Title

Kernel Knockoffs Selection for Nonparametric Additive Models

Journal

Journal of the American Statistical Association, 118(543)

ISSN

0162-1459

Authors

Dai, Xiaowu
Lyu, Xiang
Li, Lexin

Publication Date

2023-07-03

DOI

10.1080/01621459.2022.2039671

HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

J Am Stat Assoc. 2023 ; 118(543): 2158–2170. doi:10.1080/01621459.2022.2039671.

Kernel Knockoffs Selection for Nonparametric Additive Models

Xiaowu Dai,

Xiang Lyu,

Lexin Li

University of California, Berkeley

Abstract

Thanks to its fine balance between model flexibility and interpretability, the nonparametric additive model has been widely used, and variable selection for this type of model has been frequently studied. However, none of the existing solutions can control the false discovery rate (FDR) unless the sample size tends to infinity. The knockoff framework is a recent proposal that can address this issue, but few knockoff solutions are directly applicable to nonparametric models. In this article, we propose a novel kernel knockoffs selection procedure for the nonparametric additive model. We integrate three key components: the knockoffs, the subsampling for stability, and the random feature mapping for nonparametric function approximation. We show that the proposed method is guaranteed to control the FDR for any sample size, and achieves a power that approaches one as the sample size tends to infinity. We demonstrate the efficacy of our method through intensive simulations and comparisons with the alternative solutions. our proposal thus makes useful contributions to the methodology of nonparametric variable selection, FDR-based inference, as well as knockoffs.

Keywords

False discovery rate; Knockoffs; Nonparametric additive models; Reproducing kernel Hilbert space; Subsampling; Variable selection

1 Introduction

In the past decades, the nonparametric additive model has been widely used in statistics and machine learning, thanks to its fine balance between model flexibility and model interpretability (Stone, 1985; Hastie and Tibshirani, 1990; Wood, 2017). For a univariate response variable $Y \in \mathbb{R}$ and p predictor variables $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathcal{X}^p \subseteq \mathbb{R}^p$, the model postulates that,

$$Y = \mu + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{1}$$

[1]*Address for correspondence:* Lexin Li, Department of Biostatistics and Epidemiology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, USA. lexinli@berkeley.edu.

The first two authors contribute equally to the paper.

where μ is the intercept, $f_j: \mathcal{X} \mapsto \mathbb{R}$, with $\mathbb{E}_{X_j}[f_j(X_j)] = 0$, $j = 1, \ldots, p$, are the component functions that are modeled nonparametrically, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the random error, with unknown σ. Furthermore, we assume throughout this article that the component functions f_j's reside in a reproducing kernel Hilbert space (RKHS, Aronszajn, 1950; Wahba, 1990).

Variable selection for the nonparametric additive model dates back to Lin and Zhang (2006), and has seen substantial developments ever since (Meier et al., 2009; Ravikumar et al., 2009; Huang et al., 2010; Koltchinskii and Yuan, 2010; Wood, 2017, among others). In particular, Lin and Zhang (2006) proposed a component selection and smoothing operator (COSSO) penalty that extends the Lasso penalty to the nonparametric additive model, and penalized the sum of the reproducing kernel Hilbert space norms of the component functions. Meanwhile, Meier et al. (2009) and Ravikumar et al. (2009) both employed basis expansion, and penalized the sparsity and smoothness seminorms. Huang et al. (2010) employed the group Lasso penalty to obtain an initial estimator and to reduce the dimension of the problem, then employed the adaptive group Lasso to select nonzero components. This family of methods guarantee the asymptotic optimality of the function estimation and the selection consistency as the sample size tends to infinity. However, none has achieved the control of false discovery rate (FDR) unless the sample size tends to infinity. There has been another family of solutions that target simultaneous testing of multiple hypotheses and concentrate on controlling some forms of false discovery (see, e.g., Benjamini and Hochberg, 1995; Efron et al., 2001; Storey, 2007; Sun and Cai, 2007, 2009, among others); see also Cai and Sun (2017) for a review. Nevertheless, none of the existing solutions in this family directly addresses the problem of variable selection for the nonparametric additive model while controlling the false discovery at the same time.

More recently, Barber and Candes (2015) proposed a powerful framework called knockoffs that effectively controls the FDR for variable selection in the linear model under the finite-sample setting, in the sense that the sample size does not have to go to infinity. The key idea is to construct a set of so-called "knockoff variables" that are not associated with the response conditioning on the original variables, while the structure of the knockoff variables mimics that of the original ones. It then computes an importance score for each variable, and selects those that have considerably higher scores than their knockoff counterparts. There have then been numerous generalizations of this work; see Barber et al. (2020) and many references therein. Related to our target of high-dimensional nonparametric additive model, Barber and Candes (2019) considered the high-dimensional fixed design, and focused on the linear model only. Dai and Barber (2016) developed an "expansion first" strategy that performs feature expansion first then constructs the knockoffs based on the expanded features, and proposed to employ a group Lasso penalty for subsequent variable selection. They actually still studied the linear regression setting; however, their proposal can, in principle, be extended to the nonparametric additive model. Candès et al. (2018) proposed a model-X knockoffs extension that works for random designs of predictors and allows the conditional distribution of the response given the predictors to be arbitrary and unknown, though they mostly focused on the step of how to generate the knockoff variables. Fan et al. (2020) further built on the model-X knockoffs framework, and developed a knockoffs-based variable selection procedure that is applicable to the nonparametric additive

model. It employs data splitting, uses half of the data to estimate the predictor precision matrix and screen the predictors, and uses the other half to perform knockoffs based on some empirical norm of the estimated component functions. These pioneering works have opened the door for knockoffs-based selection for the nonparametric additive model. However, some may be difficult to extend beyond the linear model, and others suffer a limited power or expensive computation. In addition, there is generally a lack of theoretical power analysis for the existing knockoffs-based methods, except for Fan et al. (2020) and Weinstein et al. (2020), who made important first steps, but only studied the power behavior for the linear model.

In this article, we propose a novel kernel knockoffs selection procedure for the nonparametric additive model (1). We build on and integrate three key components: the knockoffs, the subsampling for stability, and the random feature mapping for nonparametric function approximation in RKHS. Specifically, we employ the random feature mapping (Rahimi and Recht, 2007) to approximate the component function f_j, and construct a projection operator between the RKHS and the original predictor space. Such a projection allows us to define an analog of the effect size of the individual predictor in the setting of nonparametric additive model. We then construct the importance score based on the projected component function f_j, instead of the original predictor X_j or its knockoff. Moreover, we note that the random features may introduce additional stochastic errors, which can disturb the order of the variables entering the model and lead to both false positives and false negatives. We thus further employ the subsampling strategy to improve the selection stability (Meinshausen and Buhlmann, 2010). That is, we subsample the data and apply the random feature mapping multiple times, and compute the importance score as the difference of selection frequencies over subsampling replications between each predictor and its knockoff counterpart. We show that the proposed method is guaranteed to control the FDR below the nominal level under any sample size, and achieves a power that approaches one as the sample size tends to infinity.

Our proposal makes useful contributions to the methodology and theory of nonparametric variable selection, FDR-based inference, as well as knockoffs.

First, whereas the methods such as Lin and Zhang (2006); Ravikumar et al. (2009) have obtained the variable selection consistency asymptotically, there has been no existing method that controls the FDR in the setting of nonparametric additive model for the finite-sample setting. A low FDR in such a setting assures that most of the discoveries are indeed true under any given sample size. By contrast, the asymptotic control is valid only when the sample size goes to infinity, which can be problematic for the applications with limited sample sizes. In those cases, the asymptotic control may provide little guidance on quantifying the threshold of variable selection (Barber and Candes, 2015), and it is possible that the selected variables may still include many unrelated ones (Su et al., 2017). On the other hand, the classical reproducing kernel methods usually involve non-separable variables and their knockoffs, which renders the FDR control infeasible. To address this challenge, we employ the random feature mapping to ensure the exchangeability of the null variables and their knockoffs, and in turn achieve the finite-sample FDR control. The random feature mapping nevertheless introduces an extra layer of randomness. We

further resort to subsampling, construct an importance score by averaging over multiple subsampling replications, and show these techniques can handle the extra randomness; see Section S1.2 of the Appendix for more technical details. In short, the finite-sample FDR control for nonparametric variable selection has been a long-standing and open question, and our proposal is among the first solutions for this type of question.

Second, we employ the subsampling strategy, but it is different from the existing subsampling based methods for FDR control. Specifically, Bach (2008) proposed a selection method based on bootstrap replications, which may suffer from a limited power when the sample size does not go to infinity. Meinshausen and Buhlmann (2010) proposed a stability-based procedure to select the variables with the selection frequencies exceeding a threshold level over the entire solution path, which guarantees the control of the expected number of false positives, but may fail to control the FDR. Li et al. (2013) proposed to use a mixture model for the distribution of selection frequencies, whereas Ahmed et al. (2011); He et al. (2016) suggested to estimate this distribution via permutations. However, such a distribution estimation requires either strict parametric assumptions, or expensive computations. By contrast, our method does not require estimation of the distribution of selection frequencies, but uses subsampling to tackle the extra randomness introduced by random feature mapping and to achieve the FDR control.

