Knockoffs Inference for Partially Linear Models with Automatic Structure Discovery

Hao Wang[1], **Biqin Song**[1,2,*], **Hao Deng**[1,2], **Hong Chen**[1, 2]

[1]College of Informatics, Huazhong Agricultural University, Wuhan 430062, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China
hao_wang@webmail.hzau.edu.cn, biqin.song@mail.hzau.edu.cn, dengh@mail.hzau.edu.cn, chenh@mail.hzau.edu.cn

Abstract

Partially linear models (PLM) have attracted much attention in the field of statistical machine learning. Specially, the ability of variable selection of PLM has been studied extensively due to the high requirement of model interpretability. However, few of the existing works concerns the false discovery rate (FDR) controllability of variable selection associated with PLM. To address this issue, we formulate a new Knockoffs Inference scheme for Linear And Nonlinear Discoverer (called KI-LAND), where FDR is controlled with respect to both linear and nonlinear variables for automatic structure discovery. For the proposed KI-LAND, theoretical guarantees are established for both FDR controllability and power, and experimental evaluations are provided to validate its effectiveness.

Introduction

Linear and non-parametric models are two important classes statistical modeling tools, each with their own unique advantages. The linear model is simple and has fine interpretability, but it is limited by the linearity assumption. In contrast, the non-parametric model is much more flexible and can adapt to a wide range of functional forms present in the data, but it tends to be less interpretable. Therefore, a fundamental accuracy-interpretability tradeoff is need to be taken into account. Semi-parametric model inherits the interpretation of the linear model and flexibility of non-parametric model by allowing covariates to be linear or nonlinear, which can model complex data in some scientific fields, such as econometrics, social sciences, information sciences, biomedicine, and so on (Ruppert, Wand, and Carroll 2003; Härdle et al. 2004; Fuller 2009). Partially linear model (PLM) is the classic example of Semi-parametric model (Engle et al. 1986; Härdle, Liang, and Gao 2000). For the response variable $y \in \mathcal{R}$ and input variables $(\mathbf{x}, \mathbf{t}) = (x_1, \ldots, x_p, t_1, \ldots, t_q) \in \mathcal{R}^{p+q}$, PLM is given by:

$$y = b + \mathbf{x}^\top \boldsymbol{\beta} + f(\mathbf{t}) + \varepsilon, \tag{1}$$

where b and $\boldsymbol{\beta}$ are the intercept and the vector of coefficients for linear terms, respectively. f is a nonlinear function from $\mathcal{R}^q$ to $\mathcal{R}$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

*Corresponding author.

Variable selection for PLM has drawn extensive attention over the past two decades. It has usually been studied by using a specially designed penalty function under different smooth regression conditions, such as the component selection and smoothing operator penalty (Lin and Zhang 2006), the smoothly clipped absolute deviation penalty (Xie and Huang 2009), and the doubly penalized procedure (Wang et al. 2014a). Additional methods of variable selection for PLM can be found in the works of Wang et al. (2014b); Su and Candès (2016); Lian, Zhao, and Lv (2019); Lv and Lian (2022). While these methods primarily focus on variable selection, they often lack control of error selections, such as the false discovery rate (FDR).

One very recent approach for FDR control is knockoffs which was first proposed by Barber and Candès (2015). The key idea of knockoff is to construct fake variables which have a similar structure to the original ones but have no effect on the response, then computing an importance score for each variable and selecting those with higher scores than their fake copies. This work achieved effective FDR control in the context of Gaussian linear model, where the dimensionality d does not exceed the sample size n. The knockoff filter was subsequently extended to high-dimensional linear models with the help of data splitting and feature screening techniques (Barber and Candès 2016). To achieve FDR control in general high dimensional nonlinear model, model-X knockoff (Candès et al. 2018) was proposed. It can accommodate arbitrary dependence structure of the response variable y on input variables $\mathbf{x}$ and bypass the need of calculating accurate p-values. Fan et al. (2020) expanded model-X knockoff to nonparametric additive model. Building on this work, Xiaowu Dai and Li (2023) proposed a kernel knockoffs selection method. This allowed the knockoff approach to be applied in more flexible, non-linear settings, beyond the original linear model assumptions. Su et al. (2023) first applied the generalized knockoffs framework for variable selection in PLM, but it was a linear model by nature. As far as we know, few existing works studies about knockoffs inference for PLM with nonlinear settings.

The works described above have all accomplished FDR control, but a few studies have conducted comprehensive theoretical analysis for power as in (Fan et al. 2020; Weinstein et al. 2022). However, their power analysis was limited to the linear model setting, and a more general theoretical

understanding of power across different models remains an important open challenge.

To fill the aforementioned gaps, a new variable selection procedure called *Knockoffs Inference for Linear And Nonlinear Discoverer* (KI-LAND) is proposed to study FDR control and power analysis for model (1). This is an attempt to extend knockoff framework to PLM, because there is no need for limit on p and q in (1). The main contributions of this paper are summarized as below:

i) *PLM-based knockoff inference with automatic structure discovery*. Knockoff framework is expanded to PLM, where FDR is controlled with respect to both linear and nonlinear variables for automatic structure discovery.

ii) *Theoretical guarantees and empirical effectiveness*. Theoretical analyses for FDR and power are established. Meanwhile, the experiments validate the effectiveness of the proposed method.

Methodology

Problem Setup

This paper assume that all covariates are scaled to $[0, 1]$ without loss of generality. Consider the i.i.d (independent and identically distributed) observation $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where y_i is the response, $\mathbf{x}_i = (x_{i1}, ..., x_{id})^\top$ is the input, and the true regression model is

$$y_i = b_0 + \sum_{j\in\mathcal{I}_L} x_{ij}\beta_j + \sum_{j\in\mathcal{I}_N} f_j(x_{ij}) + \sum_{j\in\mathcal{I}_O} 0(x_{ij}) + \varepsilon_i. \quad (2)$$

Here b_0 is an intercept, β_j is the coefficient for linear term, f_j is an unknown unary nonlinear function, $0(x_{ij})$ is null function, and ε_i is noise. $\mathcal{I}_L, \mathcal{I}_N, \mathcal{I}_O$ denote the index sets of nonzero linear effects, nonzero nonlinear effects, and null effects, respectively. We denote the total set as $\mathcal{I} = \{1, \ldots, d\} = \mathcal{I}_L \cup \mathcal{I}_N \cup \mathcal{I}_O$ and assume the three subgroups do not intersect each other. In applications, $\mathcal{I}_L, \mathcal{I}_N, \mathcal{I}_O$ are generally unknown.

The model (2) can also be considered as a special additive model

$$y_i = b_0 + g_1(x_{i1}) + \cdots + g_d(x_{id}) + \varepsilon_i, \quad (3)$$

where each $g_j, 1 \le j \le d$ is called the component function. Usually, we assume that each component function satisfies some smoothness conditions and $\mathbb{E}[g_j(x_{ij})] = 0, j = 1, \ldots, d$.

