure among features. So the $\mathcal{S}_0, \hat{\mathcal{S}}$ are defined as the indices of individual features, which are different from definitions in Section 2.1 . Model-X Knockoffs method assumes that there exists a subset $\mathcal{S}_0 \subset \{1, \ldots, p\}$ such that conditional on the features in $\mathcal{S}_0$, the response $\mathbf{Y}_i$ is independent of features in the complement $\mathcal{S}_0^c$. We denote $\hat{\mathcal{S}} \subset \{1, \ldots, p\}$ as the set of all selected individual features.

We start this section with the model-X Knockoffs feature definition, followed by the Knockoffs feature generation process and end with the filtering process for feature selection.

Definition 2.1 (Candes et al. (2018))**.** Suppose the family of random features $\mathbf{X} = (X_1, \cdots, X_p)^T$. Model-X Knockoffs features for $\mathbf{X}$ are a new family of random features $\widetilde{\mathbf{X}} = (\tilde{X}_1, \cdots, \tilde{X}_p)^T$ that satisfy two properties: (1) $(\mathbf{X}, \widetilde{\mathbf{X}})_{\text{swap}(\mathcal{S})} \overset{\text{d}}{=} (\mathbf{X}, \widetilde{\mathbf{X}})$ for any subset $\mathcal{S} \subset \{1, \cdots, p\}$, where $\text{swap}(\mathcal{S})$ means swapping X_j and $\tilde{X}_j$ for each $j \in \mathcal{S}$ and $\overset{\text{d}}{=}$ denotes equal in distribution, and (2) $\widetilde{\mathbf{X}} \perp \mathbf{Y} | \mathbf{X}$, i.e., $\widetilde{\mathbf{X}}$ is independent of response $\mathbf{Y}$ given feature $\mathbf{X}$.

From this definition, we can see that model-X Knockoffs feature $\tilde{X}_j$'s mimic dependency structure among the original features X_j's and are independent of the response $\mathbf{Y}$ given X_j's. By comparing the original features $\mathbf{X}$ with the Knockoffs features $\widetilde{\mathbf{X}}$, FDR can be controlled at a target level q. When $X \sim N(0, \mathbf{\Sigma})$ with the covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$, we can construct the model-X Knockoffs features $\widetilde{\mathbf{X}}$ characterized in Definition 2.1 as

$$\widetilde{\mathbf{X}} | \mathbf{X} \sim N(\mathbf{X} - \text{diag}\{\mathbf{s}\}\mathbf{\Sigma}^{-1}\mathbf{X}, 2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\mathbf{\Sigma}^{-1}\text{diag}\{\mathbf{s}\}). \tag{1}$$

Here $\text{diag}\{\mathbf{s}\}$ with all components of $\mathbf{s} \in \mathbb{R}^p$ being positive is a diagonal matrix which requires that the conditional covariance matrix in Equation 1 is positive definite. Following the above Knockoffs construction, the joint distribution of the original features and the model-X Knockoffs features is

$$(\widetilde{\mathbf{X}}, \mathbf{X}) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\} \\ \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\} & \mathbf{\Sigma} \end{pmatrix}\right). \tag{2}$$

To ensure high power in distinguishing $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, it is desired that the constructed Knockoffs features $\widetilde{\mathbf{X}}$ deviate from the original features $\mathbf{X}$ while maintaining the same correlation structure as $\mathbf{X}$. This indicates larger components of $\mathbf{s}$ are preferred since $\text{Cov}(\mathbf{X}, \widetilde{\mathbf{X}}) = \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\}$. In a setting where the features are normalized, i.e. $\mathbf{\Sigma}_{jj} = 1$ for all j, we would like to have $\text{Cor}(X_j, \tilde{X}_k) = 1 - s_j$ as close to zero as possible. One way to choose $\mathbf{s}$ is the equicorrelated construction (Barber and Candès, 2016), which uses

$$s_j^{\text{EQ}} = 2\lambda_{\min}(\mathbf{\Sigma}) \wedge 1 \text{ for all } j.$$