Last but not least, our method expands the scope of the currently fast growing area of knockoffs. Compared to Barber and Candes (2019) who focused on the high-dimensional linear model only, we generalize the knockoffs to the high-dimensional nonparametric additive model. Such an extension is far from incremental, as it has to deal with nonlinear dependency between the response and predictors, as well as the non-separability of variables of the usual reproducing kernel methods. Compared to Dai and Barber (2016) who did “expansion first”, we adopt a “knockoffs first” strategy, which leads to an easier construction of the knockoff variables, ensures a good statistical power, and is computationally more economical. Compared to Candès et al. (2018) who developed the model-X framework, but can not handle the nonparametric additive model directly, and did not provide any formal theoretical justification for the model-X knockoffs beyond the generalized linear model setting, we propose a new and effective importance score based on random feature mapping and resampling, and we explicitly study the power behavior in the nonparametric setting. Finally, compared to Fan et al. (2020) who extended the model-X framework, and proposed a novel importance score based on the basis functions and some empirical norm of the estimated component functions, but only studied the power for the linear model, we again develop a new importance score that is built on the selection probability of the variables and their knockoffs. As a result, we show that our solution is more powerful and also more robust to the data distribution. Moreover, we establish the power guarantee for the nonparametric additive model, which is more challenging than the linear model case as it involves nonlinear associations. Toward that end, we employ some functional data analysis techniques and concentration inequalities for functional empirical processes to study the spectral properties; see Section S1.4 of the Appendix for more technical details. We also briefly comment that, our theoretical tools are applicable to the FDR control for more general nonparametric models, e.g., the functional analysis of variance type models that involve higher-order interactions (Wahba et al., 1995; Lin and Zhang, 2006). Moreover,

in Section 5, we further compare with some of these key alternative knockoff solutions numerically, and demonstrate the advantages of our proposed method empirically as well.

The rest of the article is organized as follows. Section 2 formulates the problem. Section 3 develops the kernel knockoffs procedure. Section 4 establishes the theoretical guarantees on the FDR and power. Section 5 presents the simulations, and also an analysis of brain imaging data. The Supplementary Appendix collects all proofs and some additional numerical results.

2 Problem Setup

2.1 Kernel learning

Throughout this article, we consider regression functions that reside in an infinite-dimensional reproducing kernel Hilbert space. We begin with a Mercer kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

$$K(X, X') = \sum_{\upsilon=1}^{\infty} \tilde{\lambda}_\upsilon \widetilde{\psi}_\upsilon(X) \widetilde{\psi}_\upsilon(X'),$$

where $\{\widetilde{\psi}_\nu\}_{\nu=1}^{\infty}$ are eigenfunctions, $\{\tilde{\lambda}_\nu\}_{\nu=1}^{\infty}$ are eigenvalues of the integral operator defined by the kernel function, and $\tilde{\lambda}_\nu \widetilde{\psi}_\nu(X) = \int_{\mathcal{X}} K(X, X') \widetilde{\psi}_\nu(X') dX'$ (Mercer, 1909). The domain $\mathcal{X} \subseteq \mathbb{R}$ can be either a compact or an unbounded space. We consider the RKHS $\mathscr{H}_1$ generated by this kernel, which is defined as the closure of linear combinations of the basis functions $\{\widetilde{\psi}_\nu\}_{\nu=1}^{\infty}$ as follows, where $\overline{\{\cdot\}}$ denotes the closure of a function space,

$$\mathscr{H}_1 = \overline{\left\{ f: \mathcal{X} \to \mathbb{R} \middle| f(X) = \widetilde{\boldsymbol{\Psi}}(X)^\top \tilde{c}, \text{ and } \|f\|_K < \infty \text{ with } \|f\|_K^2 = \sum_{\nu=1}^{\hbar} \frac{\tilde{c}_\nu^2}{\tilde{\lambda}_\nu} \right\}}.$$

Here $\widetilde{\boldsymbol{\Psi}}(X)$ is an infinite-dimensional vector with the νth element equal to $\sqrt{\tilde{\lambda}_\nu} \widetilde{\psi}_\nu(X)$, and $\tilde{c}$ is an infinite-dimensional coefficient vector with the νth element $\tilde{c}_\upsilon$, $\nu = 1, 2, \ldots$.

Next, define the kernel $K_p: \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}$,

$$K_p\left((X_1, \ldots, X_p)^\top, (X_1', \ldots, X_p')^\top\right) = K(X_1, X_1') + \ldots + K(X_p, X_p').$$

The RKHS $\mathscr{H}_p$ generated by K_p is of the form (Aronszajn, 1950),

$$\mathscr{H}_p = \mathscr{H}_1 \oplus \ldots \oplus \mathscr{H}_1 = \{ f: \mathcal{X}^p \to \mathbb{R} | f(\boldsymbol{X}) = f(X_1, \ldots, X_p) = f_1(X_1) + \ldots + f_p(X_p), \\ f_j \in \mathscr{H}_1, \text{ and } \mathbb{E}[f_j(X_j)] = 0, \ j = 1, \ldots, p \}$$

Suppose the observed training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ consist of n i.i.d. copies of $(\boldsymbol{X}, Y)$ following the nonparametric additive model (1), with $\boldsymbol{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. The representer theorem (Wahba, 1990) shows that the solution to the kernel learning problem when restricting $f \in \mathscr{H}_p$,

$$\min_{f \in \mathscr{H}_p} \left[n^{-1} \sum_{i=1}^{n} \mathscr{L}(f(\boldsymbol{x}_i), y_i) + \lambda \|f\|_{K_p}^2 \right],$$

for some loss function $\mathscr{L}$, the kernel K_p, and the penalty parameter λ, is of the form,

$$\tilde{f}(\boldsymbol{X}) = \sum_{i=1}^{n} \alpha_i K_p(\boldsymbol{X}, \boldsymbol{x}_i),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$ are the corresponding coefficients. This in effect turns an infinity-dimensional optimization problem to an optimization problem over n parameters. This minimizer can be further written as, for any $\boldsymbol{X} \in \mathscr{X}^p$,

$$\tilde{f}(\boldsymbol{X}) = \widetilde{\boldsymbol{\Psi}}_p(\boldsymbol{X})^\top \tilde{\boldsymbol{c}}_p, \tag{2}$$

where $\widetilde{\boldsymbol{\Psi}}_p(\boldsymbol{X}) = \left[\widetilde{\boldsymbol{\Psi}}(X_1)^\top, \ldots, \widetilde{\boldsymbol{\Psi}}(X_p)^\top\right]^\top$ assembles $\widetilde{\boldsymbol{\Psi}}(X_j)'$s and is an infinite-dimensional vector, and $\tilde{\boldsymbol{c}}_p = \left[\widetilde{\boldsymbol{\Psi}}_p(\boldsymbol{x}_1), \ldots, \widetilde{\boldsymbol{\Psi}}_p(\boldsymbol{x}_n)\right]\boldsymbol{\alpha}$ is the infinite-dimensional coefficient vector.

2.2 Variable selection for nonparametric additive models

Next, we formally frame variable selection in the context of nonparametric additive models. We say a variable X_j is null if and only if Y is independent of X_j conditional on all other variables $\boldsymbol{X}_{-j} = \{X_1, \ldots, X_p\} \backslash \{X_j\}$, i.e., $Y \perp X_j | \boldsymbol{X}_{-j}$, and say X_j is non-null otherwise (Li et al., 2005). Let $\mathcal{S} \subseteq \{1, \ldots, p\}$ denote the indices of all the non-null variables, and $\mathcal{S}^\perp \subseteq \{1, \ldots, p\}$ the indices of all the null variables, or equivalently, the complement set of $\mathcal{S}$. Let $|\cdot|$ denote the cardinality, and $\widehat{\mathcal{S}}$ the indices of variables selected by some selection procedure. Our goal is to discover as many non-null variables as possible while controlling the FDR, which is defined as,

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\left| \widehat{\mathcal{S}} \cap \mathcal{S}^\perp \right|}{\left| \widehat{\mathcal{S}} \right| \vee 1} \right].$$

We next establish the identifiability of the problem under the following condition.

Assumption 1 (Irrepresentable Condition in RKHS). *For any* $j \in \{1, \ldots, p\}$, *and any functions* $g_k \in \mathscr{H}_1$, $k \neq j$, $f_j(X_j) \neq \sum_{k=1; k \neq j}^{p} g_k(X_k)$.

This condition simply says that the component function $f_j(X_j)$ in model (1) can not be strictly written as a linear combination of some functions of other variables X_k, $k \neq j$. This is

a fairly mild condition, and its parametric counterpart that $X_j \neq \sum_{k=1;k\neq j}^{p} \beta_k X_k$ for any $\beta_k \in \mathbb{R}$ has been commonly imposed in the linear model scenario (Candès et al., 2018).

Under this condition, we establish the equivalence between variable selection and selection of the component functions f_j in model (1). In other words, testing the hypothesis that X_j is null is the same as testing whether $f_j = 0$.

Proposition 1. *Suppose the nonparametric additive model* (1) *and Assumption 1 hold. Then* $j \in \mathcal{S}^{\perp}$ *if and only if* $f_j = 0$, *for* $j = 1, \ldots, p$.

Proposition 1 makes the variable selection in a nonparametric additive model comparable to that in a linear model, and is to serve as the foundation for the new kernel knockoffs procedure in Section 3, and the finite-sample FDR control in Section 4, both of which are built upon the selection of the component functions f_j's. We next develop the selection procedure for model (1) that is capable of controlling the FDR below any given nominal level $q \in (0, 1)$ under any sample size, while achieving a good power at the same time.

3 Kernel Knockoffs Procedure

3.1 Algorithm

Our kernel knockoffs selection procedure consists of six main steps. Step 1 is to generate the knockoff variables. Step 2 is to subsample without replacement half of the sample observations. Step 3 is to construct the random features for both the original and knockoff variables. Step 4 is to solve the coefficient vector through a group Lasso penalized regression based on the subsamples, which in effect leads to the selection of a set of important variables. In addition, Steps 2 to 4 are carried out repeatedly over a number of subsampling replications. Step 5 is to compute the importance score for each original variable, which is defined as the empirical selection frequency based on multiple subsampling replications. Finally, Step 6 is to apply a knockoff filter to the importance scores to produce the final set of selected variables under the given FDR level, as well as the final estimate of the component functions. We summarize our procedure in Algorithm 1 first, then discuss each step in detail.