Specially, we make assumption that $g_j \in \mathcal{H}_j$, the second-order Sobolev space on $\mathcal{X}_i = [0, 1]$, that means, $\mathcal{H}_j = \{g : g$ and g' are absolutely continuous, $g'' \in L^2[0, 1]\}$. In functional analysis, $\mathcal{H}_j$ is a reproducing kernel Hilbert space (RKHS), when having the following norm:

$$\|g_j\|^2_{\mathcal{H}_j} = \{\int_0^1 g_j(x)dx\}^2 + \{\int_0^1 g_j'(x)dx\}^2 + \int_0^1 \{g_j''(x)\}^2 dx.$$

The reproducing kernel (RK) in $\mathcal{H}_j$ is $R(x, x') = R_0(x, x') + R_1(x, x')$ with $R_0(x, x') = k_1(x)k_2(x')$ and $R_1(x, x') = k_2(x)k_2(x') - k_4(x - x')$, where $k_1(x) = x - \frac{1}{2}, k_2(x) = \frac{1}{2}\{k_1^2(x) - \frac{1}{12}\}$, and $k_4(x) = \frac{1}{24}\{k_1^4(x) - \frac{1}{2}k_1^2(x) + \frac{7}{240}\}$. More details in (Wahba 1983) and (Chong 2013). By the analysis in these two references, the space $\mathcal{H}_j$ has the following orthogonal decomposition:

$$\mathcal{H}_j = \{1\} \oplus \mathcal{H}_{0j} \oplus \mathcal{H}_{1j}, \quad (4)$$

where$\{1\}$ is the mean space, $\mathcal{H}_{0j} = \{g_j : g_j''(x) \equiv 0\}$ is the linear contrast subspace, and $\mathcal{H}_{1j} = \{g_j : \int_0^1 g_j(x)\, dx = 0, \int_0^1 g_j'(x)\, dx = 0, g_j'' \in L^2[0, 1]\}$ is the nonlinear contrast space. Both $\mathcal{H}_{0j}$ and $\mathcal{H}_{1j}$ as subspace of $\mathcal{H}_j$ are also RKHS, and have the reproducing kernels R_0 and R_1 respectively. Then, by the decomposition (4), for any function $g_j \in \mathcal{H}_j$,we can decompose it to linear and nonlinear components

$$g_j(x_{ij}) = b_j + \beta_j(x_{ij} - \frac{1}{2}) + g_{1j}(x_{ij}), \quad (5)$$

$$g(\mathbf{x}_i) = b_0 + g_1(x_{i1}) + \cdots + g_d(x_{id}), \quad (6)$$

where b_j is the intercept for component function g_j, the term $\beta_j(x_{ij} - \frac{1}{2}) = \beta_j k_1(x_{ij}) \in \mathcal{H}_{0j}$ is the linear component and $g_{1j}(x_{ij}) \in \mathcal{H}_{1j}$ is the nonlinear component. $g(\mathbf{x}_i)$ is the function estimated in space $\mathcal{H}$. Further more we have the decomposition of $\mathcal{H}$:

$$\begin{aligned}\mathcal{H} &= \bigoplus_{j=1}^d \mathcal{H}_j = \{1\} \oplus \bigoplus_{j=1}^d \mathcal{H}_{0j} \oplus \bigoplus_{j=1}^d \mathcal{H}_{1j} \\ &= \{1\} \oplus \mathcal{H}_0 \oplus \mathcal{H}_1,\end{aligned}$$

where $\mathcal{H}_0 = \bigoplus_{j=1}^d \mathcal{H}_{0j}$ and $\mathcal{H}_j = \bigoplus_{j=1}^d \mathcal{H}_{1j}$. By the above decomposition, (6) can be rewritten as the following form:

$$g(\mathbf{x}_i) = b + \sum_{j=1}^d \beta_j k_1(x_{ij}) + \sum_{j=1}^d g_{1j}(x_{ij}), \quad (7)$$

where $b = b_0 + \sum_{j=1}^d b_j$.

By the above analysis, we are able to distinguish between linear, nonlinear, and null variables by the following criteria:

Linear index set:$\mathcal{I}_L = \{j : \beta_j \neq 0, g_{1j} \equiv 0\}$,
Nonlinear index set:$\mathcal{I}_N = \{j : g_{1j} \not\equiv 0\}$,
Null index set:$\mathcal{I}_O = \{j : \beta_j = 0, g_{1j} \equiv 0\}$.

Knockoff Variable Construction

Now we will introduce how to construct a knockoff random variable. In (Candès et al. 2018), the model-X knockoff random variable $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_d) \in \mathcal{R}^d$ of a random variable $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{R}^d$ is constructed with the following two properties (exchangeability and independence):

$$i) \quad (\mathbf{x}, \tilde{\mathbf{x}}) \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}\,(\mathcal{S})}, \text{ for any } \mathcal{S} \subseteq \{1, \ldots, d\},$$
$$ii) \quad y \perp\!\!\!\perp \tilde{\mathbf{x}} \mid \mathbf{x}.$$

Above, $\overset{d}{=}$ denotes identically distributed, the vector $(\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}\,(S)}$ is obtained from $(\mathbf{x}, \tilde{\mathbf{x}})$ by swapping the entries $\mathbf{x}_j$ and $\tilde{\mathbf{x}}_j$ for each $j \in S$; for instance, with $p = 3$ and $S = \{2, 3\}$,

$$(x_1, x_2, x_3, \tilde{x}_1, \tilde{x}_2, \tilde{x}_3)_{\text{swap}\,(\{2,3\})} \overset{d}{=} (x_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_1, x_2, x_3).$$

There are two classic data-based methods to generate knockoff variables. The first method is the model-X knockoffs, which constructs knockoffs by using the expectation and variance, like the follows:

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\tilde{\mathbf{x}}],$$
$$\operatorname{cov}[(\mathbf{x}, \tilde{\mathbf{x}})] = \begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{\Psi} - \operatorname{diag}(\boldsymbol{s}) \\ \boldsymbol{\Psi} - \operatorname{diag}(\boldsymbol{s}) & \boldsymbol{\Psi} \end{bmatrix},$$

where $\boldsymbol{\Psi}$ is covariance matrix of $\mathbf{X}$, $\operatorname{diag}(\boldsymbol{s})$ is a diagonal matrix with all components of s being positive and such that $\operatorname{cov}[(\boldsymbol{X}, \tilde{\boldsymbol{X}})]$ is positive definite. To obtain a good selection power, $\operatorname{diag}(\boldsymbol{s})$ should be constructed as large as possible (Candès et al. 2018). A variant of model-X knockoffs in (Barber, Candès, and Samworth 2020) is generated by approximating the distribution of $\mathbf{X}$.