Then we define the Knockoffs statistic W_j for each feature X_j, $j \in \{1, \ldots, p\}$, which is used in the filtering process to perform feature selection. A large positive value of W_j provides evidence that X_j is important. This statistic depends on $\mathbf{X}, \widetilde{\mathbf{X}}$ and $\mathbf{Y}$, i.e. $W_j = w_j((\mathbf{X}, \widetilde{\mathbf{X}}), \mathbf{Y})$ for some function w_j. This function w_j must satisfy the following flip-sign property:

$$w_j\left([\mathbf{X}, \widetilde{\mathbf{X}}]_{\text{swap}(S)}, y\right) = \begin{cases} w_j([\mathbf{X}, \widetilde{\mathbf{X}}], y), & j \notin S \\ -w_j([\mathbf{X}, \widetilde{\mathbf{X}}], y), & j \in S \end{cases} \tag{3}$$

Candes et al. (2018) construct the Knockoffs statistic by performing Lasso on the original features $\mathbf{X}$ augmented with Knockoffs $\widetilde{\mathbf{X}}$

$$\min_{b \in \mathbb{R}^{2p}} \frac{1}{2}\|y - [\mathbf{X}, \widetilde{\mathbf{X}}]b\|_2^2 + \lambda\|b\|_1,$$

which provides Lasso coefficients $b_1, \ldots, b_{2p}$. The statistic W_j is set to be the Lasso coefficient difference given by

$$W_j = |b_j| - |b_{j+p}|.$$

After obtaining the Knockoffs statistic satisfying (3), Theorem 2.2 from Candes et al. (2018) provides feature selection procedure with controlled FDR.

Theorem 2.2 (Candes et al. (2018))**.** *Let $q \in [0,1]$. Given statistic, $W_1, \ldots, W_p$ satisfying* (3)*, let*

$$\tau = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\}.$$

Then the procedure selecting the features $\hat{S} = \{l : W_j \geq \tau\}$, controls the FDR at a target level q.

3 Deep group-feature selection using Knockoffs

3.1 Constructing Group Knockoffs features

The original Knockoffs construction (Candes et al., 2018) does not take group structure among different features into account and requires stronger constraints. When there exists high correlation between features X_j and $\tilde{X}_j$, Candes et al. (2018)'s method requires that the values of $\mathbf{s}$ to be extremely small in order to ensure the covariance matrix in Equation (2) is positive semi-definite. However, smaller values of $\mathbf{s}$ will fail to detect the difference between $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, which will lead to a decrease in the power of detecting the true positive features. In a group-sparse setting, we relax this requirement by proposing our Group Knockoffs features in Definition 3.1 to increase the power.

Definition 3.1 (Group Knockoffs features)**.** Suppose the family of random features $\mathbf{X} = (X_1, \cdots, X_p)^T$ has group structure, where the p features are partitioned into m groups, $G_1, \ldots, G_m \subset \chi = \{1, \ldots, p\}$, with group sizes $p_1, \ldots, p_m$, $\cup_{j=1}^m G_j = \chi$ and $\cap_{j=1}^m G_j = \emptyset$. Group Knockoffs features for $\mathbf{X} = (X_1, \cdots, X_p)^T$ are a new family of random features $\widetilde{\mathbf{X}} = (\tilde{X}_1, \cdots, \tilde{X}_p)^T$ that satisfy two properties: (1) $(\mathbf{X}, \widetilde{\mathbf{X}})_{\text{swap}(\mathcal{S})} \overset{\text{d}}{=} (\mathbf{X}, \widetilde{\mathbf{X}})$ for any subset $\mathcal{S} \subset \{1, \cdots, m\}$, where swap$(\mathcal{S})$ means swapping $\mathbf{X}_{G_j}$ and $\widetilde{\mathbf{X}}_{G_j}$ for each $j \in \mathcal{S}$ and $\overset{\text{d}}{=}$ denotes equal in distribution, and (2) $\widetilde{\mathbf{X}} \perp Y | \mathbf{X}$, i.e., $\widetilde{\mathbf{X}}$ is independent of the response $\mathbf{Y}$ given feature $\mathbf{X}$.

