PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

Error-based Knockoffs Inference for Controlled Feature Selection

Xuebin Zhao[1], Hong Chen[1,*], Yingjie Wang[2], Weifu Li[1], Tieliang Gong[3], Yulong Wang[2], Feng Zheng[4]

[1]College of Science, Huazhong Agricultural University, Wuhan 430062, China
[2]College of Informatics, Huazhong Agricultural University, Wuhan 430062, China
[3]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[4]Department of Computer Science and Engineering, Southern University of Science and Technology, China
chenh@mail.hzau.edu.cn;

Abstract

Recently, the scheme of model-X knockoffs was proposed as a promising solution to address controlled feature selection under high-dimensional finite-sample settings. However, the procedure of model-X knockoffs depends heavily on the coefficient-based feature importance and only concerns the control of false discovery rate (FDR). To further improve its adaptivity and flexibility, in this paper, we propose an error-based knockoff inference method by integrating the knockoff features, the error-based feature importance statistics, and the stepdown procedure together. The proposed inference procedure does not require specifying a regression model and can handle feature selection with theoretical guarantees on controlling false discovery proportion (FDP), FDR, or k-familywise error rate (k-FWER). Empirical evaluations demonstrate the competitive performance of our approach on both simulated and real data.

Introduction

Data-driven feature selection aims to uncover informative features associated with the response to tailor interpretable statistical inference. Based on regression estimation, various regularized models have been formulated for sparse feature selection (Hastie and Tibshirani 1990; Fan and Li 2001; Lin and Zhang 2007; Liu, Chen, and Huang 2020; Chen et al. 2021). Following this line, typical methods include Lasso (Tibshirani 1996), group Lasso (Yuan and Lin 2006; Bach 2008; Friedman, Hastie, and Tibshirani 2010), LassoNet (Lemhadri, Ruan, and Tibshirani 2021), SpAM (Liu et al. 2008), GroupSpAM (Yin, Chen, and Xing 2012), and regression models with automatic structure discovery (Pan and Zhu 2017; Frecon, Salzo, and Pontil 2018; Wang et al. 2020). It should be noticed that the above-mentioned works mainly concern the algorithm's power performance to select true informative features. However, it is still largely undeveloped to carry out feature selection while explicitly controlling the number of false discoveries (Hochberg and Tamhane 1987; Benjamini and Hochberg 1995; Lehmann and Romano 2005; Candès et al. 2018).

It is well known that false discovery control is crucial to enable interpretable machine learning in many real-world applications, e.g., genetic analysis where the cost of examining a falsely selected gene may be intolerable. Hochberg and Tamhane (1987) proposed a method to control the probability of selecting one or more false discoveries, while it may lead to low power in high dimensional settings. Benjamini and Hochberg (1995) formulated an approach to control the expect value of false discovery proportion (FDP), which is called FDR control. To balance the selection accuracy and the power, some trade-off models (Korn et al. 2004; Lehmann and Romano 2005) are constructed for controlling the probability of selecting k or more false discoveries (k-FWER control), or the probability of FDP exceeding a fixed level (FDP control). Besides the above works, there are extensive studies on feature selection with FWER control (Farcomeni 2008), FDR control (Benjamini and Yekutieli 2001; Efron and Tibshirani 2002; Genovese and Wasserman 2004; Fan, Guo, and Hao 2012; Liu and Shao 2014), and FDP control (Fan and Lv 2010; Delattre and Roquain 2015). However, most of them either assume a specific dependent structure between the response and argument (such as linear structure) or rely on p-value to evaluate the significance of each feature. The structure assumption may be too restrictive in many applications, where the response could depend on input features through very complicated forms. In addition, the classical p-value calculation procedures usually depend on the large-sample asymptotic theory, which may be no longer justified under high-dimensional finite-sample settings (Candès et al. 2018; Fan, Demirkaya, and Lv 2019).

Knockoff Filter

Recently, novel knockoff statistics have been constructed in (Barber and Candès 2015; Candès et al. 2018; Lu et al. 2018; Bai et al. 2020; Fan et al. 2020a,b; Liu et al. 2020; Sesia et al. 2020) to evaluate the contribution of each feature to the corresponding response. In particular, theoretical analysis demonstrates that irrelevant features' statistics are independent and symmetrically distributed without making any assumption on the sample size, the number of dimensions, or the dependent structure. This property is then used to discover informative features with FDR control. Beyond identifying informative features, a new testing procedure (called conditional randomization test) is developed in (Candès et al. 2018), which can estimate the distribution of knockoff statistics and construct the valid p-values under fi-

*Corresponding author.
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

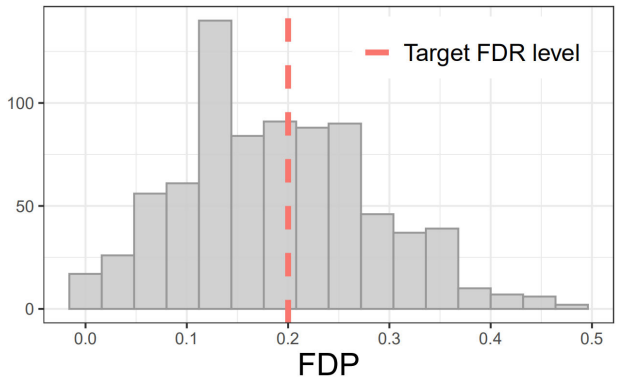

Figure 1: FDP under a given target FDR level

nite sample settings by repeatedly training the learning model.

Rapid progress has been made in recent years on understanding the theoretical behavior of knockoff techniques. Candès et al. (2018) proved that the model-X knockoffs (MX-Knockoff) framework enjoys tight FDR control when the covariant distribution is known, and feature importance statistic satisfies some mild conditions (e.g., shown in Proposition 1). Moreover, some refined works in Barber, Candès, and Samworth (2020); Fan et al. (2020a,b) have demonstrated that the knockoff procedures can also control FDR with asymptotic probability one even the covariant distribution follows some unknown Gaussian graphical model. In addition, the power of knockoff filters is guaranteed for RANK (Fan et al. 2020a), IPAD (Fan et al. 2020b) and (Weinstein et al. 2020) under the linear model assumption. Recently, a two-step approach has been proposed in (Liu et al. 2020) based on the projection correlation and the model-X knockoff features, which has both sure screening and rank consistency under weak assumptions.