The second method is deep knockoffs (Yaniv Romano and Candès 2020), a sampling machine based on deep generative models, which approximates model-X knockoffs for unspecified data distributions. The key idea is to optimize the criterion evaluating the validity of the generated knockoffs by refining the knockoff sampling mechanism iteratively. This method leverages higher-order moments, enabling a better approximation of exchangeability.

In this article, we construct knockoffs with model-X knockoff when the input data approximately follows a multivariate normal distribution, and with deep knockoffs otherwise. In the next subsection, we will illustrate how to use knockoff procedure to implement future selection.

Model Setup

Now we state the following regularization scheme originally proposed in (Zhang and Liu 2011):

$$\begin{aligned} &\min_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left[y_i - g\left(\mathbf{x}_i, \tilde{\mathbf{x}}_i\right)\right]^2 \\ &+ \lambda_1 \sum_{j=1}^{2d} w_{0j} \left\|\mathcal{P}_{0j} g\right\|_{\mathcal{H}_0} + \lambda_2 \sum_{j=1}^{2d} w_{1j} \left\|\mathcal{P}_{1j} g\right\|_{\mathcal{H}_1}, \end{aligned} \tag{8}$$

where

$$\begin{aligned} g\left(\mathbf{x}_i, \tilde{\mathbf{x}}_i\right) = b + &g_1(x_{i1}) + \cdots + g_d(x_{id}) + \\ &g_{d+1}(\tilde{x}_{i1}) + \cdots + g_{2d}(\tilde{x}_{id}), \end{aligned}$$

$\mathbf{x}_i, \tilde{\mathbf{x}}_i \in \mathcal{R}^d$ are original and knockoff variables, and $\mathcal{P}_{0j} g, \mathcal{P}_{1j} g$ project $\mathcal{H}$ into $\mathcal{H}_{0j}$ and $\mathcal{H}_{1j}$, respectively. For the linear component, we set $\left\|\mathcal{P}_{0j} g\right\|_{\mathcal{H}_0} = |\beta_j|$, which is equivalent to the lasso penalty (Tibshirani 1996). In the nonlinear counterpart, we impose $\left\|\mathcal{P}_{1j} g\right\|_{\mathcal{H}_1} = \left\{\int_0^1 \left[g_{1j}''(x)\right]^2 dx\right\}^{1/2}$ to penalize the nonlinear component of g_j. As shown in (Zhang and Liu 2011),

$$w_{0j} = \frac{1}{|\tilde{\beta}_j|^{\alpha}}, w_{1j} = \frac{1}{\|\tilde{g}_{1j}\|_2^{\gamma}}, j = 1, \ldots, 2d,$$

are weight parameters used in optimization computation, where $\tilde{\beta}_j, \tilde{g}_{1j}$ are the decomposition of $\tilde{g}$ by (5), $||\cdot||_2$ is L_2 norm, $\alpha \geq 3/2$ and $\gamma \geq 3/2$ are hyperparameters to tune.

This weight design ensures that components deemed less significant are penalized more heavily, thereby shrinking them towards zero. Conversely, components that are crucial to the function are assigned lighter penalties, which helps in preserving their non-zero values during the selection process. This nuanced weighting strategy aims to balance the preservation of essential variables with the suppression of less relevant elements. The similar strategies have been used in linear models (Zhang and Lu 2007; Zou 2006; Wang, Li, and Jiang 2007) and SS-ANOVA model (Storlie et al. 2011).

To obtain the weights, we first need to get an initial estimation of the function g, which will then allow us to determine the values of $\tilde{\beta}_j$ and $\tilde{g}_{1j}$. For this initial estimation, we choose SS-ANOVA model (9):

$$\min_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left[y_i - g\left(\mathbf{x}_i, \tilde{\mathbf{x}}_i\right)\right]^2 + \lambda \sum_{j=1}^{2d} \left\|\mathcal{P}_{1j} g\right\|_{\mathcal{H}_1}^2. \tag{9}$$

Now we could consider to solve (8). As referenced in (Zhang and Liu 2011), we don't solve (8) directly, but solve an equivalent and more convenient problem :

$$\begin{aligned} \min_{\theta\geq\mathbf{0}, g\in\mathcal{H}} &\frac{1}{n}\sum_{i=1}^{n}\left[y_i - g\left(\mathbf{x}_i, \tilde{\mathbf{x}}_i\right)\right]^2 + \lambda_1 \sum_{j=1}^{2d} w_{0j} \left\|\mathcal{P}_{0j} g\right\| \\ &+ \tau_0 \sum_{j=1}^{2d} \theta_j^{-1} w_{1j} \left\|\mathcal{P}_{1j} g\right\|_{\mathcal{H}_1}^2 + \tau_1 \sum_{j=1}^{2d} w_{1j}\theta_j, \\ &\text{subject to } \theta_j \geq 0, j = 1, \ldots, 2d, \end{aligned} \tag{10}$$

where τ_0 is a fixed constant, (λ_1, τ_1) are hyperparameters to tune. The following lemma states that optimization problems (8) and (10) are equivalent (Zhang and Liu 2011).

Lemma 1. *Set $\tau_1 = \lambda_2^2/(4\tau_0)$. (i) If $\hat{g}$ minimizes (8), set $\hat{\theta}_j = \tau_0^{1/2}\tau_1^{-1/2}\left\|\mathcal{P}_{1j}\hat{g}\right\|$; then the pair $(\hat{\boldsymbol{\theta}}, \hat{g})$ minimizes (10). (ii) If $(\hat{\boldsymbol{\theta}}, \hat{g})$ minimizes (10), then $\hat{g}$ minimizes (8).*

Supposing that the input $\{\mathbf{x}_i\}_{i=1}^n$ is given, we can use the representation theory to get the solution of (10):

$$\begin{aligned} \hat{g}(\mathbf{x}, \tilde{\mathbf{x}}) = &\hat{b} + \sum_{j=1}^{d} \hat{\beta}_j k_1\left(x_j\right) + \sum_{j=d+1}^{2d} \hat{\beta}_j k_1\left(\tilde{x}_j\right) \\ &+ \sum_{j=1}^{d} \hat{\theta}_j w_{1j}^{-1} \sum_{i=1}^{n} \hat{c}_i R_{1j}\left(x_{ij}, x_j\right) \\ &+ \sum_{j=d+1}^{2d} \hat{\theta}_j w_{1j}^{-1} \sum_{i=1}^{n} \hat{c}_i \tilde{R}_{1j}\left(\tilde{x}_{ij}, \tilde{x}_j\right). \end{aligned} \tag{11}$$

The expression (11) indicates that linearity or nonlinearity of g_j is determined by the case where two parameters $\hat{\beta}_j$ and $\hat{\theta}_j$ are 0 or not. Therefore, two parameters are used to select linear and nonlinear components in (7). It is clearly that $\hat{g}_{1j} = 0$ when $\hat{\theta}_j = 0$, we define following index sets:

$$\begin{aligned} \hat{\mathcal{I}}_L &= \left\{j = 1, \ldots, 2d : \hat{\beta}_j \neq 0, \hat{\theta}_j = 0\right\}, \\ \hat{\mathcal{I}}_N &= \left\{j = 1, \ldots, 2d : \hat{\theta}_j \neq 0\right\}, \\ \hat{\mathcal{I}}_O &= \left\{j = 1, \ldots, 2d : \hat{\beta}_j = 0, \hat{\theta}_j = 0\right\}. \end{aligned} \tag{12}$$

In the next subsection, we will show how to construct the important score and knockoff statistic which help us to select linear and nonlinear components and control the false discovery rate (FDR).