We see from this definition, that the Group Knockoffs features $\tilde{X}_j$'s mimic the group-wise dependency structure among the original features X_j's and are independent of the response $\mathbf{Y}$ given X_j's. When $X \sim N(0, \mathbf{\Sigma})$, the joint distribution obeying Definition 3.1 is

$$(\widetilde{\mathbf{X}}, \mathbf{X}) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mathbf{\Sigma} & \mathbf{\Sigma} - \mathbf{S}\\ \mathbf{\Sigma} - \mathbf{S} & \mathbf{\Sigma}\end{pmatrix}\right).$$

where $\mathbf{S} = \text{diag}(\mathbf{S}_1, \ldots, \mathbf{S}_m) \prec 2\mathbf{\Sigma}$ is a group-block-diagonal matrix. Here we use $\mathbf{A} \prec \mathbf{B}$ to denote $\mathbf{B} - \mathbf{A}$ is positive definite.

We construct the Group Knockoffs features by sampling the Knockoffs vector $\widetilde{\mathbf{X}}$ from the conditional distribution

$$\widetilde{\mathbf{X}} | \mathbf{X} \sim N(\mathbf{X} - \mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{X}, 2\mathbf{S} - \mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{S}).$$

Following Dai and Barber (2016), the group-block-diagonal matrix $\mathbf{S} = \text{diag}(\mathbf{S}_1, \ldots, \mathbf{S}_m)$ satisfying $\mathbf{S} \prec 2\mathbf{\Sigma}$ can be constructed with

$$\mathbf{S}_i = \eta \mathbf{\Sigma}_{G_i, G_i},$$

where

$$\eta = 2\lambda_{\min}(\mathbf{D}\mathbf{\Sigma}\mathbf{D}) \wedge 1,\ \mathbf{D} = \text{diag}\{\mathbf{\Sigma}_{G_1 G_1}^{-1/2}, \ldots, \mathbf{\Sigma}_{G_m G_m}^{-1/2}\}.$$

3.2 Deep neural networks for Group Knockoffs features

Once the Group Knockoffs features are constructed, following similar idea in DeepPINK (Lu et al., 2018), we feed them into a new DNN structure to obtain gKnock statistic. The structure of the network is shown in Figure 2.

In the first layer, we feed $(\mathbf{X}, \widetilde{\mathbf{X}})$ into a Group-feature Competing Layer containing m filters, $G_1, \ldots, G_m$. The jth filter G_j connects group-feature $\mathbf{X}_{G_j}$ and its Knockoffs counterpart $\widetilde{\mathbf{X}}_{G_j}$. We use a linear activation function in this layer to encourage the competition between group-feature and its Knockoffs counterpart. Intuitively, if the group-feature $\mathbf{X}_{G_j}$ is important, we expect the magnitude of S_j to be much larger than $\tilde{S}_j$, and if the the group-feature $\mathbf{X}_{G_j}$ is not important, we expect the magnitude of S_j and $\tilde{S}_j$ to be similar.

We then feed the output of the Group-feature Competing Layer into a fully connected Multilayer Perceptron (MLP) to learn a non-linear mapping to the response $\mathbf{Y}$. We use $W^{(0)} \in \mathbb{R}^{m \times 1}$ to denote the weight vector connecting the Group-features Competing Layer to the MLP. The MLP has two hidden layers, each containing m neurons, and ReLU activation and L_1-regularization are used, as shown in Figure 2. We use $W^{(1)} \in \mathbb{R}^{m \times m}$ to denote the weight matrix connecting the input vector to the first hidden layer. Similarly, we use $W^{(2)} \in \mathbb{R}^{m \times m}$ as the weight matrix connecting two hidden layers and $W^{(3)} \in \mathbb{R}^{m \times 1}$ as the weight matrix connecting the second hidden layer to the output $\mathbf{Y}$.