Despite the success of knockoff filters, some issues remain to be further investigated:

- *FDR Vs. FDP and k-FWER*. The control of FDR does not assure the control of FDP (Genovese and Wasserman 2004). To illustrate such phenomenon, we display the histogram of FDP when applying MX-Knockoff (Candès et al. 2018) in 800 randomly generated datasets in Figure 1, which shows FDP can significantly exceed the target FDR level. See *Supplementary Material B* for details of our simulated example and related discussions. In addition, as pointed out in (Farcomeni 2008), k-FWER control is more desirable than FDR control when a powerful selection result can be made.
- *Coefficient-based feature statistic Vs. Coefficient-free feature statistic*. Under MX-Knockoff framework, feature importance is usually measured by the *coefficient difference*, e.g., (Candès et al. 2018; Fan et al. 2020a,b). However, it may be difficult to obtain the feature importance from general nonlinear models (Christian 2012; Liu et al. 2020). Indeed, it is an open question to design new feature importance statistics (see Section 7.2.5 in (Candès et al. 2018)), e.g., coefficient-free statistics.
- *Conditional randomization test Vs. Computational friendly test*. Although the p-value of each feature can be calculated via the conditional randomization test, this procedure is required to train the learning model multiple times. It will cause heavy computational burdens when calculating valid p-values, especially in high-dimensional finite-sample case (Candès et al. 2018). Thus, it is an open question how to efficiently calculate valid p-values via knockoff technique (see Section 7.2.6 in (Candès et al. 2018)).

Main Contributions

To address the above issues, this paper proposes a new knockoff filter scheme, called *Error-based Knockoffs Inference* (E-Knockoff), for controlled feature selection based on the error-based feature statistics. The main contributions of this paper are summarized as below:

- *Error-based knockoffs inference*. Our model integrates the knockoff features (Candès et al. 2018), the error-based feature statistics and the stepdown procedure (Lehmann and Romano 2005) into a coherent way for FDR, FDP or k-FWER control. The error-based importance measure does not require specifying a regression model and can be used to calculate the valid p-values efficiently. The stepdown procedure, with the help of new importance statistics, provides the route to control FDP and k-FWER, respectively, which is different from the previous knockoffs inference just for FDR control (Barber and Candès 2015; Candès et al. 2018; Lu et al. 2018; Fan et al. 2020a,b; Liu et al. 2020). In particular, it is novel to design the error-based statistics of feature importance under the knockoff framework, which partially answers the open questions stated in Sections 7.2.5 and 7.2.6 (Candès et al. 2018).
- *Theoretical guarantees on k-FWER, FDP, and FDR control*. For the MX-Knockoff framework, statistical foundations on the power and FDR control have been provided in (Candès et al. 2018; Fan et al. 2020a,b), where the power analysis is limited to high-dimensional linear models with both known and unknown covariate distribution. Beyond the linear models in aforementioned literature, we state theoretical justifications on the tight k-FWER control and FDP control when the stepdown procedure is employed. In particular, the robustness of k-FWER and FDP control can also be assured even for unknown covariate distribution associated with Gaussian graphical model. Additionally, our power analysis holds for general nonlinear models, which is closely related to the open question illustrated in Section 6 (Fan et al. 2020a). Some empirical evaluations support our theoretical findings.

To better illustrate the novelty of current work, we compare it with FX-Knockoff (Barber and Candès 2015), MX-Knockoff (Candès et al. 2018), DeepPINK (Lu et al. 2018), RANK (Fan et al. 2020a), PC-Knockoff (Liu et al. 2020) in Table 1 from the lens of feature statistics, control ability, and asymptotic theory. Table 1 shows that our approach

Table 1: Algorithmic properties (✓-has the given information, ×-hasn't the given information)

Properties	FX-Knockoff	MX-Knockoff	DeepPINK	RANK	PC-Knockoff	E-Knockoff (Ours)
Coefficient-free feature statistics	×	×	×	×	✓	✓
k-FWER control	×	×	×	×	×	✓
FDP control	×	×	×	×	×	✓
FDR control	✓	✓	✓	✓	✓	✓
Robust analysis	×	×	×	✓	×	✓
Power analysis (linear model)	×	×	×	✓	✓	✓
Power analysis (nonlinear model)	×	×	×	×	✓	✓

enjoys theoretical guarantees on robustness and power for FDR, FDP, and k-FWER control.

Preliminaries

This section introduces some necessary backgrounds including the problem setup, the knockoff filter (Candès et al. 2018) and the stepdown procedure (Lehmann and Romano 2005).

Problem Statement

Let $\mathcal{X} \subset \mathbb{R}^p$ be the compact input space and let $\mathcal{Y} \subset \mathbb{R}$ be the output set. We have n independent identically distributed (i.i.d.) observations $\{(\mathbf{x}_i, Y_i)\}_{i=1}^n$ from the population $(\mathbf{x}, Y)$, where $\mathbf{x} = (X_1, \ldots, X_p) \in \mathcal{X}$ and $Y \in \mathcal{Y}$. Suppose that the conditional distribution of Y is only relevant with a small subset of p covariates.

The definition of irrelevant features is given in Candès et al. (2018).