KI-LAND

In this article, we employ a subsampling strategy analogous to the one described in (Meinshausen and Bühlmann 2010; Dümbgen, Samworth, and Schuhmacher 2013). Let $I \subset \{1, \ldots, n\}$ denote the subsample indices with size $\lfloor n/2 \rfloor$, where $\lfloor \cdot \rfloor$ means round down. we make set I as training samples on model (9) and model (10). Then, we will get the following index sets:

$$\begin{aligned} \hat{\mathcal{S}}(I)^L &= \left\{ j = 1, \ldots, 2d : \hat{\beta}_j \neq 0 \right\}, \\ \hat{\mathcal{S}}(I)^N &= \left\{ j = 1, \ldots, 2d : \hat{\theta}_j \neq 0 \right\}. \end{aligned} \tag{13}$$

Where $\hat{\mathcal{S}}(I)^L$ and $\hat{\mathcal{S}}(I)^N$ are the index sets of the linear and nonlinear components based on training data I, respectively. We repeat subsampling subset I without replacement U times. The probabilities of the linear component and the nonlinear component being selected can be represented as:

$$\Pi_j^L = \mathcal{P}(j \in \hat{\mathcal{S}}(I)^L), \Pi_j^N = \mathcal{P}(j \in \hat{\mathcal{S}}(I)^N), j = 1, \ldots, 2d. \tag{14}$$

Where the symbol $\mathcal{P}$ denotes the probability. The selection probabilities Π_j^L and Π_j^N are defined as important scores that reflect the importance of linear and nonlinear components of the original and knockoff variables, respectively. Specifically, the larger the probability value, the more important the corresponding component is.

Due to the selection probabilities are unknown, they can be estimated accurately by the empirical selection probabilities. More specifically, repeating the above subsampling and estimating model (8) and (10) U times, we denote I_u as the uth time subsampling training set. Then, Π_j^L and Π_j^N in (14) can be transform into the follows:

$$\begin{aligned} \widehat{\Pi}_j^L &= \frac{1}{U} \sum_{u=1}^{U} \mathbf{1} \left(j \in \widehat{\mathcal{S}}(I_u)^L \right), \quad j = 1, \ldots, 2d, \\ \widehat{\Pi}_j^N &= \frac{1}{U} \sum_{u=1}^{U} \mathbf{1} \left(j \in \widehat{\mathcal{S}}(I_u)^N \right), \quad j = 1, \ldots, 2d. \end{aligned} \tag{15}$$

Based on the selection probabilities, we can define the knockoff statistics for linear and nonlinear components of the original variable $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{R}^d$.

$$\Delta_j^L = \Pi_j^L - \Pi_{j+d}^L, \Delta_j^N = \Pi_j^N - \Pi_{j+d}^N, j = 1, \ldots, d. \tag{16}$$

The higher the value of knockoff statistic Δ_j^L or Δ_j^N, the more important the jth linear component or nonlinear component of original variables is.

The most important step is to control FDR of the selected linear and nonlinear components. Given the target nominal FDR q, we need to choose data-dependent threshold values T^L and T^N for linear and nonlinear components selection respectively. There are two ways to construct threshold, one is the classic knockoff filter (Barber and Candès 2015; Xiaowu Dai and Li 2023) to choose thresholds as the follows:

$$T = \min \left\{ t \in \{|\Delta_j| : |\Delta_j| > 0\} : \frac{\#\{j : \Delta_j \leqslant -t\}}{\#\{j : \Delta_j \geqslant t\}} \leqslant q \right\}. \tag{17}$$

Δ_j is the knockoff statistic and we set $T = \infty$ if the above set is null set. Another one is a more conservative knockoff filter but being used commonly as follows:

$$T_+ = \min \left\{ t \in \{|\Delta_j| : |\Delta_j| > 0\} : \frac{\#\{j : \Delta_j \leqslant -t\} + 1}{\#\{j : \Delta_j \geqslant t\}} \leqslant q \right\}. \tag{18}$$

After threshold values T^L and T^N are constructed via (17) or (18), the final sets of selected linear and nonlinear components are $\hat{\mathcal{S}}^L$ and $\hat{\mathcal{S}}^N$ respectively:

$$\begin{aligned} \hat{\mathcal{S}}^L &= \left\{ j \in \{1, \ldots, p\} : \Delta_j^L \geqslant T^L \right\}, \\ \hat{\mathcal{S}}^N &= \left\{ j \in \{1, \ldots, p\} : \Delta_j^N \geqslant T^N \right\}. \end{aligned} \tag{19}$$

Now we can choose linear variables and nonlinear variables based on (19). The index set (12) can be redefined as follow:

$$\begin{aligned} \hat{\mathcal{I}}_L &= \left\{ j = 1, \ldots, d : j \in \hat{\mathcal{S}}^L, j \notin \hat{\mathcal{S}}^N \right\}, \\ \hat{\mathcal{I}}_N &= \left\{ j = 1, \ldots, d : j \in \hat{\mathcal{S}}^N \right\}, \\ \hat{\mathcal{I}}_O &= \left\{ j = 1, \ldots, d : j \notin \hat{\mathcal{S}}^L, j \notin \hat{\mathcal{S}}^N \right\}. \end{aligned} \tag{20}$$

Algorithm

The knockoff procedure described before can be summarized in 10 steps, please see Algorithm 1. Step 1 is to generate the knockoff variables via model-X knockoff or deep knockoffs. Steps 2 to 5 are carried out U times repeatedly over a number of subsampling replications. By doing this, it not only can calculate the importance scores, but can also reduce the impact of randomness. Steps 6 to 9 are for variables selection and the final step is output.

Theoretical Analysis

In this section, we build theoretical guarantees for the KI-LAND procedure in terms of controlling FDR and power asymptomatic property. All proofs can be seen in supplementary material.