3.3 gKnock statistic

After the DNN is trained, we compute the gKnock statistic based on the weights to evaluate the importance of the group-feature. Firstly, we use $\mathbf{z} = (\|S_1\|_2^2/p_1, \ldots, \|S_m\|_2^2/p_m)^T$ and $\tilde{\mathbf{z}} =$

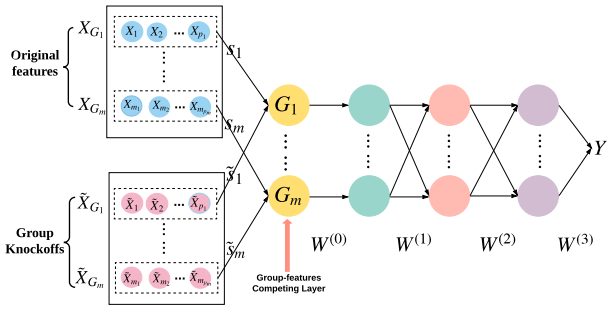

Figure 2: A graphical demonstration of the DNN structure for Deep-gKnock. This figure is best viewed in color.

$(\|\tilde{S}_1\|_2^2/p_1, \ldots, \|\tilde{S}_m\|_2^2/p_m)^T$ to represent the relative importance between $\mathbf{X}_{G_j}$ and $\widetilde{\mathbf{X}}_{G_j}$, $j = 1, \ldots, m$. Secondly, we assess the relative importance of the jth group-feature among all m group-feature by $\mathbf{w} = W^{(0)} \circ (W^{(1)}W^{(2)}W^{(3)}) \in \mathbb{R}^{m\times 1}$, where $\circ$ denotes the Schur (entrywise) matrix product. Thirdly, the importance measures for $\mathbf{X}_{G_j}$ and $\widetilde{\mathbf{X}}_{G_j}$ are provided by

$$Z_j = \|S_j\|_2^2 \times w_j \quad \text{and} \quad \tilde{Z}_j = \|\tilde{S}_j\|_2^2 \times \tilde{w}_j.$$

Finally, we define the gKnock statistic as

$$W_j = Z_j^2 - \tilde{Z}_j^2, \quad j = 1, \ldots, m$$

and the same filtering process as Theorem 2.2 is applied to the W_j's to select group-feature.

4 Simulation studies

In this section, to properly characterize the performance of our approach and, more generally, determine when it would be advantageous to use it, we investigate the effects of key parameters and different model architectures on the group-feature selection performance of our approach in both a Gaussian linear regression model (4) and a nonlinear Single-Index model (5).

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{4}$$
$$Y_i = g(\mathbf{X}_i^T\boldsymbol{\beta}) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{5}$$

where $Y_i \in \mathbb{R}$ is the ith response, $\mathbf{X}_i \in \mathbb{R}^p$ is the feature vector of the ith observation, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient vector, $\epsilon_i \in \mathbb{R}^p$ is the noise of ith observation, and g is some unknown link function.

To be specific, we explore (1) the effects of key parameters including the sample size, the number of associated groups (sparsity), between-group correlation, within-group correlation and the noise level on the feature selection performance; (2) the robustness of the group-wise knockoffs construction to model misspecification; (3) the effects of model architecture changes (such as the model gets deeper) on the feature selection performance.

We provide comprehensive simulation experiments by comparing the performance of Deep-gKnock with multiple state-of-the-art methods including groupSlope (Gossmann et al., 2016), DeepPink (Lu et al., 2018), group-knockoff (Dai and Barber, 2016), knockoff (Candes et al., 2018), and multilayer (Katsevich et al., 2019).