Definition 1 *(Candès et al. 2018) A feature X_j is said to be "irrelevant" if Y is independent of X_j conditionally on*

$$\mathbf{x}_{-j} := (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p).$$

For simplicity, we denote it as

$$Y \perp\!\!\!\perp X_j | \mathbf{x}_{-j}.$$

Remark 1 *If feature X_j is irrelevant according to Definition 1, it satisfies that the conditional distribution $Y|\mathbf{x} \overset{d}{=} Y|\mathbf{x}_{-j}$, where $\overset{d}{=}$ denotes equality in distribution. That is, the conditional distribution of Y remains invariant when removing X_j from $\mathbf{x}$.*

Let $\mathcal{S}_1 \subset \{1, \ldots, p\}$ be the index set with respect to irrelevant features. Naturally, the index set of true informative features $\mathcal{S}_0$ is the complement set of $\mathcal{S}_1$, i.e., $\mathcal{S}_0 = \mathcal{S}_1^c$. This paper aims to find $\widehat{\mathcal{S}}$, the data dependent estimation of $\mathcal{S}_0$, while controlling k-FWER, FDP, or FDR. Recall that

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}_1|}{|\widehat{\mathcal{S}}| \vee 1}\right]$$

and

$$k\text{-FWER} = Prob\{|\widehat{\mathcal{S}} \cap \mathcal{S}_1| \geq k\},$$

where $|\cdot|$ is the cardinality of a set.

Model-X Knockoff Framework

Model-X knockoff framework (Candès et al. 2018) aims to identify informative features while controlling FDR. The key point is to construct the knockoff copy of $\mathbf{x}$ which looks like real ones without contribution to the response.

Definition 2 *(Candès et al. 2018) Model-X knockoffs for the family of random variables $\mathbf{x} = (X_1, \ldots, X_p)$ is a new family of random variables $\tilde{\mathbf{x}} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ satisfying*

$$\tilde{\mathbf{x}} \perp\!\!\!\perp Y | \mathbf{x} \tag{1}$$

and

$$(\mathbf{x}, \tilde{\mathbf{x}})_{swap(s)} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}}), \forall s \subset \{1, \ldots, p\}. \tag{2}$$

Here, $(\mathbf{x}, \tilde{\mathbf{x}}) = (X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p)$, and $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(s)}$ is swapping X_j with $\tilde{X}_j$ for all $j \in s$, e.g., when $p = 3$, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(\{2,3\})} = (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$.

Remark 2 *The properties of knockoff features have been well investigated in (Candès et al. 2018). The property* (1) *illustrates that all knockoff features are noise features, and* (2) *assures the similarity between $\mathbf{x}$ and $\tilde{\mathbf{x}}$.*

Candès et al. (2018) state the construction of knockoffs when $\mathbf{x}$ obeys a known Gaussian graphical model $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where the covariance matrix $\mathbf{\Sigma}$ is positive definite. Model-X knockoff can construct $\tilde{\mathbf{x}}$ conditionally on $\mathbf{x}$ w.r.t. $\tilde{\mathbf{x}}|\mathbf{x} \overset{d}{=} \mathcal{N}(\mu, \mathbf{V})$, where

$$\mu = \mathbf{x}(I_p - \mathbf{\Sigma^{-1}}\text{diag}\{\mathbf{s}\}),$$

$$\mathbf{V} = 2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\mathbf{\Sigma^{-1}}\text{diag}\{\mathbf{s}\},$$

and the joined distribution of $(\mathbf{x}, \tilde{\mathbf{x}})$ satisfies

$$(\mathbf{x}, \tilde{\mathbf{x}}) \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$$

with

$$\mathbf{G} = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\} \\ \mathbf{\Sigma} - \text{diag}\{\mathbf{s}\} & \mathbf{\Sigma} \end{pmatrix}. \tag{3}$$

Some strategies have been provided in (Candès et al. 2018) for selecting diagonal matrix $\text{diag}\{\mathbf{s}\}$.

Given i.i.d. observations $\{(\mathbf{x}_i, Y_i)\}_{i=1}^n$, denote

$$\mathbf{X} = (\mathbf{x}_i)_{i=1}^n \in \mathbb{R}^{n \times p} \text{ and } \mathbf{y} = (Y_i)_{i=1}^n \in \mathbb{R}^n,$$

where each $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})$. The knockoff data matrix $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_i)_{i=1}^n$ is constructed by row w.r.t. $\tilde{\mathbf{x}}|\mathbf{x}$, where $\tilde{\mathbf{x}}_i$ is the knockoff copy of $\mathbf{x}_i$. The $n \times 2p$ matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$

is obtained by connecting $\mathbf{X}$ and $\tilde{\mathbf{X}}$. To identify active features, a paired-input filter is trained on $\big([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}\big)$, e.g, Lasso (Hastie and Tibshirani 1990) in (Candès et al. 2018). Then, each feature (including its knockoff) is assigned with a coefficient-based score, e.g., the absolute value of Lasso coefficient (Candès et al. 2018; Fan et al. 2020a).

Let Z_j and $\tilde{Z}_j$ be the score of X_j and $\tilde{X}_j$ respectively. The importance measure of feature X_j is defined by

$$W_j := w_j\big([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}\big) = Z_j - \tilde{Z}_j,$$

where w_j is a model-driven function associated with $[\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}$. Typical example of W_j is the *Lasso coefficient difference* (LCD) used in (Candès et al. 2018; Fan et al. 2020a). The distribution of W_j enjoys the following *flip-sign* property.

Proposition 1 *(Candès et al. 2018) Assume swapping X_j with $\tilde{X}_j$ has the effect of changing the sign of W_j, i.e.,*

$$w_j\big([\mathbf{X}, \tilde{\mathbf{X}}]_{swap(\{j\})}, \mathbf{y}\big) = -w_j\big([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}\big).$$

Then, each W_j associated with irrelevant feature is independent and symmetrically distributed.

The above property is helpful to obtain the theoretical guarantee on the FDR control.

Lemma 1 *(Candès et al. 2018) If W_j is independent and symmetrically distributed for each $j \in \mathcal{S}_1$. For any given target FDR level $q \in (0, 1)$, let*

$$\tau = \min\left\{\tau > 0 : \frac{1 + |\{j : W_j \le -\tau\}|}{|\{j : W_j \ge \tau\}| \vee 1} \le q\right\}. \quad (4)$$

Then the procedure selecting the variables $\widehat{\mathcal{S}} = \{j : W_j \ge \tau\}$ can control FDR $\le q$.