3.2 Knockoff variable construction

A random vector $\widetilde{\boldsymbol{X}} \in \mathbb{R}^p$ is said to be a knockoff copy of $\boldsymbol{X} \in \mathbb{R}^p$ (Candès et al., 2018) if

$$(\boldsymbol{X}, \widetilde{\boldsymbol{X}}) \overset{d}{=} (\boldsymbol{X}, \widetilde{\boldsymbol{X}})_{\text{swap}(\mathscr{A})}, \quad \text{for any} \quad \mathscr{A} \subseteq \left\{1, \ldots, p\right\}, \quad \text{and} \quad Y \perp \widetilde{\boldsymbol{X}} | \boldsymbol{X}, \tag{3}$$

where the symbol $\overset{d}{=}$ denotes the equality in distribution, and swap(j) is the operator swapping X_j with $\widetilde{X}_j$. For instance, if $p = 3$ and $\mathscr{A} = \{1, 3\}$, then $\left(X_1, X_2, X_3, \widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3\right)_{\text{swap}(\mathscr{A})}$ becomes $\left(\widetilde{X}_1, X_2, \widetilde{X}_3, X_1, \widetilde{X}_2, X_3\right)$. In the variable selection literature, there are alternative methods that add pseudo-variables to help control the false positives in selection, e.g., by generating independent features, or permuting entries of the existing features (Miller, 2002; Wu et al., 2007). Different from those methods, the knockoff framework has a unique property of exchangeability as given by (3).

There have been numerous ways proposed to construct the knockoff variables. We adopt two particular constructions, depending on the data.

The first is the second-order knockoffs construction (Candès et al., 2018), which generates the knockoffs by matching only the first two moments of the two distributions. In this case, $\widetilde{\boldsymbol{X}}$ is a second-order knockoff copy of $\boldsymbol{X}$ if

$$\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[\widetilde{\boldsymbol{X}}], \quad \text{and} \quad \operatorname{cov}[(\boldsymbol{X}, \widetilde{\boldsymbol{X}})] = \begin{bmatrix} \Sigma & \Sigma - \operatorname{diag}(s) \\ \Sigma - \operatorname{diag}(s) & \Sigma \end{bmatrix},$$

where Σ is the covariance matrix of $\boldsymbol{X}$, and s is a p-dimensional vector such that $\operatorname{cov}[(\boldsymbol{X}, \widetilde{\boldsymbol{X}})]$ is positive semi-definite. To ensure a good statistical power, s should be chosen as large as possible, so that the original and knockoff variables are differentiable (Candès et al., 2018). This strategy is implemented in practice by approximating the distribution of $\boldsymbol{X}$ as the multivariate normal, and is employed in numerous knockoffs-based applications (Barber et al., 2020).

Algorithm 1

Kernel knockoffs selection procedure for nonparametric additive models

Input: Training data $\{(x_i, y_i)\}_{i=1}^n$, the number of random features r, the number of subsampling replications L, and the nominal FDR level $q \in [0, 1]$.

Step 1: Construct the knockoff variables $\{\tilde{x}_i\}_{i=1}^n$ to augment the original variables $\{x_i\}_{i=1}^n$ using the second-order knockoffs or the deep knockoffs machine.

for $\ell = 1$ to L **do**

  Step 2: Draw without replacement to obtain a subsample $I_\ell \subset \{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$.

  Step 3: Sample $2p$ of i.i.d. r-dimensional random features $\{w_v, b_v\}_{v=1}^r$ by (5), and construct the augmented random feature vector $\boldsymbol{\Psi}_{2p}(\boldsymbol{X})$ by (6).

  Step 4: Solve the coefficient vector $\hat{\boldsymbol{c}}_{2p}(I_\ell)$ by (8), and record the selected variables.

end for

Step 5: Compute the importance score by (10), i.e., the empirical selection frequency, $\{\widehat{\Pi}_j\}_{j \in [2p]}$ based on the L estimates of $\{\hat{\boldsymbol{c}}_{2p}(I_\ell)\}_{\ell \in [L]}$.

Step 6: Apply the knockoff filter by (11) at the nominal FDR level q.

Output: the set of selected variables $\widehat{\mathcal{S}}$, and the function estimate $\hat{f}^{\mathrm{RF}}(\boldsymbol{X})$.

The second is the deep knockoffs machine (Romano et al., 2019), which generates the knockoff variables using deep generative models. The key idea is to iteratively refine a knockoff sampling mechanism until a criterion measuring the validity of the produced knockoffs is optimized. This strategy is shown to be able to match higher-order moments, and also achieve a better approximation of exchangeability.

In our construction of knockoff variables, we employ the second-order knockoffs when there is clear evidence that the predictor variables approximately follow a multivariate normal distribution, and employ the deep knockoffs machine otherwise. Given the training

samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, we first augment with the knockoff samples $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^n$, and form the data $\{(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})^\top \in \mathcal{X}^p$, and $\tilde{\boldsymbol{x}}_i = (\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,p})^\top \in \mathbb{R}^p$.

3.3 Random feature mapping

We next construct the random features for both original and knockoff variables. The key idea is to employ the random feature mapping (Rahimi and Recht, 2007; Båzåvan et al., 2012) to approximate the kernel function, which enables us to construct a projection operator between the RKHS and the original predictor space. Specifically, if the kernel functions that generate $\mathscr{H}_1$ are shift-invariant, i.e., $K(X, X') = K(X - X')$, and integrate to one, i.e., $\int_{\mathcal{X}} K(X - X')d(X - X') = 1$, then the Bochner's theorem (Bochner, 1934) states that such kernel functions satisfy the Fourier expansion:

$$K(X - X') = \int_{\mathbb{R}} p(w)\exp\left\{\sqrt{-1}w(X - X')\right\}dw,$$

where $p(w)$ is a probability density defined by

$$p(w) = \int_{\mathcal{X}} K(X)e^{-2\pi\sqrt{-1}wX}dX\,.$$

We note that many kernel functions are shift-invariant and integrate to one. Examples include the Laplacian kernel, $K(X, X') = c_1 e^{-|X-X'|/b_1}$, the Gaussian kernel, $K(X, X') = c_2 e^{-b_2^2|X - X'|^2/2}$, and the Cauchy kernel, $K(X, X') = c_3\left(1 + b_3^2|X - X'|^2\right)^{-1}$, where c_1, c_2, c_3 are the normalization constants, and b_1, b_2, b_3 are the scaling parameters. It is then shown that (Rahimi and Recht, 2007; Båzåvan et al., 2012) the minimizer in (2) can be approximated by,

$$\hat{f}^{\mathrm{RF}}(\boldsymbol{X}) = \boldsymbol{\Psi}_p(\boldsymbol{X})^\top \boldsymbol{c}_p, \tag{4}$$

where $\boldsymbol{\Psi}_p(\boldsymbol{X}) = \left[\boldsymbol{\Psi}(X_1)^\top, \ldots, \boldsymbol{\Psi}(X_p)^\top\right]^\top \in \mathbb{R}^{pr}$, and $\boldsymbol{\Psi}(X_j) = [\psi_1(X_j), \ldots, \psi_r(X_j)]^\top \in \mathbb{R}^r$ is a vector of r Fourier bases with the frequencies drawn from the density $p(w)$, i.e.,

$$\begin{aligned} &\omega_{j,v} \overset{\text{i.i.d.}}{\sim} p(\omega), \quad b_{j,v} \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}[0, 2\pi], \\ &\psi_v(X_j) = \sqrt{\frac{2}{r}}\cos(X_j\omega_{j,v} + b_{j,v}), \quad j = 1, \ldots, p,\ v = 1, \ldots, r\,. \end{aligned} \tag{5}$$

The use of random feature mapping achieves potentially substantially dimension reduction. More specifically, the estimator in (4) only requires to learn the pr-dimensional coefficient $\boldsymbol{c}_p$, compared to the estimator in (2) that involves an infinite-dimensional vector $\tilde{c}_p$. Rudi and Rosasco (2017) showed that the random feature mapping obtains an optimal bias-variance tradeoff if r scales at a certain rate and $r/n \to 0$ when n grows. They further proved that the estimator in (4) can achieve the minimax optimal estimation error. Beyond the estimation optimality, we note that the random feature mapping also efficiently reduces the computational complexity. That is, the computation complexity of the estimator in (4) is

only $O(nr_2)$, compared to the computation complexity of the kernel estimator in (2) that is $O(n^3)$. The saving of the computation is substantial if $r/n \to 0$ as n grows.

In our setting of kernel knockoffs selection, we construct the random features for both original and knockoff variables and obtain the augmented random feature vector as,

$$\boldsymbol{\Psi}_{2p}(\boldsymbol{X}) = \left(\boldsymbol{\Psi}(X_1)^\top, \ldots, \boldsymbol{\Psi}(X_p)^\top, \boldsymbol{\Psi}(\widetilde{X}_1)^\top, \ldots, \boldsymbol{\Psi}(\widetilde{X}_p)^\top\right)^\top \in \mathbb{R}^{2pr}, \tag{6}$$

where $\boldsymbol{\Psi}(X_j) = [\psi_1(X_j), \ldots, \psi_r(X_j)]^\top \in \mathbb{R}^r$, and $\boldsymbol{\Psi}(\widetilde{X}_j) = \left[\psi_1(\widetilde{X}_j), \ldots, \psi_r(\widetilde{X}_j)\right]^\top \in \mathbb{R}^r$, $j = 1, \ldots, p$, are two sets of r-dimensional random features that are independently sampled from (5). Then the minimizer in (2) can be approximated by,

$$\widehat{f}(\boldsymbol{X}) = \Psi_{2p}(\boldsymbol{X})^\top \boldsymbol{c}_{2p}. \tag{7}$$

Meanwhile, we note that the randomness of the features generated from (5) may alter the ranking of variable significances. As such, we couple the random feature mapping with knockoffs and subsampling to achieve the desired FDR control and power.