To generate the synthetic data, we set the number of features $p = 1000$ and the number of groups $g = 100$ with the number of features per group as $p_i = 10$. The true regression coefficient vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is group sparse with $k = 20$ groups of nonzero signals, and the nonzero coefficients are randomly chosen from $\{\pm 1.5\}$. We draw X_i independently from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, with diagonal entries $\Sigma_{ii} = 1$, within-group correlations $\Sigma_{ij} = \rho$ for $i \neq j$ in the same group, between-group correlations $\Sigma_{ij} = \gamma\rho$ for i, j in the different groups. The errors ϵ_i are i.i.d. from a standard normal distribution. The true link function is $g(x) = (x/20)^3 + 4(x/20)^2$.

In our default setting, we set $n = 1000$, $\rho = \gamma = 0$. To study the effects of key parameters including the sample size, between-group correlation and within-group correlation on the group-feature selection performance, we vary one setting and keep the others remain at their default level in each experiment.

- Sample size n: we vary the number of observations from 500, 750, 1000, 1250 to 1500.
- Number of associated groups (sparsity) k: we fix $\rho = \gamma = 0.5$ and vary k from 10, 15, 20, 25, 30.
- Group correlation: we fix the within-group correlation $\rho = 0.5$, and set the between-group correlation to be $\gamma\rho$, with $\gamma \in \{0, 0.2, \ldots, 0.8\}$.
- Within-group correlation: we vary within-group correlation with $\rho \in \{0, 0.2, \ldots, 0.8\}$ and fix $\gamma = 0.4$.
- Noise level σ^2: we vary σ^2 values from 1, 2, 3, 4.

For each setting, we run each experiment for 100 replications and set the target gFDR level at $q = 0.2$. The empirical gFDR and power are reported in Figure 3 and Figure 4.

In the linear model setting shown in Figure 3, group-SLOPE fails to control the gFDR at the target level

gFDR = 0.2 in each of the following three situations: (1) $p > n$; (2) between-group correlation γ is large; (3) within group correlation ρ is large. In contrast, Deep-gKnock can precisely control the gFDR in all settings. All methods except multilayer can achieve high power with varying sample sizes. Only DeepPink can achieve comparable power as our method in all settings. When the between-group correlation and the within-group correlation increases, we observe degradation performance in power for all methods except DeepPink and Deep-gKnock. All methods are robust to varying noise levels while Deep-gKnock achieves high power and successfully controls gFDR. Knockoff and DeepPink are not able to control gFDR near the target level since they are not specifically designed for a group-feature selection setting.

In the Single-Index model setting shown in Figure 4, Deep-gKnock achieves highe power and consistently controls gFDR in all settings, which demonstrates the advantages of our Deep-gKnock by using DNNs to model the non-linear relationship between features and the response. The power of gknockoff, grpSLOPE, multilayer, knockoff is relatively low compared with DeepPink and Deep-gKnock since these methods are not designed for nonlinear models. In comparison, the power of DeepPink and Deep-gKnock is markedly better than other methods. The power of DeepPink and Deep-gKnock increases with (1) increasing the sample size increases; (2) increasing the number of associated groups (3) increasing between-group correlation and (4) increasing within group correlation.

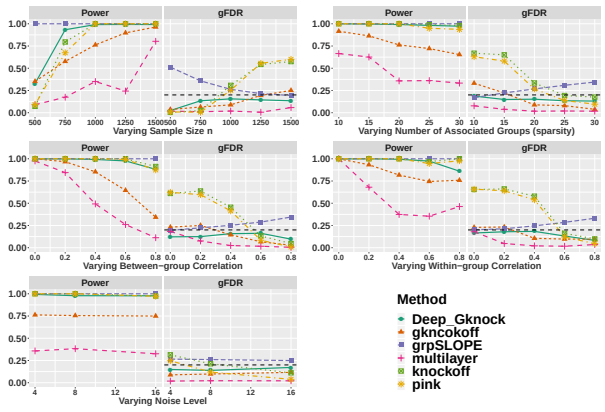

Figure 3: Simulation results for a linear model.