Usually, the FDR control under model-X knockoff framework depends heavily on the coefficient difference derived from Lasso (Candès et al. 2018; Fan et al. 2020a), group Lasso (Sesia et al. 2020), and paired-input deep neural networks (Lu et al. 2018). In many applications involving complex function relationships, this coefficient-based property may hinder the flexibility and accuracy of model-X knockoff framework.

Stepdown Procedure

Denote $\mathcal{P}_j$ as the p-value associated with the significance of feature $X_j, j = 1, \ldots, p$. Let $\mathcal{P}_{k_j}$, $j = 1, \ldots, p$, be the p-values with $\mathcal{P}_{k_1} \le \cdots \le \mathcal{P}_{k_p}$ and let $\alpha_j, j = 1, \ldots, p$, be the significance threshold values with $\alpha_1 \le \cdots \le \alpha_p$. Naturally, the first m features with lower p-values are selected as informative variables $\widehat{\mathcal{S}} = \{k_1, \ldots, k_m\}$, where

$$m = \max\{M : \mathcal{P}_{k_j} \le \alpha_j, \forall j \le M\}.$$

Lehmann and Romano (2005) have stated the following results about k-FWER and FDP control.

Lemma 2 *(Lehmann and Romano 2005) For any given $\alpha \in (0, 1)$ and $k = 1, \ldots, p$, the stepdown procedure with*

$$\alpha_j = \begin{cases} \dfrac{k\alpha}{p}, & j \le k \\ \dfrac{k\alpha}{p + k - j}, & j > k \end{cases}$$

can control k-FWER $\le \alpha$.

Lemma 3 *(Lehmann and Romano 2005) For any given $\alpha, q \in (0, 1)$, if the p-value of any irrelevant feature is independent of the p-values of informative features, the stepdown procedure with*

$$\alpha_j = \frac{(\lfloor qj \rfloor + 1)\alpha}{p + \lfloor qj \rfloor + 1 - j}$$

satisfies $Prob\{FDP > q\} \le \alpha$.

Error-based Knockoff Inference

This paper exploits the idea of "feature replacing" for controlled feature selection, i.e., replacing a feature with its knockoff and see whether there is a significant difference in the estimation error or not. We first assume the distribution of $\mathbf{x}$ is known as prior and propose an error-based feature statistic for k-FWER, FDP, or FDR control. Then, we extend the theoretical result to a more general setting where the distribution of $\mathbf{x}$ is unknown. Finally, the power analysis is stated for the proposed approach.

Error-based Feature Importance

To construct the error-based feature importance $W_j, j = 1, \ldots, p$, we need to divide n samples into two disjointed parts: $(\mathbf{X}^*, \mathbf{y}^*)$ and $(\mathbf{X}', \mathbf{y}')$, containing n_1 and n_2 samples, respectively.

Let f be the regression estimator trained on $(\mathbf{X}^*, \mathbf{y}^*)$ and let $\tilde{\mathbf{X}}'$ be the knockoff copy of $\mathbf{X}'$. Denote $\big((\mathbf{x}'_i, \tilde{\mathbf{x}}'_i), Y'_i\big)$ as the i-th column of $([\mathbf{X}', \tilde{\mathbf{X}}'], \mathbf{y}')$. Given an estimator f and random sample $(\mathbf{x}, Y) \in \mathcal{X} \times \mathcal{Y}$, define the error-based random variable

$$\xi := \xi(\mathbf{x}, Y) = |f(\mathbf{x}) - Y| \quad (5)$$

and

$$\xi^j := \xi^j(\mathbf{x}, Y) = |f\big(\mathcal{R}_j(\mathbf{x})\big) - Y|, \quad (6)$$

where

$$\mathcal{R}_j(\mathbf{x}) = (X_1, \ldots, X_{j-1}, \tilde{X}_j, X_{j+1}, \ldots, X_p).$$

For observation $(\mathbf{x}'_i, Y'_i)$ associated with $(\mathbf{X}', \mathbf{y}')$, we define

$$\xi_i = |f(\mathbf{x}'_i) - Y'_i| \ \text{ and } \ \xi_i^j = \big|f\big(\mathcal{R}_j(\mathbf{x}'_i)\big) - Y'_i\big|.$$

The feature importance can be measured by the *error difference* between ξ_i^j and ξ_i, i.e.,

$$T_i^j := \xi_i^j - \xi_i.$$

The error-based feature importance can be characterized by

$$W_j := \frac{1}{n_2}\Big(\sum_{i=1}^{n_2} \mathrm{I}_{\{T_i^j > 0\}}\Big) - 0.5, \quad (7)$$

where indicator function $\mathrm{I}_{\{A\}} = 1$ if A is true and 0 otherwise.

The basic properties of W_j are stated as below, which are obtained from the definition of knockoff features (e.g., Definition 2) and the *flip-sign* property of MX-Knockoff feature importance (e.g., Proposition 1). The corresponding proof can be found in *Supplementary Material C*.

Proposition 2 *For each $j \in \mathcal{S}_1$, W_j defined in (7) is independent and symmetrically distributed around zero, and satisfies $n_2(W_j + 0.5) \sim \mathcal{B}(n_2, 0.5), \forall j \in \mathcal{S}_1$.*

Remark 3 *The first conclusion of Theorem 2 differs from Proposition 1 (See also Lemma 3.3 in (Candès et al. 2018)) in that, we remove the assumption on feature importance via replacing strategy. Since the error-based importance measure has no requirement on the structure of learning machine, it may be much flexible for applications. The second conclusion reveals the distribution information of the proposed irrelevant feature's importance, which gives us the opportunity to realize k-FWER control and FDP control via combining the knockoff technique with the stepdown procedure.*

Combining Lemma 1 and Proposition 2 yields the following result for FDR control.