FDR Analysis

In this subsection, theoretical analysis shows that KI-LAND procedure is able to control FDR at any given nominal level q and sample size. In subsection Problem Setup, we employ kernel function method to learn the function g_j in (3) of each dimension's variable, and build a mapping between the RKHS and the original function space. By construction, the kernel matrix of the knockoff variables is designed to mimic the structure of the kernel matrix of the original variables, despite the knockoffs being unassociated with the response when conditioned on the original variables (Xiaowu Dai and Li 2023). For the stability of the results, we use a subsampling technique. Finally, KI-LAND controls the

Algorithm 1: KI-LAND procedure

Input: training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the number of subsampling replications U and the nominal FDR level $q \in (0,1)$.

Output: $\hat{\mathcal{I}}_L, \hat{\mathcal{I}}_N, \hat{\mathcal{I}}_O$.

1: **Step 1:** construct the knockoff $\{\tilde{\mathbf{x}}_\mathbf{i}\}_{i=1}^n$ via model-X knockoff or deep knockoffs.
2: **for** $u = 1$ to U **do**
3: **Step 2:** sampling without replacement to obtain a training set $I_u \subset \{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$.
4: **Step 3:** estimate the weights w_{0j} and w_{1j} by the SS-ANOVA model (9) .
5: **Step 4:** solve the equivalence problem (10) to get $\hat{\beta}_j$ and $\hat{\theta}_j$.
6: **Step 5:** record the selected index sets (13) of the linear and nonlinear components.
7: **end for**
8: **Step 6:** compute the important scores by (15), i.e., the empirical selection frequency based on U times estimation of model (9) and (10).
9: **Step 7:** compute knockoff statistics based on the important scores Δ_j^L and Δ_j^N via (16).
10: **Step 8:** select data-dependent thresholds T^L and T^N by (17) or (18).
11: **Step 9:** choose the final sets of linear and nonlinear components via (19).
12: **Step 10:** output linear variable set $\hat{\mathcal{I}}_L$, nonlinear variable set $\hat{\mathcal{I}}_N$ and null variable set $\hat{\mathcal{I}}_O$.

finite-sample FDR, the same to the existing knockoff methods (Barber and Candès 2015; Candès et al. 2018).

We first show that the knockoff statistics Δ_j^L and Δ_j^N in (16) have symmetric properties for a null linear or nonlinear component $j \in \hat{\mathcal{S}}^L$ or $\hat{\mathcal{S}}^N$ respectively. The symmetric property of the null components are crucial for the KI-LAND procedure, which then selects a data-dependent threshold while controlling the FDR (Barber and Candès 2015).

Assumption 1 (PLM decomposition)**.** *For input* $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{R}^d$*, the response* y *can be represented as follow by the decomposition (5)*

$$\begin{aligned} y &= y^L + y^N + \varepsilon^L + \varepsilon^N \\ &= \sum_{j=1}^d \beta_j k_1(x_j) + \sum_{j=1}^d g_{1j}(x_j) + \varepsilon^L + \varepsilon^N. \end{aligned}$$

Assumption 2 (Irrepresentable Condition in RKHS)**.** *For* $j \in \{1, \ldots, d\}, x_j \neq \sum_{k=1;k\neq j}^d \beta_k x_k$*, for any* $\beta_k \in \mathcal{R}$*; and* $g_j(x_j) \neq \sum_{k=1;k\neq j}^d g_k(x_k)$*, for any functions* $g_k \in \mathcal{H}_1$.

Proposition 1. *Suppose Assumption 1 and 2 hold.*
i) $j \in \mathcal{S}_0^L$*, if and only if* $\beta_j = 0$*, for* $j = 1, \ldots, d$*;*
ii) $j \in \mathcal{S}_0^N$*, if and only if* $g_j = 0$*, for* $j = 1, \ldots, d$.

Lemma 2. *Take any subset* $\mathcal{S}$ *of null components, then*
i) For $\mathcal{S} \subset \mathcal{S}_0^L$*,* $[\mathbf{X}, \tilde{\mathbf{X}}] \mid Y^L \overset{\mathrm{d}}{=} [\mathbf{X}, \tilde{\mathbf{X}}]_{swap(\mathcal{S})} \mid Y^L$*;*
ii) For $\mathcal{S} \subset \mathcal{S}_0^N$*,* $[\mathbf{R}, \tilde{\mathbf{R}}] \mid Y^N \overset{\mathrm{d}}{=} [\mathbf{R}, \tilde{\mathbf{R}}]_{swap(\mathcal{S})} \mid Y^N$.

Remark 1. $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathcal{R}^{n\times 2d}$ *is the gram matrix of original and knockoff variables.* $[\mathbf{R}, \tilde{\mathbf{R}}] \in \mathcal{R}^{n\times 2nd}$ *is consist of kernel matrix of original and knockoff variables. For more details, please see the appendix.* $\mathcal{S}_0^L$*,* $\mathcal{S}_0^N$ *are true nulls of nonlinear and nonlinear component index set.* $\mathcal{S}^L$*,* $\mathcal{S}^N$ *are true linear and nonlinear component index set.*

Lemma 3 (Sign-flip property for the nulls)**.** *Suppose Assumptions (1) and (2) hold. Let* $(\epsilon_1, \ldots, \epsilon_d)$ *be a set of independent random variables,*
i) such that $\epsilon_j = 1$ *if* $j \in \mathcal{S}^L$*, and* $\epsilon_j = \pm 1$ *with equal probability 1 / 2 if* $j \in \mathcal{S}_0^L$ *Then,* $\left(\Delta_1^L, \ldots, \Delta_d^L\right) \overset{d}{=} \left(\Delta_1^L \cdot \epsilon_1, \ldots, \Delta_d^L \cdot \epsilon_d\right)$.
ii) such that $\epsilon_j = 1$ *if* $j \in \mathcal{S}^N$*, and* $\epsilon_j = \pm 1$ *with equal probability 1 / 2 if* $j \in \mathcal{S}_0^N$ *Then,* $\left(\Delta_1^N, \ldots, \Delta_d^N\right) \overset{d}{=} \left(\Delta_1^N \cdot \epsilon_1, \ldots, \Delta_d^N \cdot \epsilon_d\right)$.