To check the robustness to model misspecification, under the linear model, we also generate $\mathbf{X}$ from a multivariate t distribution with df $v = 3$. The results are summarized in Figure 5. The performance of each method is similar to the performance as shown in Figure 3 when the design matrix is from a multivariate normal distribution.

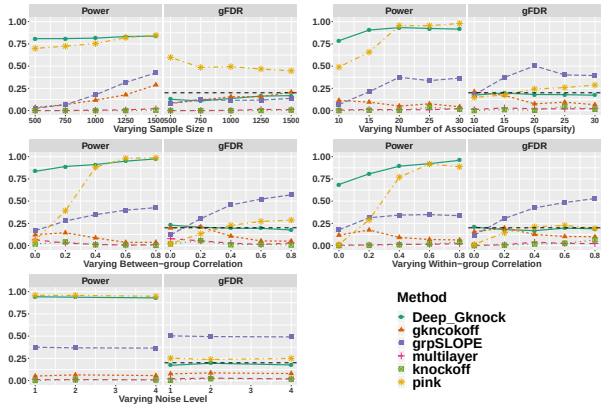

Figure 4: Simulation results for a Single-Index model.

To explore the effects of model architecture changes on the feature selection performance, we report the results in Figure 6. It shows that the group-feature selection performance is robust to model architecture changes in terms of different numbers of layers in the model.

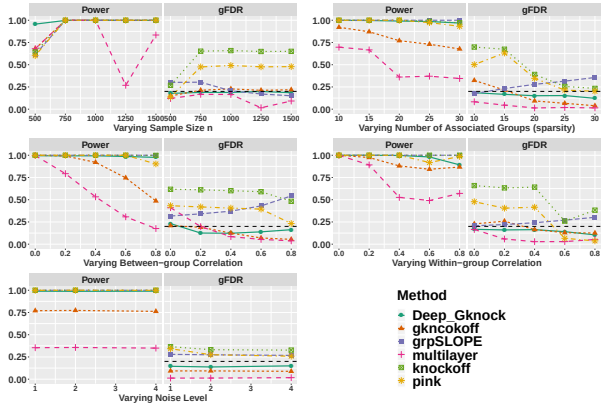

Figure 5: Model misspecification: multivariate t distributed design matrix under a linear model.

5 Real data analysis

In addition to the simulation studies presented in Section 4, we also demonstrate the performance of Deep-gKnock on three real datasets. The third dataset is a representative of the real-world genome-wide association studies in the $p \gg n$ regime. The gFDR level is set to $q = 0.2$.

5.1 Application to prostate cancer data

The prostate cancer data contains clinical measurements for 97 male patients who were about to receive a radical prostatectomy. It was analyzed in Hastie

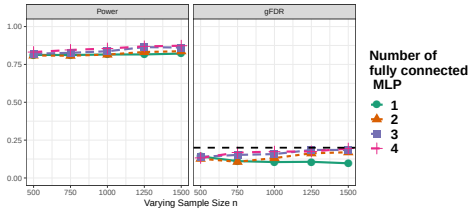

Figure 6: Feature selection robustness to model architecture changes.

et al. (2013) to study the correlation between the response $\mathbf{Y}$, the level of prostate-specific antigen (lpsa) and other eight features. The features are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45). For the categorical variable svi with two levels, we coded it as one dummy variable and treated it as one group. For each continuous variable, we used five B-Spline basis functions to represent its effect and treated those five basis functions as a group. This provides us eight groups with a total of 36 features. We summarize the group-feature selection results in Table 1. The features selected by Deep-gKnock are the same as using Lasso in Hastie et al. (2013).