Theorem 1 *For any given target FDR level $q \in (0, 1)$, the error-based knockoff procedure, with feature importance* (7) *and knockoff threshold* (4)*, satisfies FDR $\leq q$.*

Assume the null-hypothesis that the feature is irrelevant. Let $M_j := \max\{n_2(W_j+0.5), n_2(0.5-W_j)\}$. The p-values are defined as

$$\mathcal{P}_j := 2 \sum_{i=M_j}^{n_2} C(n_2, i) \frac{1}{2^{n_2}}, j = 1, \ldots, p \qquad (8)$$

are used to evaluate the feature significance, where $C(n_2, m)$ is the combinatorial number.

The following theoretical results on k-FWER and FDP control can be established by combining Proposition 2 with Lemmas 2 and 3.

Theorem 2 *For any given $\alpha \in (0, 1)$, the stepdown procedure, constructed in Lemma 2 and associated with knockoff-based p-values* (8)*, satisfies k-FWER $\leq \alpha$.*

Theorem 3 *For any given $q, \alpha \in (0, 1)$, the FDP of $\widehat{\mathcal{S}}$, associated with the stepdown procedure in Lemma 3 and p-values in* (8)*, satisfies $Prob\{FDP > q\} \leq \alpha$.*

Remark 4 *Different from the previous knockoff filters relied on the **coefficient difference** (Barber and Candès 2015; Candès et al. 2018; Lu et al. 2018; Barber and Candès 2019), the current knockoff procedure rooted in the **error difference**. The error-based knockoff strategy is model-free (no structure restriction on estimator f), and gives us the opportunity to tackle FDR, FDP, and k-FWER control.*

Robustness Analysis

This section further establishes the asymptotic properties of k-FWER control and FDP control when the distribution of $\mathbf{x}$ is characterized by some unknown Gaussian graphical model, i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. All proofs of this section have been provided in *Supplementary Material C*.

Let $\widehat{\mathbf{\Sigma}}$ be the empirical estimation of covariance matrix obtained by $(\mathbf{X}^*, \mathbf{y}^*)$. To ease the presentation, for any notation $\mathbf{A}$ associated with the unknown covariance matrix $\mathbf{\Sigma}$, the notation $\widehat{\mathbf{A}}$ stand for its empirical estimation constructed via $\widehat{\mathbf{\Sigma}}$. Inspired from Fan et al. (2020a), we introduce the following conditions for our robustness analysis.

Algorithm 1: Construct feature importance statistic W_j

Input: Data $(\mathbf{X}', \mathbf{y}')$, trained filter f, feature index j
Output: Feature importance statistic W_j

Construct $\tilde{\mathbf{X}}'$, i.e., the knockoff copy of $\mathbf{X}'$.
for $i = 1, \ldots, n_2$ **do**
  Obtain $\mathcal{R}_j(\mathbf{x}'_i)$ by replacing j-th feature in $\mathbf{x}'_i$ with its knockoff copy.
  $T_i^j \leftarrow \left|f\big(\mathcal{R}_j(\mathbf{x}'_i)\big) - Y'_i\right| - |f(\mathbf{x}'_i) - Y'_i|$
end for
$W_j \leftarrow \frac{1}{n_2}\left(\sum_{i=1}^{n_2} \mathrm{I}_{\{T_i^j > 0\}}\right) - 0.5.$
return W_j

The following condition on density function is required, which holds true for bounded regression problem with Gaussian noise assumption (Tibshirani 1996; Yuan and Lin 2006; Meier, Van De Geer, and Bü hlmann 2008; Christian 2012).

Condition 1 *Let $\eta(Y|\mathbf{x})$ be the probability density function of Y conditioned on $\mathbf{x}$. There holds $\max_{(\mathbf{x},Y)} \eta(Y|\mathbf{x}) \leq C_1$ for some constant C_1.*

Without loss of generality, assume the covariance matrix $\mathbf{G}$ defined in (3) to be positive definite (Fan et al. 2020a). The following condition is used to characterize the relationship between $\mathbf{G}$ and its empirical estimation $\widehat{\mathbf{G}}$, and to rule out some extreme case of these matrixs, e.g., $\lambda_{max}(\mathbf{G}) = \infty$. Similar condition has been used in (Fan et al. 2020a) for robust analysis.

Condition 2 *Let $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ be the minimum and the maximum matrix eigenvalues, respectively. There exist some positive sequence a_{n_1}, b_{n_1} satisfying $a_{n_1} \to 0, b_{n_1} \to 0$ as $n_1 \to \infty$, and a positive constant C_2 such that*

$$\|\widehat{\mathbf{G}} - \mathbf{G}\|_2 \leq a_{n_1}$$

and

$$\begin{aligned}\frac{1}{C_2} &\leq \min\left\{\lambda_{\min}(\mathbf{G}), \lambda_{\min}(\widehat{\mathbf{G}})\right\} \\ &\leq \max\left\{\lambda_{\max}(\mathbf{G}), \lambda_{\max}(\widehat{\mathbf{G}})\right\} \leq C_2\end{aligned}$$

with probability at least $1 - p^{-\frac{1}{b_{n_1}}}$.

Let $\tilde{\mathbf{x}}^{\widehat{\mathbf{\Sigma}}}$ be the knockoff feature based on the distribution $\mathcal{N}(\mathbf{0}, \widehat{\mathbf{\Sigma}})$. For feasibility, denote $\eta_{\mathbf{\Sigma}}$ and $\eta_{\widehat{\mathbf{\Sigma}}}$ be the distribution density function of $(\mathbf{x}, \tilde{\mathbf{x}})$ and $(\mathbf{x}, \tilde{\mathbf{x}}^{\widehat{\mathbf{\Sigma}}})$ respectivly. The relationship between $\eta_{\mathbf{\Sigma}}$ and $\eta_{\widehat{\mathbf{\Sigma}}}$ is described as below.