3.4 Resampling, importance score, and knockoff filtering

We adopt the subsampling scheme similarly as that in Meinshausen and Bühlmann (2010); Dümbgen et al. (2013). Specifically, we subsample a subset of the training samples without replacement with size n_s, and let I denote the corresponding subsample indices out of $\{1, \ldots, n\}$. We set $n_s = [n/2]$, where $[n/2]$ is the largest integer no greater than $n/2$. We then estimate the coefficient vector $\boldsymbol{c}_{2p} = \left(c_1^\top, \ldots, c_{2p}^\top\right)^\top \in \mathbb{R}^{2pr}$ in (7), in which each $c_j \in \mathbb{R}^r$ for $j = 1, \ldots, 2p$, via a group Lasso penalized regression based on the subsample I of the observations,

$$\min_{\substack{c_j \in \mathbb{R}^r \\ j = 1, \ldots, 2p}} \frac{1}{|I|} \sum_{i \in I} \left[y_i - \bar{y}(I) - \sum_{j=1}^{p} \boldsymbol{\Psi}(x_{i,j})^\top c_j - \sum_{j=p+1}^{2p} \boldsymbol{\Psi}(\tilde{x}_{i,j-p})^\top c_j \right]^2 + \tau \sum_{j=1}^{2p} \|c_j\|_2, \tag{8}$$

where $\bar{y}(I) = \sum_{i \in I} y_i/|I|$ is the empirical mean, and $\tau \geqslant 0$ is the penalty parameter. Let $\hat{\boldsymbol{c}}_{2p}(I) = \left(\hat{c}_1^\top(I), \ldots, \hat{c}_{2p}^\top(I)\right)^\top$ denote the minimizer of (8). We remark that the group Lasso penalty in (8) encourages the entire vector $c_j \in \mathbb{R}^r$ to be shrunk to zero, for $j = 1, \ldots, 2p$. Consequently, estimating $\boldsymbol{c}_{2p}$ via (8) in effect leads to the selection of important variables among all $2p$ candidate variables $\left(X_1, \ldots, X_p, \widetilde{X}_1, \ldots, \widetilde{X}_p\right)$. We also remark that, our use of the group Lasso penalty in (8) is different from Huang et al. (2010). Specifically, Huang et al. (2010) used the B-spline basis for nonparametric function approximation, and used the group Lasso twice, first for obtaining an initial estimator and reducing the dimension of the problem, then for selecting the nonzero components. By contrast, we use the random feature mapping for nonparametric function approximation, and apply group Lasso for selection and finite-sample FDR control. These differences have different theoretical implications; for instance, the B-spline basis is usually orthonormal, whereas the random feature mapping is generally not orthogonal. Our group Lasso penalty is also different from the COSSO penalty used in Lin and Zhang (2006), which takes the

form $\sum_{j=1}^{p}\left\|\boldsymbol{\Psi}(x_{i,j})^{\top}c_j\right\|_K+\sum_{j=p+1}^{2p}\left\|\boldsymbol{\Psi}(\tilde{x}_{i,j-p})^{\top}c_j\right\|_K$. Since the random feature mapping generally cannot form an orthogonal basis, there is no closed-form representation of the RKHS norms $\left\|\boldsymbol{\Psi}(x_{i,j})^{\top}c_j\right\|_K$ and $\left\|\boldsymbol{\Psi}(\tilde{x}_{i,j-p})^{\top}c_j\right\|_K$ in our setting. As a result, the COSSO penalty is difficult to implement, and instead we adopt the group Lasso penalty in (8) that also yields the desired theoretical properties.

Given the penalized estimate $\hat{c}_{2p}(I)$, we obtain an estimate of the selected variable indices $\widehat{\mathcal{S}}(I)\subseteq\{1,\ldots,2p\}$. That is, for each $j\in\{1,\ldots,p\}$, $j\in\widehat{\mathcal{S}}(I)$ if $\hat{c}_j(I)\neq\mathbf{0}$ and the original variable X_j is selected, and $(j+p)\in\widehat{\mathcal{S}}(I)$ if $\hat{c}_{j+p}(I)\neq\mathbf{0}$ and the knockoff variable $\tilde{X}_j$ is selected. Then the probability of being in the selected set $\widehat{\mathcal{S}}(I)$ is

$$\widehat{\Pi}_j=\mathbb{P}\left\{j\in\widehat{\mathcal{S}}(I)\right\},\quad\text{for } j=1,\ldots,2p, \tag{9}$$

where $\mathbb{P}$ is with respect to both subsampling I and the random features. We note that $\widehat{\Pi}_j$ can be estimated accurately using the empirical selection frequencies (Meinshausen and Bühlmann, 2010). Specifically, we repeat the above subsampling and coefficient estimation procedure L times, each time for a subsample I_ℓ, $\ell=1,\ldots,L$. We then obtain the selected variable indices $\widehat{\mathcal{S}}(I_\ell)$ for I_ℓ and compute (9) using the empirical selection frequency as the percentage of times the jth variable, $j=1,\ldots,2p$, is included in $\left\{\widehat{\mathcal{S}}(I_\ell)\right\}_{\ell=1}^{L}$.

Next, we define the importance score for the original variable X_j, $j=1,\ldots,p$, as,

$$\Delta_j=\widehat{\Pi}_j-\widehat{\Pi}_{j+p}. \tag{10}$$

We comment that Δ_j in (10) is calculated for only one run of the knockoffs procedure, i.e., we generate the knockoffs only once. This is different from the derandomized knockoffs method recently proposed by Ren et al. (2020), which aggregates the selection results across multiple runs of knockoffs to reduce the randomness of the knockoff generation.

Finally, given the target nominal FDR level q, we apply a knockoff filter (Barber and Candès, 2015) to the importance scores to produce the final set of selected variables,

$$T=\min\left\{t\in\{|\Delta_j|:|\Delta_j|>0\}:\frac{\#\{j:\Delta_j\leqslant -t\}}{\#\{j:\Delta_j\geqslant t\}}\leqslant q\right\}\quad(\text{knockoffs}). \tag{11}$$

Set $T=\infty$ if the above set is empty. Another commonly used but slightly more conservative knockoff filter (Barber and Candès, 2015; Candès et al., 2018) is,

$$T_+=\min\left\{t\in\{|\Delta_j|:|\Delta_j|>0\}:\frac{\#\{j:\Delta_j\leqslant -t\}+1}{\#\{j:\Delta_j\geqslant t\}}\leqslant q\right\}\quad(\text{knockoffs+}). \tag{12}$$

In our simulations, we have experimented with both filers, which produce very similar results, so we only present the results based on T.

Given the threshold value T, the final set of selected variables is,

$$\widehat{\mathcal{S}} = \{j \in \{1, \ldots, p\} : \Delta_j \geqslant T\}. \tag{13}$$

We then reestimate $\boldsymbol{c}_p$ in (4) using all the sample observations as,

$$\hat{\boldsymbol{c}}_p^{\mathrm{RF}} = \left((\hat{c}_1^{\mathrm{RF}})^\top, \ldots, (\hat{c}_p^{\mathrm{RF}})^\top\right)^\top = \underset{c_j \in \mathbb{R}^r,\, j \in \widehat{\mathcal{S}}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{1}{n}\sum_{i=1}^{n} y_i - \sum_{j \in \widehat{\mathcal{S}}} \boldsymbol{\Psi}(x_{i,j})^\top c_j\right]^2.$$

We obtain the final knockoffs-based kernel regression estimator as,

$$\hat{f}^{\mathrm{RF}}(\boldsymbol{X}) = \boldsymbol{\Psi}_p(\boldsymbol{X})^\top \hat{\boldsymbol{c}}_p^{\mathrm{RF}}. \tag{14}$$

3.5 Parameter tuning

We next discuss the parameter tuning. We further carry out a sensitivity analysis in Section S2.2, and a parallelization experiment in Section S2.5 of the Appendix.

For the number of random features r, we start with an initial set Ξ of candidate values for r. For each working rank $r \in \Xi$, we calculate the selection frequencies $\left\{\widehat{\Pi}_{j,r}\right\}_{j \in [2p], r \in \Xi}$, and the standard deviation $\hat{\sigma}_r = \mathrm{sd}\left(\left\{\widehat{\Pi}_{j,r}\right\}_{j \in [2p]}\right)$, with a relatively small number for the subsampling replications. We then choose the value of $r \in \Xi$ that maximizes the following criterion that balances the selection standard deviation and model complexity,

$$\hat{r} = \underset{r \in \Xi}{\mathrm{argmax}} 2p\hat{\sigma}_r - \ln(r).$$

We have observed through our numerical simulations that, when we start from a small value of r, the selection frequencies of both original variables and their knockoffs counterparts are close to zero. As r increases, it starts to separate the truly important variables from the null variables and knockoffs, where the selection frequencies of those truly important variables grow positively, and correspondingly, the standard deviation $\hat{\sigma}_r$ increases. Meanwhile, the log penalty term helps balance the model complexity.

For the regularization parameter $\boldsymbol{\tau}$ in (8), we choose it by minimizing the BIC criterion,

$$\hat{\tau} = \underset{\tau \geqslant 0}{\mathrm{argmin}} \log[\mathrm{RSS}(\tau)] + r\frac{\log n}{n}\left|\widehat{\mathcal{S}}(\tau)\right|,$$

where RSS($\boldsymbol{\tau}$) is the cross-validation residual sum of squares, and $\left|\widehat{\mathcal{S}}(\tau)\right|$ is the cardinality.

For the number of subsampling replications L, our numerical experiments have found that L = 100 results in a competitive performance in FDR control and power. For the subsampling sample size n_s, we have found that, when n_s is no smaller than $\lfloor n/2 \rfloor$, the method performs well. We also comment that, the computation of our method can be easily parallelized, since it requires no information sharing across different subsamples.