Table 1: Group-feature selection results for prostate cancer data

Method	group-feature selected
group-SLOPE	lcavol, lweight, svi, gleason
Deep-gKnock	lcavol, lweight

5.2 Application to yeast cell cycle data

We apply Deep-gKonck to the task of identifying the important transcription factors (TFs), which are related to regulation of the cell cycle. TFs belong to a class of proteins called binding proteins, and control the rate at which DNA is transcribed into mRNA. We utilize a yeast cell cycle dataset from Spellman et al. (1998) and Lee et al. (2002). The response $\mathbf{Y}$ is the messenger ribonucleic acid (mRNA) levels on 542 genes, and are measured at 28 minutes during a cell cycle. The features $\mathbf{X}$ is the measurements of binding information of 106 TFs . Out of the 106 TFs, 21 TFs are known and experimentally confirmed cell cycle related TFs (Wang et al., 2007).

It has been studied that groups of TFs function in a coordinated fashion to direct cell division, growth and death (Latchman, 1997). Following Ma et al. (2007), we use the K-means method to cluster the 106 TFs, and determine the optimal number of clusters using the Gap statistic (Tibshirani et al., 1999). The Gap statistic suggests the 106 TFs can be clustered into 20 groups. To visualize the clustering results, we use Principal Component Analysis (PCA) algorithm to reduce the dimensionality to its first two principal components, which results in a scatter plot of data points colored by their cluster labels in Figure 7. One of the clusters contains four TFs and all of them are experimentally verified.

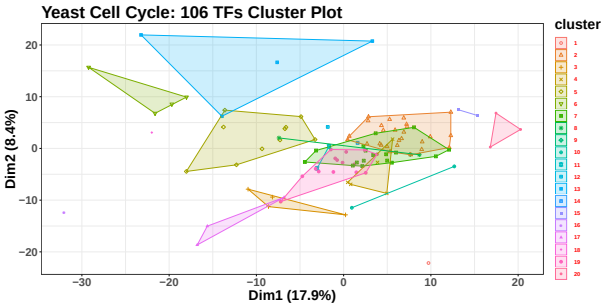

Figure 7: Cluster plot for 106 TFs in Yeast Cell Cycle data

Group-SLOPE identified 7 groups which contains 41 TFs, 12 of which are confirmed. Deep-gKnock identified 5 groups which contains 26 TFs, 11 of which are confirmed. To demonstrate the selection performance, following Zhu and Su (2019), we also compute the probability of containing at least q confirmed TFs from a s randomly chosen TFs from a hypergeometric distribution in Table 2. We included the results for Lasso and group-SLOPE in Table 2 as benchmarks. Smaller probability values suggest better feature selection performance. The small probability of Deep-gKnock suggests that the large number of confirmed TFs selected is not due to chance. Deep-gKnock also outperforms group-SLOPE.

Table 2: Probability of containing at least q confirmed TFs out of 85 unconfirmed and 21 confirmed TFs in a random draw of s TFs.

Method	s	q	$P(Q \geq q)$
Lasso	100	21	0.256
group-SLOPE	41	12	0.04673
Deep-gKnock	26	11	0.00192

5.3 Application to colon cancer data

Development in microarray techniques makes it possible to identifying influential genes that are associated with occurrence or progression of diseases from a whole

genome scale. We apply Deep-gKnock to the task of identifying a small subset of genes that are linked with the development of colon cancer. Alon et al. (1999) used Affymetrix Oligonucleotide Array to measure expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes. The data was processed, filtered, and reduced to a subset of 2,000 gene expression values with the largest minimal intensity over the 62 tissue samples. We standardized each tissue sample to zero mean and unit variance across the genes.

Gene data often presents group structure where the groups consist of co-regulated genes with coordinated functions. We use the K-means method to cluster the 2000 genes. The Gap statistic suggests the 2000 genes can be clustered into 20 groups. A visualization of the clustering results is provided in Figure 8.