Lemma 4 *Under Condition 2, there holds*

$$|\eta_{\mathbf{\Sigma}}(\mathbf{x}, \tilde{\mathbf{x}}) - \eta_{\widehat{\mathbf{\Sigma}}}(\mathbf{x}, \tilde{\mathbf{x}})| \leq O(a_{n_1}), \forall (\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X}^2$$

with probability at least $1 - p^{-\frac{1}{b_{n_1}}}$.

Table 2: Results on the simulated data for controlled feature selection (different dimension p)

p	E-Knockoff(k-FWER)			E-Knockoff(FDP)			E-Knockoff(FDR)			MX-Knockoff			DeepPINK		
	$\text{FDP}_{\max}$	FDR	Power	$\text{FDP}_{\max}$	FDR	Power	$\text{FDP}_{\max}$	FDR	Power	$\text{FDP}_{\max}$	FDR	Power	$\text{FDP}_{\max}$	FDR	Power
50	0.03	0.01	1.00	0.12	0.03	1.00	0.35	0.19	1.00	0.33	0.20	1.00	0.32	0.20	1.00
100	0.03	0.00	0.97	0.17	0.04	1.00	0.40	0.19	1.00	0.45	0.19	1.00	0.52	0.17	1.00
200	0.07	0.00	0.95	0.14	0.04	0.99	0.48	0.20	1.00	0.45	0.20	1.00	0.36	0.16	1.00
400	0.04	0.01	0.91	0.13	0.04	0.98	0.43	0.19	1.00	0.39	0.19	1.00	0.50	0.22	1.00
800	0.07	0.01	0.63	0.13	0.04	0.83	0.36	0.19	0.95	0.40	0.20	1.00	0.33	0.18	1.00
1200	0.06	0.00	0.60	0.08	0.03	0.79	0.42	0.16	0.93	0.40	0.16	1.00	0.38	0.19	1.00
1600	0.07	0.01	0.59	0.17	0.04	0.79	0.40	0.18	0.94	0.35	0.19	1.00	0.44	0.18	1.00
2000	0.07	0.01	0.57	0.13	0.03	0.77	0.52	0.19	0.92	0.41	0.16	1.00	0.44	0.24	0.94

Table 3: Maximum number of false discoveries in 50 trials

method	50	100	200	400	800	1200	1600	2000
E-Knockoff(k-FWER)	1	1	2	1	1	1	1	1

It is a position to present the main results on robustness analysis for controlled variable selection.

Theorem 4 *Let Conditions 1 and 2 be true. For any given $\alpha \in (0,1), k = 1, \ldots, p$ and $n_2 \in \mathbb{R}$, the feature seleciton procedure described in Theorem 2 satisfies*

$$\widehat{k\text{-}FWER} \leq \alpha + O(p^{-\frac{1}{b_{n_1}}}) + O(a_{n_1}).$$

Theorem 5 *Let Conditions 1 and 2 be true. For any given $q, \alpha \in (0,1)$, $n_2 \in \mathbb{R}$, the feature selection procedure described in Theorem 3 satisfies*

$$Prob\{\widehat{FDP} > q\} \leq \alpha + O(p^{-\frac{1}{b_{n_1}}}) + O(a_{n_1}).$$

Remark 5 *Theorems 4 and 5 guarantee the robustness of our error-based knockoff inference under mild conditions. Here, we omit the robust analysis for FDR control since it can be derived directly from (Fan et al. 2020a). To the best of our knowledge, there is no robust analysis for FDP and k-FWER control under the knockoff filtering framework.*

Power Analysis

The following restriction on the predictor f is involved for the power property of our error-based knockoff inference.

Condition 3 *For the error-based random variables ξ in* (5) *and ξ^j in* (6)*, there holds*

$$Prob\{\xi^j > \xi\} > Prob\{\xi^j < \xi\}, \forall j \in \mathcal{S}_0.$$

Condition 3 ensures that the predictor f would have degraded performance when replacing an informative feature with a knockoff feature. Recall that the construction procedure of knockoffs is independent of the response Y, see e.g., Candès et al. (2018); Romano, Sesia, and Candès (2019); Jordon, Yoon, and Mihaela (2019). Therefore, the current restriction on f is mild since less information is more likely to result in additional prediction loss.

The central limit theorem assures that the variance of proposed feature importance in (7) would converge to zero. Thus, we get the following result.

Theorem 6 *Under Condition 3, the feature selection procedures described in Corollaries 1-3 satisfy*

$$\textit{Power} := \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}_0|}{|\mathcal{S}_0|}\right] \to 1 \ \ as \ \ n_2 \to \infty.$$

Remark 6 *Theorem 6 demonstrates that the power of proposed procedures depends on the sample size n_2 of $(\mathbf{X}', \mathbf{y}')$. Theorems 4-6 imply that, for our error-based knockoff inference, there is a tradeoff between n_1 (associated with getting the predictor f and covarince matrix $\widehat{\mathbf{\Sigma}}$) and n_2 (related to generate knockoffs and error-based feature statistic W_j).*

Experimental Analysis

This section states empirical evaluations of our error-based knockoff inference on both synthetic data and HIV dataset (Rhee et al. 2006) to valid our theoretical claims about controlled feature selection and power analysis. The detailed experiment settings and some additional experiments are provided in *Supplementary Material D*.

Simulated Data Evaluation

Inspired by (Lu et al. 2018), we draw $\mathbf{x}$ independently from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}^{-1} = (0.5^{|j-k|})_{1 \leq j,k \leq p}$. Then, we simulate the response from single index model:

$$Y = g(\mathbf{x}\beta) + \epsilon, \ \ \epsilon \sim \mathcal{N}(0, 0.01)$$

where the linkage function

$$g(a) = \sqrt{|a|} + a + a^2 + \sin(a) + \arctan(a), \forall a \in \mathbb{R},$$

$\beta = (\beta_1, \ldots, \beta_p)^T$ satisfying $\beta_j = 0, \forall j \in \mathcal{S}_1$ and $\beta_j = 1/|\mathcal{S}_0|$ otherwise. Here, the sample size $n = 2000$ and the number of features $p \in \{50, 100, 200, 400, 800, 1200, 1600, 2000\}$ with $|\mathcal{S}_0| = 30$ (Lu et al. 2018).