4 Theoretical Guarantees

4.1 FDR Control

We show that our proposed procedure controls the FDR under any given nominal level and any given sample size. Due to the intrinsic difficulty of the nonparametric additive model, we employ the random feature mapping to approximate the component function f_j, and construct a projection operator between the RKHS and the original predictor space. By construction, the random features of the knockoff variables have a similar structure mimicking the random features of the original variables, even though the knockoffs are not associated with the response conditioning on the original ones. Since the random features may introduce additional stochastic errors, which can disturb the order of the variables entering the model, we further employ the subsampling strategy to improve the selection stability. Finally, we compute the importance score for each variable based on the projected component function f_j, and select the variables that have considerably higher scores than their knockoff counterparts. Intuitively, such a procedure enjoys the finite-sample FDR control, similarly as the existing knockoff solutions (Barber and Candès, 2015; Candès et al., 2018).

We first show that the importance score Δ_j in (10) has a symmetric distribution for a null variable $X_j \in \mathcal{S}^{\perp}$, and is equally likely to be positive or negative. The symmetric property of the null variables is crucial for the knockoffs procedure, which then chooses a data-dependent threshold while having the FDR under control (Barber and Candes, 2015).

Theorem 1. *Suppose Assumption 1 holds. Let* $(s_1, \ldots, s_p)$ *be a set of independent random variables, such that* $s_j = \pm 1$ *with probability* $1/2$ *if* $j \in \mathcal{S}^{\perp}$, *and* $S_j = 1$ *if* $j \in \mathcal{S}$. *Then,*

$$(\Delta_1, \ldots, \Delta_p) \overset{d}{=} (\Delta_1 \cdot s_1, \ldots, \Delta_p \cdot s_p).$$

Next, we show that our selection procedure successfully controls the false discovery under any sample size. The result holds regardless of the distribution or the number of predictors, and does not require any knowledge of the noise level. The false discovery here is measured by both the FDR, and the modified FDR, which is defined as,

$$\mathrm{mFDR} = \mathbb{E}\left[\frac{\left|\widehat{\mathcal{S}} \cap \mathcal{S}^{\perp}\right|}{\left|\widehat{\mathcal{S}}\right| + 1/q}\right].$$

The definition of mFDR follows the knockoffs literature (Barber and Candès, 2015), with 1 replaced by $1/q$ in the denominator compared to FDR. Meanwhile, it is close to FDR in the setting when there are a large number of variables selected, i.e., when $\left|\widehat{\mathcal{S}}\right|$ is large, and it is less conservative than FDR, in that mFDR is always under control if FDR is.

Theorem 2. *For any* $q \in [0, 1]$ *and any sample size n, the selected set* $\widehat{\mathcal{S}}$ *in* (13) *based on the knockoff filter T in* (11) *satisfies that mFDR* $\leqslant q$. *Meanwhile, the selected set* $\widehat{\mathcal{S}}$ *based on the knockoff filter* T_+ *in* (12) *satisfies that FDR* $\leqslant q$.

We remark that, Theorem 2 achieves the valid FDR control with no restriction on the dimension p relative to the sample size n. As such, the proposed method works for both settings of $p < n$ and $p > n$. In particular, the FDR control under $p > n$ is achieved by building upon the model-X knockoffs (Candès et al., 2018), which treats the variables as random and utilizes the stochasticity of the random variables. This is different from the original knockoff solution (Barber and Candès, 2015), which treats the variables as fixed and relies on specific stochastic properties of the linear model, and thus excludes the setting of $p > n$ or nonlinear models.

4.2 Power Analysis

Next, we show that our proposed kernel knockoffs selection procedure achieves a power that approaches one as the sample size tends to infinity. We first note that, the theoretical power analysis for the knockoff methods is largely missing in the current literature, with a few exceptions such as Fan et al. (2020) and Weinstein et al. (2020). Fan et al. (2020) studied the power for linear regressions under the model-X knockoff framework. Weinstein et al. (2020) studied the power of knockoffs with thresholded Lasso for linear models. By contrast, we study the power for nonparametric models. We also remark that, as is common for all knockoffs selection methods, the power of our knockoffs-based method is usually no greater than that of the group Lasso-based selection. This is because the proposed knockoffs procedure is built on top of the group Lasso selection in (8). In a sense, the knockoffs procedure further selects variables from the set of variables that are identified by group Lasso for the augmented predictors. Therefore, the key of our power analysis is to investigate how much power loss that the knockoffs procedure would induce. We introduce some regularity conditions.

Assumption 2. *The number of nonzero component functions, i.e.,* $|\mathcal{S}|$, *is bounded.*

Assumption 3. *Suppose there exists a constant* $C_{\min} > 0$, *such that the minimal eigenvalue of matrix* $\mathbb{E}\left[n^{-1}\Sigma_{\mathcal{S}}^{\top}\Sigma_{\mathcal{S}}\right]$ *satisfies that,*

$$\Lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}\Sigma_{\mathcal{S}}^{\top}\Sigma_{\mathcal{S}}\right]\right) \geq \frac{1}{2}C_{\min},$$

where the expectation is taken over the random features and $\Sigma_{\mathcal{S}} \in \mathbb{R}^{n \times 2r|\mathcal{S}|}$ *is the design matrix with the ith row equal to* $\left[\boldsymbol{\Psi}(x_{i,j_1})^{\top}, \ldots, \boldsymbol{\Psi}\left(x_{i,j_{|\mathcal{S}|}}\right)^{\top}, \boldsymbol{\Psi}(\tilde{x}_{i,j_1})^{\top}, \ldots, \boldsymbol{\Psi}\left(\tilde{x}_{i,j_{|\mathcal{S}|}}\right)^{\top}\right]$, $i = 1, \ldots, n$, $\mathcal{S} = \{j_1, \ldots, j_{|\mathcal{S}|}\}$.

Assumption 4. *Suppose* $p < e^n$. *Let* $\eta_R \equiv c_\eta\left\{n^{-\beta/(2\beta+1)} + [(\log p)/n]^{1/2}\right\}$ *for some constant* $c_\eta > 0$. *Suppose* $\min_{j \in \mathcal{S}} \|f_j(X_j)\|_{L_2(X_j)} \geq \kappa_n \eta_R$, *for some slowly diverging sequence* $\kappa_n \to \infty$, *as* $n \to \infty$, *where the RKHS* $\mathcal{H}_1$ *is embedded to a* β*th order Sobolev space with* $\beta > 1$.

Assumption 5. *Suppose there exists a constant* $0 \leqslant \xi_\Sigma < 1$ *such that,*

$$\max_{j \notin \mathcal{S}} \left\| \{\Sigma_j(I)\}^\top \Sigma_{\mathcal{S}}(I) \left[\Sigma_{\mathcal{S}}(I)^\top \Sigma_{\mathcal{S}}(I) \right]^{-1} \right\|_2 \leqslant \xi_\Sigma, \quad \textit{and} \quad \frac{\xi_\Sigma \sqrt{|\mathcal{S}|} + 1}{\tau} \eta_R + \xi_\Sigma \sqrt{|\mathcal{S}|} < 1.$$

All these conditions are reasonable and are commonly imposed in the literature. Specifically, Assumption 2 concerns the overall complexity in that it upper bounds the total number of nonzero component functions. Similar conditions have been commonly adopted in sparse additive models over RKHS (e.g., Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2016; Dai and Li, 2021). Moreover, we carry out a numerical experiment in Section S2.4 of the Appendix, and show empirically that our method still works reasonably well when the number of nonzero components $|\mathcal{S}|$ increases along with the sample size. We speculate that it is possible to allow $|\mathcal{S}|$ to diverge, but leave the full theoretical investigation as future research. Assumption 3 ensures the identifiability among the $|\mathcal{S}|$ submatrices of $\Sigma_{\mathcal{S}}$. The same condition has been used in Zhao and Yu (2006); Ravikumar et al. (2010). Assumption 4 imposes some regularity on the minimum regulatory effect. Similar conditions have been used in Lasso regressions (Ravikumar et al., 2010; Raskutti et al., 2012; Fan et al., 2020). In addition, similar to the treatment of high-dimensional linear and additive models (Raskutti et al., 2011; Yuan and Zhou, 2016), we assume that $p < e^n$ to ensure nontrivial probabilistic bounds. In other words, we allow the dimension p to diverge at the exponential order of the sample size n. Assumption 5 reflects the intuition that the large number of irrelevant variables cannot exert an overly strong effect on the relevant variables. Besides, the second inequality characterizes the relationship between ξ_Σ, the sparse tuning parameter τ, and the sparsity level $|\mathcal{S}|$. This condition is again standard for Lasso regressions (Zhao and Yu, 2006; Ravikumar et al., 2010).

Next, we characterize the statistical power of the proposed kernel knockoffs procedure. For the true set $\mathcal{S}$ and the selected set $\widehat{\mathcal{S}}$, the power is defined as

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

Theorem 3. *Suppose Assumptions 1–5 hold, and the number of random features* $r \geqslant c_r n^{2\beta/(2\beta+1)}$ *for some* $c_r > 0$. *Then, the selected set* $\widehat{\mathcal{S}}$ *in* (13) *satisfies that,* $\textit{power}\left(\widehat{\mathcal{S}}\right) \to 1$, *as* $n \to \infty$.

We again remark that, Theorem 3 holds for both settings of $p < n$ and $p > n$, or more specifically, $n < p < e^n$, which is implied by Assumption 4. The power property under $p > n$ is achieved by integrating Rademacher processes and the concentration inequalities for empirical processes (van de Geer, 2002; Yuan and Zhou, 2016) with the deviation conditions for nonparametric regressions (Loh and Wainwright, 2012; Dai and Li, 2021).

Together, Theorems 2 and 3 show that our proposed selection method is able to achieve both the finite-sample FDR control and the asymptotic power that approaches one.