The following part introduces that KI-LAND procedure controls the false discovery for any sample size well. The results which are established without the need for prior knowledge of the noise level are robust to the distribution and number of predictor variables. we adopt a modified FDR (mFDR) to approximate FDR when using the threshold T (17) (Barber and Candès 2015). For a true set $\mathcal{S}$, selected set $\hat{\mathcal{S}}$ and true nulls $\mathcal{S}_0$, FDR and mFDR are defined as,

$$\mathrm{FDR}(\hat{\mathcal{S}}) = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{S}_0|}{|\hat{\mathcal{S}}| \vee 1}\right], \mathrm{mFDR}(\hat{\mathcal{S}}) = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{S}_0|}{|\hat{\mathcal{S}}| + 1/q}\right].$$

Theorem 1 (FDR control of KI-LAND)**.** *For any* $q \in (0,1)$ *and any sample size n, choose the threshold* $T^L > 0$ *and* $T^N > 0$ *via T (17) . Then* $\mathbf{mFDR}(\widehat{\mathcal{I}}_L) \leq q$ *and* $\mathbf{mFDR}(\widehat{\mathcal{I}}_N) \leq q$*; Meanwhile, choose the threshold* $T^L > 0$ *and* $T^N > 0$ *via* T_+ *(18). Then* $\mathbf{FDR}(\widehat{\mathcal{I}}_L) \leq q$ *and* $\mathbf{FDR}(\widehat{\mathcal{I}}_N) \leq q$.

Remark 2. *Theorem 1 controls the valid FDR with no restriction on the dimension* d *and the sample size* n.

Power Analysis

As far as we know, the current literature lacks thorough theoretical analysis about power for knockoff methods, except (Fan et al. 2020; Weinstein et al. 2022) which analyzed the power for linear regressions under the model-X knockoff framework and for thresholded Lasso knockoffs respectively. In this section, we will show that the KI-LAND procedure has a full power as the $n \to \infty$. To obtain the theoretical results, some basic regularity conditions are introduced.

Assumption 3 (Minimum signal)**.** *For some slowly diverging sequence* $\kappa_n \to \infty$*, as* $n \to \infty$*, let* $\eta \equiv c_\eta\{n^{-\beta/(2\beta+1)} + [(\log d)/n]^{1/2}\}$*. For some constant* $c_\eta > 0$*, such that,* $\min_{j\in\mathcal{S}^L} |\beta_j| \geq \kappa_n\{(\log d)/n\}^{1/2}, \min_{j\in\mathcal{S}^N} \|g_{1j}(x_j)\|_2 \geqslant \kappa_n\eta$*, where the RKHS* $\mathcal{H}_{1j}$ *is embedded to a* β*th order Sobolev space with* $\beta > 1$.

In order to introduce the following conditions, some notations are given. $\mathbf{\Sigma} = [\mathbf{R}, \tilde{\mathbf{R}}] \in \mathcal{R}^{n\times 2nd}$, $\mathbf{\Sigma}_{\mathcal{S}^N} \in \mathcal{R}^{n\times 2n|\mathcal{S}^N|}$

is the design matrix consisted of the jth and $j+d$th columns of $\mathbf{\Sigma}$ for $j \in \mathcal{S}^N$. The same to other matrixs $[\mathbf{X}, \tilde{\mathbf{X}}]^\top_{\mathcal{S}^L}, \mathbf{X}_{\mathcal{S}^L}$.

Assumption 4 (Minimal eigenvalue)**.** *Suppose here is a constant $C_{\min} > 0$, such that the minimal eigenvalue λ_{min} of matrix $\mathbb{E}[n^{-1}[\mathbf{X}, \tilde{\mathbf{X}}]^\top_{\mathcal{S}^L}[\mathbf{X}, \tilde{\mathbf{X}}]_{\mathcal{S}^L}]$ and $\mathbb{E}[n^{-1}\mathbf{\Sigma}^\top_{\mathcal{S}^N}\mathbf{\Sigma}_{\mathcal{S}^N}]$ satisfies that,*

$$\lambda_{\min}(\mathbb{E}[n^{-1}[\mathbf{X}, \tilde{\mathbf{X}}]^\top_{\mathcal{S}^L}[\mathbf{X}, \tilde{\mathbf{X}}]_{\mathcal{S}^L}]) \geqslant C_{\min},$$
$$\lambda_{\min}(\mathbb{E}[n^{-1}\mathbf{\Sigma}^\top_{\mathcal{S}^N}\mathbf{\Sigma}_{\mathcal{S}^N}]) \geqslant C_{\min}.$$

Assumption 5 (Mutual incoherence)**.** *Suppose there exists a constant $0 \leqslant \xi < 1$, such that,*

$$\max_{j \notin \mathcal{S}^L} \|[\mathbf{X}, \tilde{\mathbf{X}}]^{\mathrm{T}}_j [\mathbf{X}_{\mathcal{S}^L}, \tilde{\mathbf{X}}_{\mathcal{S}^L}]$$
$$([\mathbf{X}_{\mathcal{S}^L}, \tilde{\mathbf{X}}_{\mathcal{S}^L}]^{\mathrm{T}}[\mathbf{X}_{\mathcal{S}^L}, \tilde{\mathbf{X}}_{\mathcal{S}^L}])^{-1}\|_2 \leq 1 - \xi.$$

Assumption 6 (Mutual incoherence)**.** *Suppose $d < e^n$ and there exists a constant $0 \leqslant \xi_{\mathbf{\Sigma}} < 1$, such that,*

$$\max_{j \notin \mathcal{S}^N} \|\{\mathbf{\Sigma}_j\}^\top \mathbf{\Sigma}_{\mathcal{S}^N}[\mathbf{\Sigma}^\top_{\mathcal{S}^N}\mathbf{\Sigma}_{\mathcal{S}^N}]^{-1}\|_2 \leqslant \xi_{\mathbf{\Sigma}},$$
$$\frac{\xi_{\mathbf{\Sigma}}\sqrt{|\mathcal{S}^N|}+1}{\lambda_2}\eta + \xi_{\mathbf{\Sigma}}\sqrt{|\mathcal{S}^N|} < 1,$$

where $\mathbf{\Sigma}_j$ is the the jth columns of $\mathbf{\Sigma}$.

Remark 3. *All these assumptions are reasonable and used frequently in current literature. In detail, Assumption 3 is a minimum regulatory effect condition which ensures that the solution of LAND model (8) does not overlook a great portion of important variables. We can find the same assumption in (Zhao and Yu 2006; Ravikumar, Wainwright, and Lafferty 2010). Assumption 4 is the minimal eigenvalue condition, which states that the Gram matrix of the important set on the augment design matrix is invertible. Similar assumptions have been imposed in Lasso regressions (Ravikumar, Wainwright, and Lafferty 2010; Raskutti, J Wainwright, and Yu 2012; Fan et al. 2020). Assumptions 5 and 6 indicate that the correlation between the true signals and nulls should not exert an strong effect (Zhao and Yu 2006; Ravikumar, Wainwright, and Lafferty 2010; Wainwright 2019).*

With these regularity conditions in place, we are now ready to characterize the statistical power of the KI-LAND procedure. For a true set $\mathcal{S}$ and a selected set $\hat{\mathcal{S}}$, the power is defined as,

$$\mathrm{Power}(\widehat{\mathcal{S}}) = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{|\mathcal{S}| \vee 1}\right].$$

Theorem 2 (Power of KI-LAND)**.** *Suppose Assumptions 3-6 hold, Then, the oracle KI-LAND procedure has a asymptomatic property that $\mathrm{Power}(\hat{\mathcal{I}}_L) \to 1$ and $\mathrm{Power}(\hat{\mathcal{I}}_N) \to 1$, as $n \to \infty$.*

Remark 4. *Theorems 1 and 2 show that the KI-LAND procedure can simultaneously provide good false discovery rate (FDR) control and statistical power properties under some regularity settings. In contrast to Theorem 1, Theorem 2 holds for $d < n$ and $n < d < e^n$.*

Experiments

In this section, we conduct experiments to evaluate the empirical performance of the proposed KI-LAND, where various data settings are considered including the different correlation, the different sample size and the different dimension of variables. In all cases, we set $L = 100$, and select the tuning parameter $(\lambda_1, \tau_0, \tau_1)$ by cross-validation. And the target FDR of $\hat{\mathcal{I}}_L$ and $\hat{\mathcal{I}}_N$ are 0.2.