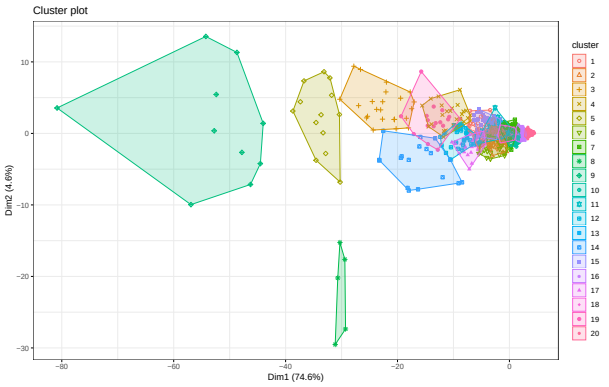

Figure 8: Clustering plot for 2000 genes in colon cancer data.

Deep-gKnock identified 13 groups which contain 621 genes. In order to get a more parsimonious model and identify important individual genes, we use Deep-gKnock again to select representative genes within each group. There are 26 genes presented in the final model, and many of them have been confirmed by previous studies. Gene Hsa.462 is found to be related to cancer cell proliferation. Gene Hsa. 37937 is identified as a tumor suppressor (Yam et al., 2001). Gene Hsa.627 is also identified as a Colon cancer biomarker in Zhang et al. (2005).

6 Conclusion

We have introduced a novel group-feature selection method Deep-gKnock combining Knockoffs with DNNs. It is an end-to-end group-wise feature selection approach with controlled gFDR for high-dimensional data. With the flexibility of DNN, we also provide deep representations with enhanced interpretability and reproducibility. Both synthetic and real data analysis is provided to demonstrate that Deep-gKnock can achieve superior power and accurate gFDR control compared with state-of-the-art methods. Moreover, Deep-gKnock achieves scientifically meaningful group-feature selection results for cutting-edge real world datasets.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750.

Barber, R. F. and Candès, E. J. (2016). A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574*.

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

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103.

Brzyski, D., Gossmann, A., Su, W., and Bogdan, M. (2018). Group slope–adaptive selection of groups of predictors. *Journal of the American Statistical Association*, pages 1–15.

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.

Dai, R. and Barber, R. (2016). The knockoff filter for fdr control in group-sparse and multitask regression. In *International Conference on Machine Learning*, pages 1851–1859.

Gossmann, A., Brzyski, D., Su, W., and Bogdan, M. (2016). *grpSLOPE: Group Sorted L1 Penalized Estimation*. R package version 0.2.1.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4).

Katsevich, E., Sabatti, C., et al. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1–33.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Latchman, D. S. (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *science*, 298(5594):799–804.

Li, Z., Xie, W., and Liu, T. (2018). Efficient feature selection and classification for microarray data. *PloS one*, 13(8):e0202167.

Lu, Y., Fan, Y., Lv, J., and Noble, W. S. (2018). Deeppink: reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 8690–8700.

Ma, S., Song, X., and Huang, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC bioinformatics*, 8(1):60.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297.

Su, Z., Zhu, G., Chen, X., and Yang, Y. (2016). Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika*, 103(3):579–593.

Tang, J. and Liu, H. (2014). Feature selection for social media data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(4):19.

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

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., Brown, P., et al. (1999). Clustering methods for the analysis of dna microarray data. *Dept. Statist., Stanford Univ., Stanford, CA, Tech. Rep.*

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Yam, J. W. P., Chan, K. W., and Hsiao, W.-L. W. (2001). Suppression of the tumorigenicity of mutant p53-transformed rat embryo fibroblasts through expression of a newly cloned rat nonmuscle myosin heavy chain-b. *Oncogene*, 20(1):58.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, X. W., Yap, Y. L., Wei, D., Chen, F., and Danchin, A. (2005). Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. *European Journal of Human Genetics*, 13(12):1303.

Zhu, G. and Su, Z. (2019). Envelope-based sparse partial least squares. *The Annals of Statistics*, (in press).