5 Numerical Studies

We carry out intensive simulations to examine the empirical performance of our proposed method under the varying signal strength, the predictor distribution, the nonparametric component function, the sample size and the number of predictors. We compare with several alternative solutions. We also illustrate our method with an analysis of brain imaging data for Alzheimer's disease. We report additional simulation results in Section S2 of the Appendix.

5.1 Alternative methods for comparison

We abbreviate our proposed kernel knockoffs selection method as KKO. We solve the group Lasso penalized problem in (8) using the R package `grpreg`. We employ the Laplacian kernel with $r \in \Xi = \{2, 3, 4\}$, and tune the hyperparameters following Section 3.5. We set the target FDR level at $q = 0.2$ following Fan et al. (2020). We also briefly comment that, in addition to the reproducing kernel approach, the spline basis expansion is another commonly used approach in the nonparametric additive modeling. But it involves a totally different set of methodological tools and theoretical analysis, and we leave it as future research.

We compare our method with three main competitors. The first competitor is the nonparametric selection method for sparse additive models (SPAM) of Ravikumar et al. (2009), which combines B-spline basis expansion with grouped Lasso. We set the number of B-spline expansions at $[n^{1/5}]$, i.e., the largest integer no greater than $n^{1/5}$, and tune the sparsity penalty by generalized cross-validation. We implement the method using the R package `SAM`. We did not compare with the COSSO method of Lin and Zhang (2006), due to that the code is not available, and SPAM usually achieves a similar and sometimes more competitive performance than COSSO. The second competitor is the linear knockoffs (LKO) selection method, and we implement it using the R package `knockoffs`. The third competitor is the graphical nonlinear knockoffs (RANK) selection method of Fan et al. (2020). We follow the same parameter setup as in Fan et al. (2020), and implement their method based on the R package `gamsel`.

We also compare our method with that of Dai and Barber (2016). More specifically, Dai and Barber (2016) adopted an "expansion first" strategy, which first performs feature expansion $\widetilde{\boldsymbol{\Psi}}(X_j)$, $j = 1, \ldots, p$, then constructs the knockoffs based on the expanded features. By contrast, we adopt the "knockoffs first" approach, which constructs the knockoffs directly for the variables $\{X_j\}_{j=1}^p$, then performs the random feature expansion on both original variables and their knockoffs. There are two advantages of doing the knockoffs first. First, it ensures a better knockoffs construction and eventually a better statistical power. That is, to construct a good knockoff variable using either the second-order knockoffs or the deep knockoffs, it requires a reasonably slow eigenvalue decay of the covariance Σ, so that the original variables and their knockoffs are differentiable (Candès et al., 2018). We consider a simulation example replicated 200 times, where the predictors follow a multivariate normal distribution with the zero mean and the identity covariance matrix, $n = 500$, $p = 5$, and we employ the Laplacian kernel with $r = 3$. Figure 1(a) shows the eigenvalue decay of the sample covariance matrix (blue line), and the random kernel expansion (red line). It is seen

that the former decays more slowly than the latter, and therefore it is better to construct the knockoffs based on the original variables. Second, the “expansion first” leads to larger correlations between the original variables and their knockoffs, compared to the “knockoffs first”, as shown in Figure 1(b), which would in turn make the subsequent group Lasso selection harder. Finally, the “expansion first” is computationally more expensive. This is because the “expansion first” approach requires generating the knockoffs for pr-dimensional variables, whereas our “knockoffs first” approach only requires generating the knockoffs for p-dimensional variables.

5.2 Varying signal strength, predictor design, and component functions

We first study the performance with the varying signal strength and the predictor distribution.

We simulate the response, $Y = \sum_{j \in \mathcal{S}} \theta_j f_j(X_j) + \epsilon$, where $\mathcal{S}$ is the set of relevant predictors with $|\mathcal{S}| = 10$, and ϵ is a standard normal error. We sample θ_j independently from a uniform distribution $(-\theta, \theta)$ for some positive constant θ. The magnitude of θ reflects the strength of the signal, and we vary $\theta = \{0.1, 1, 10, 100, 200\}$. We simulate the predictors independently from three different distributions, a multivariate normal distribution with mean zero and covariance $\Sigma_{ij} = 0.3^{|i-j|}$, a mixture normal distribution, with an equal probability from three multivariate normal distributions, all with mean zero, and different covariances where $\Sigma_{1,ij} = 0.1^{|i-j|}$, $\Sigma_{2,ij} = 0.3^{|i-j|}$, and $\Sigma_{3,ij} = 0.5^{|i-j|}$, and a uniform $[-2, 2]$ distribution. We employ the second-order knockoffs when the predictor distribution is normal, and the deep knockoffs otherwise, to generate the knockoff variables. We first consider a trigonometric polynomial component function, and fix the number of predictors at $p = 50$, and the sample size at $n = 900$. We later consider other forms of component functions, and different (p, n).

$$f_j(x) = u_{j,1}\sin(c_{j,1}x) + u_{j,2}\cos(c_{j,2}x) + u_{j,3}\sin^2(c_{j,3}x) + u_{j,4}\cos^2(c_{j,4}x), \tag{15}$$

where $u_{j,k}$ follows a uniform (1, 2) distribution, and $c_{j,k}$ follows a uniform (1, 10), for $k = 1, 2, 3, 4$. Figure 2 reports the FDR and power over 200 data replications for the four methods with the varying signal strength θ and three different predictor distributions. It is seen that our method successfully controls the FDR blow the expected level, and at the same time achieves the best power. Besides, the performance is robust with respect to different predictor distributions. By comparison, the alternative methods are much more sensitive in terms of the FDR control, and the powers are consistently lower. Moreover, the linear knockoffs method often fails to control the FDR.

Next, we consider more forms of component functions. The first is a sin-ratio function,

$$f_j(x) = \frac{\sin(c_{j,1}x)}{2 - \sin(c_{j,2}x)}, \tag{16}$$

where $c_{j,k}$ follows a uniform (1, 10) distribution for $k = 1, 2$, and $|S| = 10$. We note that it is generally more difficult to estimate the sin-ratio function (16) compared to the trigonometric polynomial function (15). The second is a mixed additive model, where we sample the component function with an equal probability from (15) or (16). We fix the

predictor distribution as the multivariate normal, $p = 50$ and $n = 900$. We continue to vary the signal strength $\theta = \{0.1, 1, 10, 100, 200\}$. Figure 3 reports the FDR and power based on 200 data replications. It is seen again that our method achieves the best power while controlling the FDR under the nominal level for the new component functions.

5.3 Varying sample size and dimension

Next, we investigate the empirical performance with the varying n and p.

For the varying n, we consider the trigonometric polynomial function (15) with $p = 50$, $|\mathcal{S}| = 10$, $\theta = 100$, and the multivariate normal predictor distribution with mean zero and covariance $\Sigma_{ij} = 0.3^{|i-j|}$. We vary the sample size $n = \{400, 900, 1500, 2500\}$. Figure 4 reports the FDR and power based on 200 data replications. It is seen that our method successfully control the FDR at all sample sizes, while its power quickly increases as n increases, and dominates the powers of all the competitive methods considerably. Besides, both LKO and RANK have an inflated FDR especially when the sample size is small.

For the varying p, we consider the trigonometric polynomial function (15) with $n = 900$, $|\mathcal{S}| = 10$, $\theta = 100$, and the multivariate normal predictor distribution with mean zero and covariance $\Sigma_{ij} = 0.3^{|i-j|}$. We vary the number of predictors p = {30, 50, 150, 500, 1500, 2000}. As such, we have considered both cases that $p < n$ and $p > n$. Recall that the proposed method can handle both the low-dimensional and high-dimensional regimes, and the theoretical guarantees in Section 4 are established for both $p < n$ and $p > n$. Moreover, for the $p > n$ case, we construct the knockoff variables for all the original variables; i.e., we construct the p-dimensional knockoffs $\widetilde{X}$ for the p-dimensional $\boldsymbol{X}$. This follows a similar strategy as in Fan et al. (2020), but is different from Barber and Candes (2019), where a precedent step of feature screening is carried out first, and the knockoff variables are constructed only for those selected variables. Our approach avoids to require the sure screening property; see also the discussion in Fan et al. (2020). Figure 5 reports the FDR and power based on 200 data replications. It is seen that our method again achieves the best performance in terms of FDR and power in both regimes. By contrast, LKO and RANK have a low power and inflated FDR. Although SPAM sometimes obtains a power similar to ours, its FDR is much inflated and is far above the nominal level.

5.4 Brain imaging data analysis

We illustrate the proposed method with a brain imaging data analysis to study the Alzheimer's disease (AD). AD is an irreversible neurodegenerative disorder, and is characterized by progressive impairment of cognitive and memory functions, loss of bodily functions, and ultimately death. It is the leading form of dementia in elderly subjects, and is the sixth leading cause of death in the United States. Over 5.5 million Americans were affected by AD in 2018, and without any effective treatment or prevention, this number is projected to triple by 2050 (Alzheimer's Association, 2020). Brain atrophy as reflected by brain grey matter cortical thickness is a well-known biomarker for AD. We study a dataset with $n = 697$ subjects, each of whom received an anatomical magnetic resonance imaging (MRI) scan that measures cortical thickness. The data is publicly available at http://adni.loni.usc.edu/. The MRI image has been preprocessed by the standard pipeline,

and is summarized in the form of a vector of cortical thickness measurements for a set of parcellated brain regions-of-interest. There are $p = 68$ regions in total. Brain parcellation is particularly useful to facilitate the interpretation, and has been frequently employed in brain imaging analysis (Fornito et al., 2013; Kang et al., 2016). In addition to the MRI image, for each subject, the data also records a composite cognitive score, which combines numerous tests that assess episodic memory, timed executive function, and global cognition (Donohue et al., 2014). Our goal is to study the association between the composite cognitive score and the vector of brain cortical thickness, and identify individual brain regions with strong associations. A linear model is inadequate to capture such an association, and we turn to the nonparametric additive model instead. We apply the proposed kernel knockoffs selection procedure. As the distribution of the predictors is not necessarily normal, we employ the deep knockoffs machine to generate the knockoff variables. We continue to set the target FDR level at $q = 0.2$.