The simulated data is generated to demonstrate the empirical performance of the KI-LAND procedure. We compare the KI-LAND with LAND (Zhang and Liu 2011), KKO (Xiaowu Dai and Li 2023), and Model-X (Candès et al. 2018). The following functions on [0, 1] are used for building components of response y:

$$\begin{aligned}
f_1(x) =& x, \quad f_2(x) = \cos(2\pi x),\\
f_3(x) =& \sin(2\pi x)/(2-\sin(2\pi x)),\\
f_4(x) =& \cos(2\pi x)/(2-\cos(2\pi x)),\\
f_5(x) =& 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3(\sin(2\pi x))^2\\
&+ 0.4(\cos(2\pi x))^3 + 0.5(\sin(2\pi x))^3,\\
f_6(x) =& (3x-1)^2, \quad f_7(x) = (3x-1)^3.
\end{aligned}$$

where f_1 is a linear function, $f_2, ..., f_5$ are pure nonlinear function, and f_6, f_7 are consist of linear and nonlinear functions. For $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{R}^d$, y is generated by the model:

$$\begin{aligned}
y =& \sum_{j=1}^{10} f_1(x_j) + f_2(x_{11}) + 3f_2(x_{12}) + f_3(x_{13})\\
&+ 2f_3(x_{14}) + f_4(x_{15}) + 4f_4(x_{16}) + f_5(x_{17})\\
&+ f_5(x_{18}) + f_6(x_{19}) + f_7(x_{20}) + \varepsilon,
\end{aligned}$$

where $\varepsilon \sim \mathcal{N}(0,1)$, the designed matrix $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega})$, $\mathbf{\Omega} = \left(\rho^{|i-j|}\right)_{1 \leq i,j \leq p}$, $\rho = \mathrm{corr}\,(X_i, X_j)$ for all $i \neq j$. It is clearly that $x_1, ..., x_{10}$ are the linear variables, $x_{11}, ..., x_{20}$ are the nonlinear variables, and the rest is $d-20$ noise variables. So $|\mathcal{I}_L| = 10$, $|\mathcal{I}_N| = 10$, $|\mathcal{I}_O| = d - 20$.

For the different variable dimensions $d = \{50, 100, 200, 400\}$, we set $n = 200$, $\rho = 0.2$, these cases examine whether the proposed method performs well in terms of FDR and statistical power when the important variables become more sparse. For the different sample size $n = \{50, 100, 200, 400\}$, we set $d = 100, \rho = 0.2$, these cases examine whether the proposed method performs well in term of power when the sample size becomes larger. The above cases include both $d > n$ and $d < n$. For the different correlations $\rho = \{0, 0.2, 0.5, 1\}$, we set $d = 100, n = 200$, these cases examine whether the proposed method performs well when the correlation between variables is increasing.

Table 1 and Figure 1 show that KI-LAND procedure can achieve a good performance in terms of FDR control and power. As the sample size increasing, the power of both the linear and nonlinear components exhibits an upward trend. Conversely, the power shows an opposite trend of change when the variable dimension decreases or the correlation increases. This phenomenon is in line with expectations. Although KI-LAND has a weaker power than KKO, it

		FDR					Power				
		KI-LAND (L)	KI-LAND (N)	KKO	Model-X	LAND	KI-LAND (L)	KI-LAND (N)	KKO	Model-X	LAND
n	50	0.1667	0.2000	0.1875	0.2000	0.3333	50.00%	40.00%	65.00%	20.00%	60.00%
	100	0.1250	0.1429	0.1500	0.1667	0.3684	70.00%	60.00%	85.00%	25.00%	60.00%
	200	0.2000	0.1429	0.1429	0.1429	0.3182	80.00%	60.00%	**90.00%**	30.00%	75.00%
	400	0.1818	0.1111	0.1429	0.2000	0.2963	**90.00%**	**80.00%**	**90.00%**	**40.00%**	**95.00%**
d	50	0.1111	0.1111	0.1429	0.2000	0.2727	80.00%	**80.00%**	**90.00%**	**40.00%**	**80.00%**
	100	0.1818	0.1111	0.1905	0.1667	0.3000	**90.00%**	80.00%	85.00%	25.00%	70.00%
	200	0.1250	0.1429	0.1500	0.1667	0.3684	70.00%	60.00%	85.00%	25.00%	60.00%
	400	0.1429	0.1429	0.2000	0.2000	0.4211	60.00%	60.00%	80.00%	20.00%	55.00%
ρ	0	0.1000	0.2000	0.0500	0.1429	0.2500	**90.00%**	**80.00%**	**95.00%**	**30.00%**	**75.00%**
	0.2	0.1250	0.1429	0.1500	0.1667	0.3684	70.00%	60.00%	85.00%	25.00%	60.00%
	0.5	0.1429	0.1429	0.1000	0.2000	0.3000	60.00%	60.00%	90.00%	20.00%	70.00%
	1	0.1667	0.1667	0.2000	0.2000	0.4000	50.00%	50.00%	80.00%	20.00%	60.00%

Table 1: Comparisons in terms of FDR and power with the different sample sizes, variable dimensions and correlations.