This paper employes the coefficient-based model-X knockoff (Candès et al. 2018) and DeepPink (Lu et al. 2018) as the baselines. We set the target FDR level $q = 0.2$ for all FDR controlled methods, set $q = 0.2$ and $\alpha = 0.2$ for FDP control version of E-Knockoff (*E-Knockoff (FDP)*), and set $k = 2$ and $\alpha = 0.1$ for k-FWER control version of E-Knockoff (*E-Knockoff (k-FWER)*). The feature importance is measured by the coefficient difference associated with

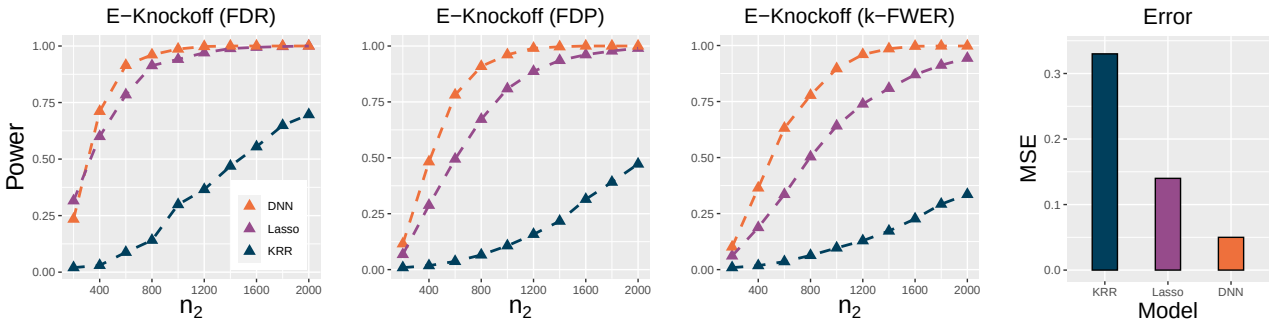

Figure 2: Power analysis (Power Vs. n_2) on the simulated data for different learning models

Lasso for MX-Knockoff (Candès et al. 2018) and associated with paired-input DNNs for DeepPink (Lu et al. 2018). We use Lasso as the base estimator of our E-Knockoff inference with $n_1 = n_2 = 1000$. Table 2 summaries the estimation of FDR, Power and the maximum value of FDP ($\text{FDP}_{\max}$) with 50 repetitions. In addition, Table 3 reports the max number of false discoveries for E-Knockoff (k-FWER) in these trials. These experimental results show that our error-based knockoff inference can reach the FDR control, FDP control, and k-FWER control flexibily, while MX-Knockoff and DeepPINK just can control the FDR. Meanwhile, E-Knockoff (FDP) and E-Knockoff (k-FWER) also enjoy the promising selection accuracy in almost all settings.The results of Table 2 also verify the tradeoff between accuracy and power discussed in (Korn et al. 2004; Lehmann and Romano 2005; Farcomeni 2008).

To verify the model-free property and power ability of our approach, we provide an experiment to illusrate the influence of n_2 and f on selection results. We set $p = 800, n_1 = 1000$ and select n_2 from $\{200, 400, 600, 800, \ldots, 2000\}$. Three classic learning machines are used to get f including Deep neural networks (DNN) (Hinton and Salakhutdinov 2006), Lasso (Tibshirani 1996) and Kernel ridge regression (KRR) (Christian 2012). Experimental results of power and mean square error (MSE) are displaced in Figure 2 after repeating the each experiment 30 times. Full simulated results are presented in *Supplementary Material D*. It can be observed that a powerful selection result can be made with the increase of n_2, which supports our conclusion in Theorem 6. Also, the result implies that a better-trained filter can select true active features with less samples.

Real Data Evaluation

We next apply E-Knockoff to identify key mutations of HIV associated with the drug resistance (Rhee et al. 2006). The HIV-1 dataset consists of the data of the drug resistance level, mutations, and the treatment-selected mutations (TSM) associated with drug resistance. For each drug, the response Y is the log-transformed drug resistance level, and the j-th feature of argument $\mathbf{x}$ indicates the presence or absence of the j-th mutation (Lu et al. 2018; Li et al. 2021). Figure 3 summarizes the experimental results related to FPV drugs resistance (with 1809 samples and 224 dimensions), and *Supplementary Material D* reports the results of other drugs.

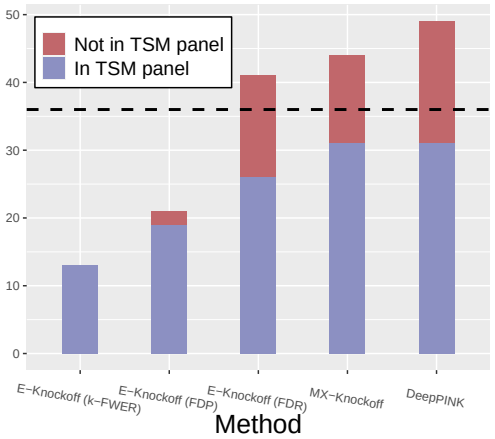

Figure 3: Results on the HIV-1 drug resistance dataset. For FPV drug class, we show the number of mutation positions for PI identified by different knockoff filters. The color indicates whether or not the selected position appears in the TSM panel, and the horizontal line shows the total number of positions on the TSM panel.

Here, we use Lasso as the base estimator for E-Knockoff inference ($n_1 = \frac{n}{3}, n_2 = \frac{2n}{3}$). We set $k = 2, \alpha = 0.1$ for E-Knockoff (k-FWER), $q = 0.2, \alpha = 0.2$ for E-Knockoff (FDP), and $q = 0.2$ for E-Knockoff (FDR), MX-Knockoff, and DeepPINK. Empirical results demonstrate that the error-based knockoff inference can usually control the false discovery efficiently.