Table 1 reports the ten brain regions selected by our method. These findings agree with and support the current literature on AD research. Particularly, the middle temporal gyrus is located on the temporal lobe, and is associated with processes of recognition of known faces and accessing word meaning while reading. Middle temporal lobe atrophy is common in AD as well as its prodromal stage, mild cognitive impairment (Visser et al., 2002). The superior parietal lobe is involved with attention, visuospatial perception, and spatial orientation. Damage to the parietal lobe is common in AD, and leads to problems with performing gestures and skilled movements (Pini et al., 2016). The fusiform is linked with various neural pathways related to recognition. The inferior parietal lobe is involved in perception of emotions. The superior temporal gyrus is involved in auditory processing, and has also been implicated as a critical structure in social cognition. Numerous studies have found involvement of these brain regions in the development of AD (Convit et al., 2000; Du et al., 2007; Pini et al., 2016). The precuneus is associated with episodic memory, visuospatial processing, reflections upon self, and aspects of consciousness, and is found to be an AD-signature region (Bakkour et al., 2013). Finally, the entorhinal cortex functions as a hub in a widespread network for memory, navigation and the perception of time. It is found implicated in the early stages of AD, and is one of the most heavily damaged cortices in AD (van Hoesen et al., 1991).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Editor, the Associate Editor, and three referees for their constructive comments and insightful suggestions. Li's research was partially supported by the NSF grant CIF-2102227, and the NIH grants R01AG061303, R01AG062542, and R01AG034570.

References

Ahmed I, Hartikainen A-L, Järvelin M-R and Richardson S (2011). False discovery rate estimation for stability selection: application to genome-wide association studies. Statistical Applications in Genetics and Molecular Biology 10.

Alzheimer's Association (2020). 2020 Alzheimer's disease facts and figures. Alzheimer's & Dementia 16 391–460.

Aronszajn N (1950). Theory of reproducing kernels. Transactions of the American Mathematical Society 68 337–404.

Bach FR (2008). Bolasso: model consistent lasso estimation through the bootstrap. In Proceedings of the 25th International Conference on Machine Learning.

Bakkour A, Morris JC, Wolk DA and Dickerson BC (2013). The effects of aging and alzheimer's disease on cerebral cortical anatomy: Specificity and differential relationships with cognition. NeuroImage 76 332–344. [PubMed: 23507382]

Barber RF and Candès EJ (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics 43 2055–2085.

Barber RF and Candès EJ (2019). A knockoff filter for high-dimensional selective inference. The Annals of Statistics 47 2504–2537.

Barber RF, Candès EJ and Samworth RJ (2020). Robust inference with knockoffs. The Annals of Statistics 48 1409–1431.

Båzåvan EG, Li F and Sminchisescu C (2012). Fourier Kernel Learning. New York: Springer.

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) 57 289–300.

Bochner S (1934). A theorem on fourier-stieltjes integrals. Bulletin of the American Mathematical Society 40 271–276.

Cai TT and Sun W (2017). Large-scale global and simultaneous inference: Estimation and testing in very high dimensions. Annual Review of Economics 9 411–439.

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 80 551–577.

Convit A, DE Asis J et al. (2000). Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to alzheimer's disease. Neurobiology of Aging 21 19–26. [PubMed: 10794844]

Dai R and Barber R (2016). The knockoff filter for fdr control in group-sparse and multitask regression. In International Conference on Machine Learning. PMLR.

Dai X and Li L (2021). Kernel ordinary differential equations. Journal of the American Statistical Association 1–35. [PubMed: 35757777]

Donohue MC, Sperling RA and Others (2014). The Preclinical Alzheimer Cognitive Composite: Measuring Amyloid-Related Decline. Journal of the American Medical Association: Neurology 71 961–970. [PubMed: 24886908]

Du A-T, Schuff N, Kramer JH, Rosen HJ, Gorno-Tempini ML, Rankin K, Miller BL and Weiner MW (2007). Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130 1159–1166. [PubMed: 17353226]

Dümbgen L, Samworth RJ, Schuhmacher D et al. (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 Institute of Mathematical Statistics, 78–90.

Efron B, Tibshirani R, Storey JD and Tusher V (2001). Empirical bayes analysis of a microarray experiment. Journal of the American Statistical Association 96 1151–1160.

Fan Y, Demirkaya E, Li G and Lv J (2020). Rank: large-scale inference with graphical nonlinear knockoffs. Journal of the American Statistical Association 115 362–379. [PubMed: 32742045]

Fornito A, Zalesky A and Breakspear M (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. NeuroImage 80 426–444. [PubMed: 23643999]

Hastie TJ and Tibshirani RJ (1990). Generalized Additive Models, vol. 43. London: Chapman and Hall.

He K, Li Y, Zhu J, Liu H, Lee JE, Amos CI, Hyslop T, Jin J, Lin H and Wei Q (2016). Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. Bioinformatics 32 50–57. [PubMed: 26382192]

Huang J, Horowitz JL and Wei F (2010). Variable selection in nonparametric additive models. Annals of Statistics 38 2282. [PubMed: 21127739]

Kang J, Bowman FD, Mayberg H and Liu H (2016). A depression network of functionally connected regions discovered via multi-attribute canonical correlation graphs. NeuroImage 141 431–441. [PubMed: 27474522]

Koltchinskii V and Yuan M (2010). Sparsity in multiple kernel learning. The Annals of Statistics 38 3660–3695.

Li L, Cook RD and Nachtsheim CJ (2005). Model-free variable selection. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 67 285–299.

Li S, Hsu L, Peng J and Wang P (2013). Bootstrap inference for network construction with an application to a breast cancer microarray study. The Annals of Applied Statistics 7 391. [PubMed: 24563684]

Lin Y and Zhang HH (2006). Component selection and smoothing in multivariate nonparametric regression. The Annals of Statistics 34 2272–2297.

Loh P-L and Wainwright MJ (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. Annals of Statistics 40 1637–1664.

Meier L, Van de Geer S and Bühlmann P (2009). High-dimensional additive modeling. The Annals of Statistics 37 3779–3821.

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

Mercer J (1909). Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society, London A 415–446.

Miller A (2002). Subset Selection in Regression. CRC Press.

Pini L, Pievani M, Bocchetta M, Altomare D, Bosco P, Cavedo E, Galluzzi S, Marizzoni M and Frisoni GB (2016). Brain atrophy in alzheimer's disease and aging. Aging Research Reviews 30 25–48. Brain Imaging and Aging. [PubMed: 26827786]

Rahimi A and Recht B (2007). Random features for large-scale kernel machines. Advances in Neural Information Processing systems 20 1177–1184.

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.

Raskutti G, Wainwright MJ and Yu B (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Transactions on Information Theory 57 6976–6994.

Ravikumar P, Lafferty J, Liu H and Wasserman L (2009). Sparse additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 1009–1030.

Ravikumar P, Wainwright MJ and Lafferty JD (2010). High-dimensional ising model selection using l_1-regularized logistic regression. The Annals of Statistics 38 1287–1319.

Ren Z, Wei Y and Candès E (2020). Derandomizing knockoffs. arXiv preprint arXiv:2012.02717.

Romano Y, Sesia M and Candès E (2019). Deep knockoffs. Journal of the American Statistical Association 1–12. [PubMed: 34012183]

Rudi A and Rosasco L (2017). Generalization properties of learning with random features. In NIPS.

Stone CJ (1985). Additive regression and other nonparametric models. The Annals of Statistics 689–705.

Storey JD (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 347–368.

Su W, Bogdan M and Candes E (2017). False discoveries occur early on the lasso path. The Annals of statistics 2133–2150.

Sun W and Cai TT (2007). Oracle and adaptive compound decision rules for false discovery rate control. Journal of the American Statistical Association 102 901–912.

Sun W and Cai TT (2009). Large-scale multiple testing under dependence. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 393–424.

Van de Geer SA (2002). On hoeffding's inequality for dependent random variables. In Empirical process techniques for dependent data. Springer, 161–169.

Van Hoesen GW, Hyman BT and Damasio AR (1991). Entorhinal cortex pathology in alzheimer's disease. Hippocampus 1 1–8. [PubMed: 1669339]

Visser PJ, Verhey FRJ, Hofman PAM, Scheltens P and Jolles J (2002). Medial temporal lobe atrophy predicts alzheimer's disease in patients with minor cognitive impairment. Journal of Neurology, Neurosurgery & Psychiatry 72 491–497. [PubMed: 11909909]

Wahba G (1990). Spline Models for Observational Data. SIAM.

Wahba G, Wang Y, Gu C, Klein R and Klein B (1995). Smoothing spline anova for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. The Annals of Statistics 23 1865–1895.

Weinstein A, Su WJ, Bogdan M, Barber RF and Candès EJ (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. arXiv preprint arXiv:2007.15346.

Wood SN (2017). Generalized Additive Models: An Introduction With R. CRC press.

Wu Y, Boos DD and Stefanski LA (2007). Controlling variable selection by the addition of pseudovariables. Journal of the American Statistical Association 102 235–243.

Yuan M and Zhou D-X (2016). Minimax optimal rates of estimation in high dimensional additive models. Annals of Statistics 44 2564–2593.

Zhao P and Yu B (2006). On model selection consistency of lasso. Journal of Machine Learning Research 7 2541–2563.