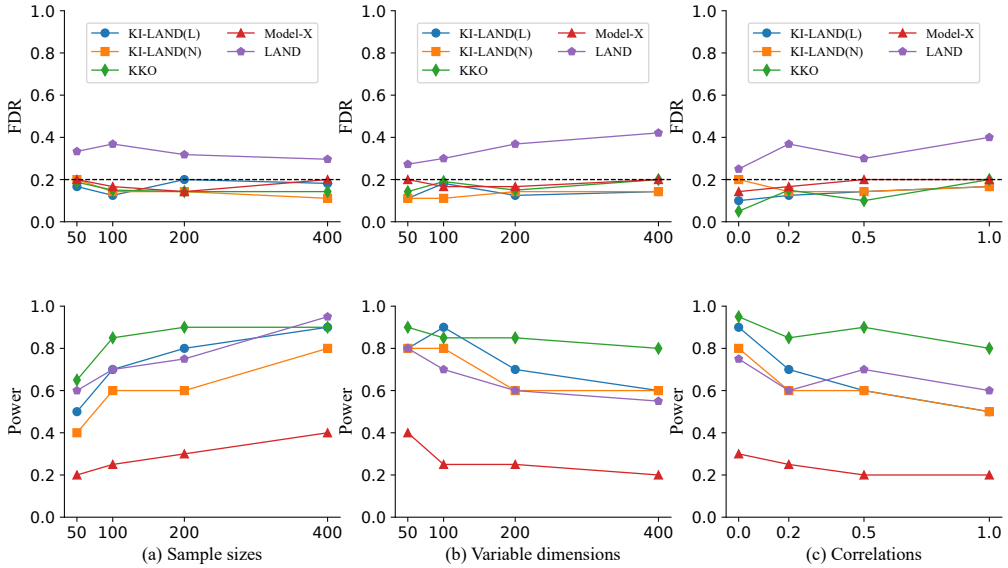

Figure 1: Empirical performance and comparisons in terms of FDR and power with the different (a) sample sizes, (b) variable dimensions and (c) correlations. (L) and (N) denote the experiments of KI-LAND on the set $\hat{\mathcal{I}}_L$ and $\hat{\mathcal{I}}_N$, respectively.

can identify linear and nonlinear variables while KKO just implements variable selection. The low power of model-X requires further explanation, as it stems from the use of Lasso which is unable to capture nonlinear relationships. The changing trend of FDR value are more complex, but we focus on whether it can be controlled below the target level $q = 0.2$. Table 1 shows that FDR can be controlled by all knockoffs methods including KI-LAND, KKO and model-X. Compared to LAND, KI-LAND can control FDR successfully while LAND inflates FDR. The experiments with real-world dataset are in the supplementary materials.

Conclusion

This paper formulates a new Knockoffs Inference scheme for Linear And Nonlinear Discoverer (KI-LAND), where FDR is controlled with respect to both linear and nonlinear variables for automatic structure discovery. Compared with the most related works (Zhang and Liu 2011; Xiaowu Dai and Li 2023), experimental results have shown that the proposed KI-LAND exhibits strong competitiveness. In theory, we have established the fine-grained characterizations on the FDR controllability and the power performance. The full version of the paper (including supplementary materials) can be found at arXiv.

Acknowledgments

The work was supported in part by the National Natural Science Foundation of China under Grants 62376104, 12426512, and 12071166, in part by the Fundamental Research Funds for the Central Universities of China (No. 2662023LXPY005), and in part by Hubei Provincial Natural Science Foundation of China (No.2023AFB523).

References

Barber, R. F.; and Candès, E. J. 2015. Controlling the false discovery rate via knockoffs. *The Annals of statistics*, 2055–2085.

Barber, R. F.; and Candès, E. J. 2016. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*.

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

Candès, 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.

Chong, G. 2013. *Smoothing spline ANOVA models*, volume 297. Springer.

Dümbgen, L.; Samworth, R. J.; and Schuhmacher, D. 2013. Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, volume 9, 78–91. Institute of Mathematical Statistics.

Engle, R. F.; Granger, C. W.; Rice, J.; and Weiss, A. 1986. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394): 310–320.

Fan, Y.; Demirkaya, E.; Li, G.; and Lv, J. 2020. RANK: Large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*.

Fuller, W. A. 2009. *Measurement error models*. John Wiley & Sons.

Härdle, W.; Liang, H.; and Gao, J. 2000. *Partially linear models*. Springer Science & Business Media.

Härdle, W.; Müller, M.; Sperlich, S.; Werwatz, A.; et al. 2004. *Nonparametric and semiparametric models*, volume 1. Springer.

Lian, H.; Zhao, K.; and Lv, S. 2019. Projected spline estimation of the nonparametric function in high-dimensional partially linear models for massive data. *The Annals of Statistics*, 47(5): 2922–2949.

Lin, Y.; and Zhang, H. H. 2006. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5): 2272 – 2297.

Lv, S.; and Lian, H. 2022. Debiased distributed learning for sparse partial linear models in high dimensions. *Journal of Machine Learning Research*, 23(2): 1–32.

Meinshausen, N.; and Bühlmann, P. 2010. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4): 417–473.

Raskutti, G.; J Wainwright, M.; and Yu, B. 2012. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of machine learning research*, 13(2).

Ravikumar, P.; Wainwright, M. J.; and Lafferty, J. D. 2010. High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287 – 1319.

Ruppert, D.; Wand, M. P.; and Carroll, R. J. 2003. *Semiparametric regression*. 12. Cambridge university press.

Storlie, C. B.; Bondell, H. D.; Reich, B. J.; and Zhang, H. H. 2011. Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21(2): 679.

Su, H.; Yuan, P.; Sun, Q.; Yi, M.; and Li, G. 2023. StabGKnock: Controlled variable selection for partially linear models using generalized knockoffs. *arXiv preprint arXiv:2311.15982*.

Su, W.; and Candès, E. 2016. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3): 1038 – 1068.

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

Wahba, G. 1983. *Cross validated spline methods for the estimation of multivariate functions from data on functionals*. University of Wisconsin, Department of Statistics.

Wainwright, M. J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, H.; Li, G.; and Jiang, G. 2007. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3): 347–355.

Wang, L.; Xue, L.; Qu, A.; and Liang, H. 2014a. Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, 42(2): 592 – 624.

Wang, L.; Xue, L.; Qu, A.; and Liang, H. 2014b. Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, 42(2): 592 – 624.

Weinstein, A.; Su, W. J.; Bogdan, M.; Barber, R. F.; and Candès, E. J. 2022. A Power Analysis for Model-X Knockoffs with ℓ_p-Regularized Statistics. arXiv:2007.15346.

Xiaowu Dai, X. L.; and Li, L. 2023. Kernel Knockoffs Selection for Nonparametric Additive Models. *Journal of the American Statistical Association*, 118(543): 2158–2170. PMID: 38143786.

Xie, H.; and Huang, J. 2009. SCAD-penalized regression in high-dimensional partially linear models. *Annals of Statistics*, 37: 673–696.

Yaniv Romano, M. S.; and Candès, E. 2020. Deep Knockoffs. *Journal of the American Statistical Association*, 115(532): 1861–1872.

Zhang, G. C.; and Liu, Y. 2011. Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models. *Journal of the American Statistical Association*, 106(495): 1099–1112. PMID: 22121305.

Zhang, H. H.; and Lu, W. 2007. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3): 691–703.

Zhao, P.; and Yu, B. 2006. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476): 1418–1429.