Conclusion

To improve the adaptivity and flexibility of the model-X knockoff framework, this paper proposes a new error-based knockoff inference method for controlled feature selection. We establish the statistical asymptotic analysis and power analysis of the proposed approach. Empirical evaluations demonstrate the competitive performance of the proposed procedure on simulated and real data, which support our research motivation and theoretical findings. In the future, it is interesting to extend the current work for multi-environment controlled feature selection (Li et al. 2021).

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant Nos. 12071166, 62076041, 61702057, 61806027, 61972188, 62106191 and by the Fundamental Research Funds for the Central Universities of China under Grant 2662020LXQD002. We are grateful to the anonymous AAAI reviewers for their constructive comments.

References

Bach, F. 2008. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(40): 1179 – 1225.

Bai, X.; Ren, J.; Fan, Y.; and Sun, F. 2020. KIMI: Knockoff inference for motif identification from molecular sequences with controlled false discovery rate. *Bioinformatics*, 37(6): 759 – 766.

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

Barber, R. F.; and Candès, E. J. 2019. A knockoff filter for high-dimensional selective inference. *Annals of Statistics*, 47(5): 2504 – 2537.

Barber, R. F.; Candès, E. J.; and Samworth, R. J. 2020. Robust inference with knockoffs. *Annals of Statistics*, 48(3): 1409 – 1431.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289 – 300.

Benjamini, Y.; and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4): 1165 – 1188.

Candès, E. J.; 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.

Chen, H.; Wang, Y.; Zheng, F.; Deng, C.; and Huang, H. 2021. Sparse modal additive model. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6): 2373 – 2387.

Christian, R. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Delattre, S.; and Roquain, E. 2015. New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *The Annals of Statistics*, 43(3): 1141 – 1177.

Efron, B.; and Tibshirani, R. 2002. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1): 70 – 86.

Fan, J.; Guo, S.; and Hao, N. 2012. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1): 37 – 65.

Fan, J.; and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348 – 1360.

Fan, J.; and Lv, J. 2010. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1): 101 – 148.

Fan, Y.; Demirkaya, E.; Li, G.; and Lv, J. 2020a. RANK: Large-Scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 115(529): 362 – 379.

Fan, Y.; Demirkaya, E.; and Lv, J. 2019. Nonuniformity of p-values can occur early in diverging dimensions. *Journal of Machine Learning Research*, 20(77): 1 – 33.

Fan, Y.; Lv, J.; Sharifvaghefi, M.; and Uematsu, Y. 2020b. IPAD: Stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association*, 115(532): 1822 – 1834.

Farcomeni, A. 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17(4): 347 – 388.

Frecon, J.; Salzo, S.; and Pontil, M. 2018. Bilevel learning of the group Lasso structure. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Friedman, J. H.; Hastie, T.; and Tibshirani, R. 2010. A note on the group lasso and a sparse group lasso. arXiv:1001.0736.

Genovese, C.; and Wasserman, L. 2004. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3): 1035 – 1061.

Hastie, T. J.; and Tibshirani, R. 1990. *Generalized Additive Models*. London: Chapman and Hall.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786): 504 – 507.

Hochberg, Y.; and Tamhane, A. C. 1987. *Multiple Comparison Procedures*. Wiley.

Jordon, J.; Yoon, J.; and Mihaela, v. d. S. 2019. KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.

Korn, E. L.; Troendle, J. F.; McShane, L. M.; and Simon, R. 2004. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2): 379 – 398.

Lehmann, E. L.; and Romano, J. P. 2005. Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3): 1138 – 1154.

Lemhadri, I.; Ruan, F.; and Tibshirani, R. 2021. LassoNet: Neural networks with feature sparsity. *Journal of Machine Learning Research*, 130: 10 – 18.

Li, S.; Sesia, M.; Romano, Y.; Candès, E. J.; and Sabatti, C. 2021. Searching for consistent associations with a multi-environment knockoff filter. arXiv:2106.04118.

Lin, Y.; and Zhang, H. H. 2007. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5): 2272 – 2297.

Liu, G.; Chen, H.; and Huang, H. 2020. Sparse shrunk additive models. In *International Conference on Machine Learning (ICML)*.

Liu, H.; Wasserman, L.; Lafferty, J.; and Ravikumar, P. 2008. SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*.

Liu, W.; Ke, Y.; Liu, J.; and Li, R. 2020. Model-free feature screening and FDR control with knockoff features. *Journal of the American Statistical Association*, 1 – 16.

Liu, W.; and Shao, Q. 2014. Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics*, 42(5): 2003 – 2025.

Lu, Y.; Fan, Y.; Lv, J.; and Stafford Noble, W. 2018. DeepPINK: reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Pan, C.; and Zhu, M. 2017. Group additive structure identification for kernel nonparametric regression. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rhee, S.; Taylor, J.; Wadhera, G.; Ben-Hur, A.; Brutlag, D. L.; and Shafer, R. W. 2006. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46): 17355 – 17360.

Romano, Y.; Sesia, M.; and Candès, E. J. 2019. Deep knockoffs. *Journal of the American Statistical Association*, 115(532): 1861–1872.

Sesia, M.; Katsevich, E.; Bates, S.; Candès, E. J.; and Sabatti, C. 2020. Multi-resolution localization of causal variants across the genome. *Nature Communication*, 11(1093).

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

Wang, Y.; Chen, H.; Zheng, F.; Xu, C.; Gong, T.; and Chen, Y. 2020. Multi-task additive models for robust estimation and automatic structure discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Weinstein, A.; Su, W. J.; Bogdan, M.; Barber, R. F.; and Candès, E. J. 2020. A Power Analysis for Knockoffs with the Lasso Coefficient-Difference Statistic. arXiv:2007.15346.

Yin, J.; Chen, X.; and Xing, E. P. 2012. Group sparse additive models. In *International Conference on Machine Learning (ICML)*.

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.