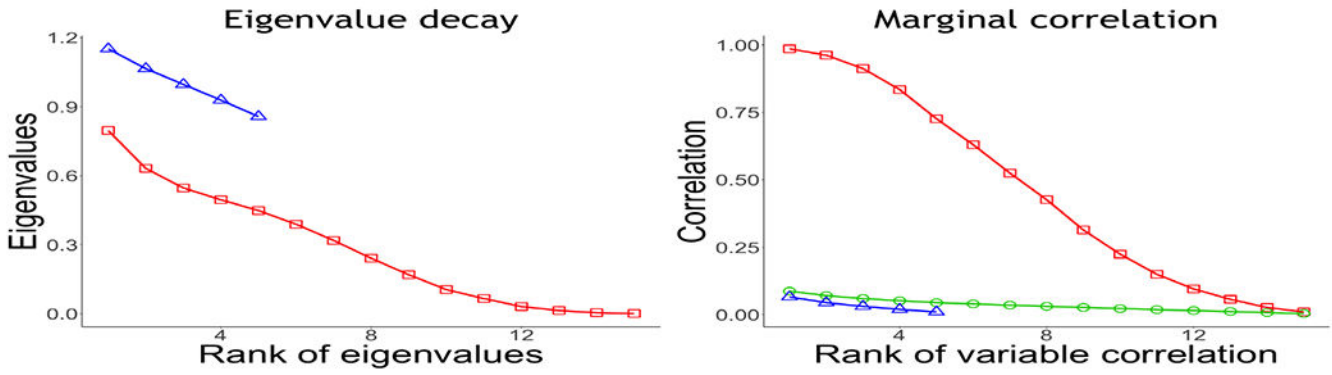

Figure 1:

Comparison between the “expansion first” strategy and the “knockoffs first” strategy. Left panel: the eigenvalue decay of the sample covariance matrix (blue line) and the random kernel expansion (red line). Right panel: the marginal sample correlations between the original variables and the knockoff variables (blue line), between the random kernel expansion and the “expansion first” knockoff variables (red line), and between the random kernel expansion and the “knockoffs first” knockoff variables (green line).

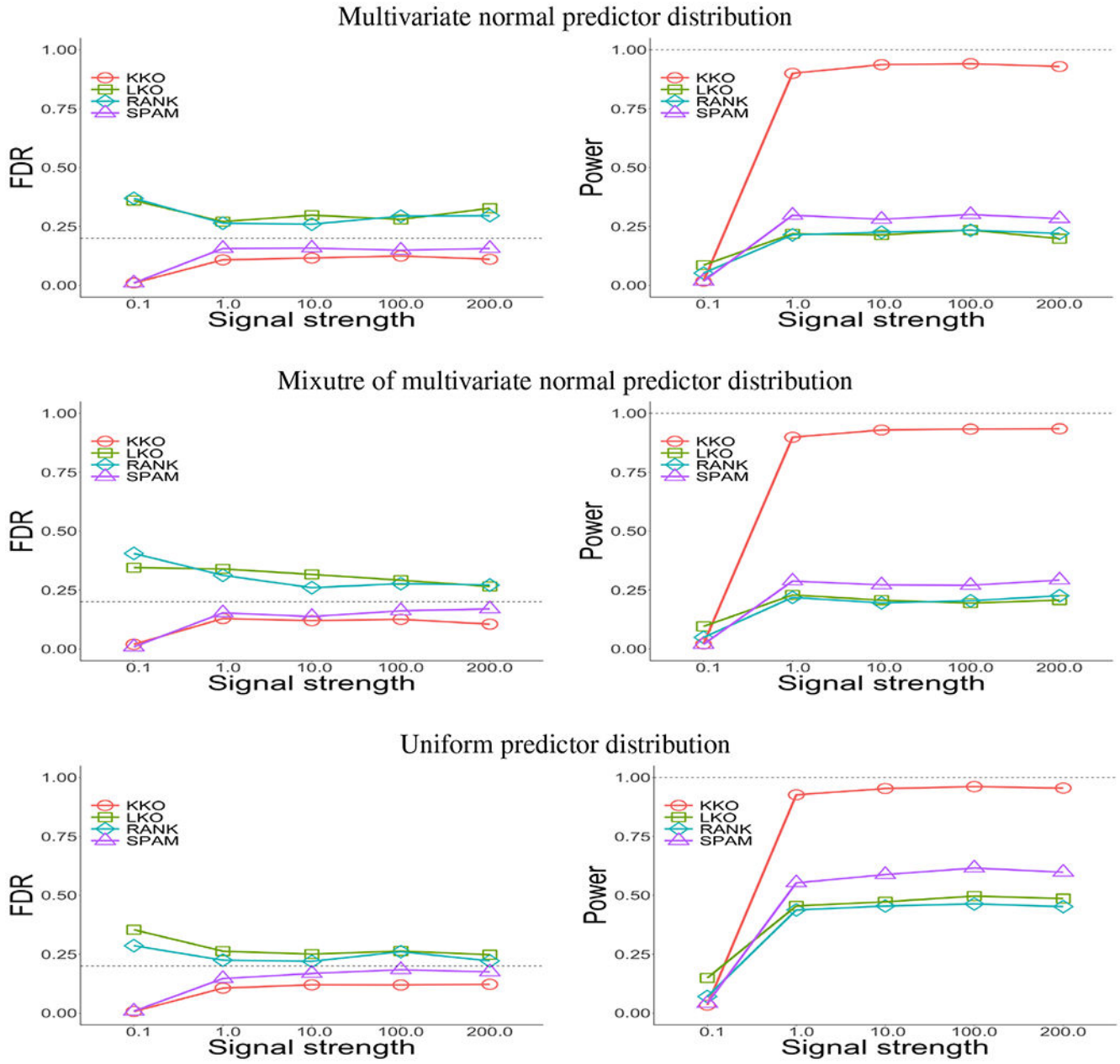

Figure 2:

Empirical performance and comparison in terms of FDR and power with the varying signal strength and predictor distribution. Four methods are compared: the nonparametric selection method for sparse additive models (SPAM) of Ravikumar et al. (2009), the linear knockoffs (LKO) of Barber and Candès (2015), the graphical nonlinear knockoffs (RANK) of Fan et al. (2020), and our proposed kernel knockoffs (KKO).

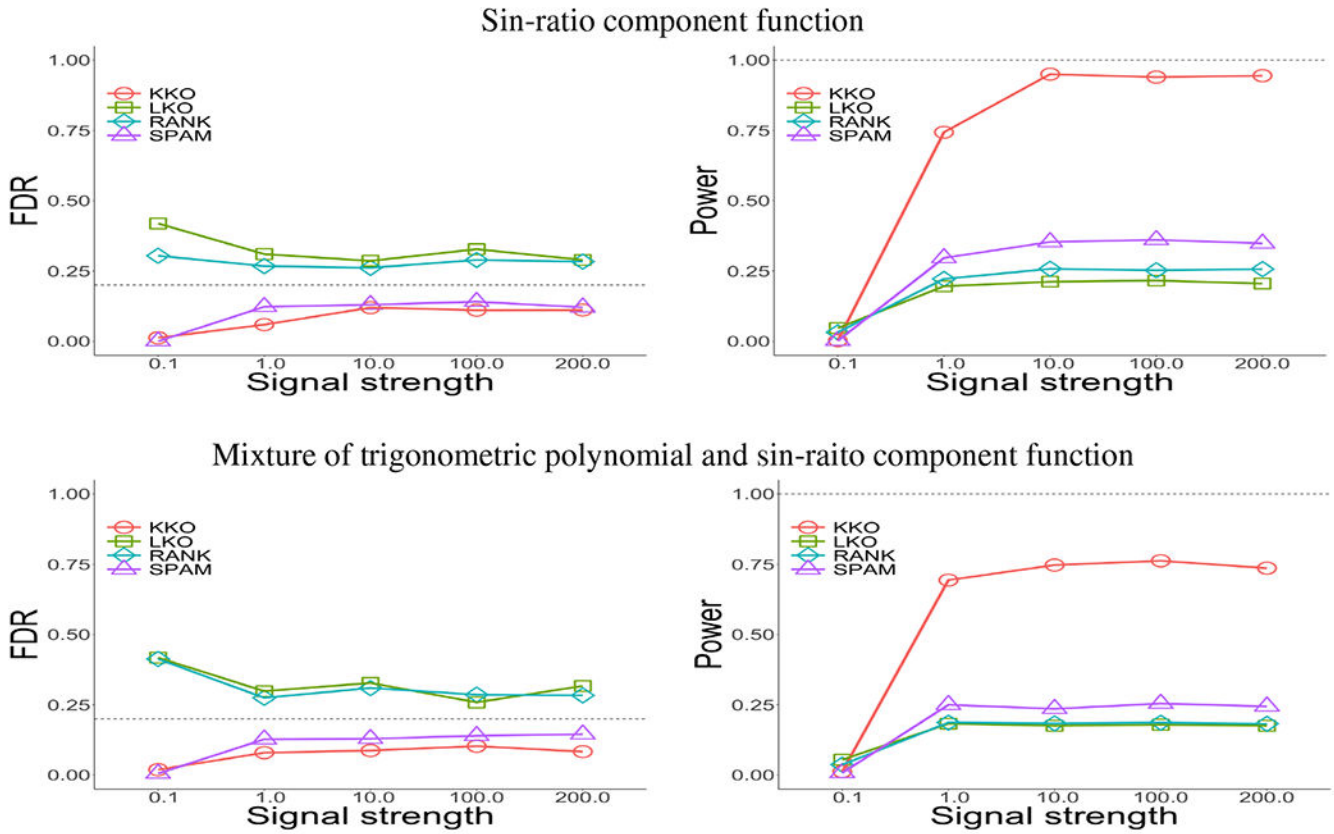

Figure 3:

Empirical performance and comparison in terms of FDR and power with the varying signal strength and component function. The same four methods as in Figure 2 are compared.

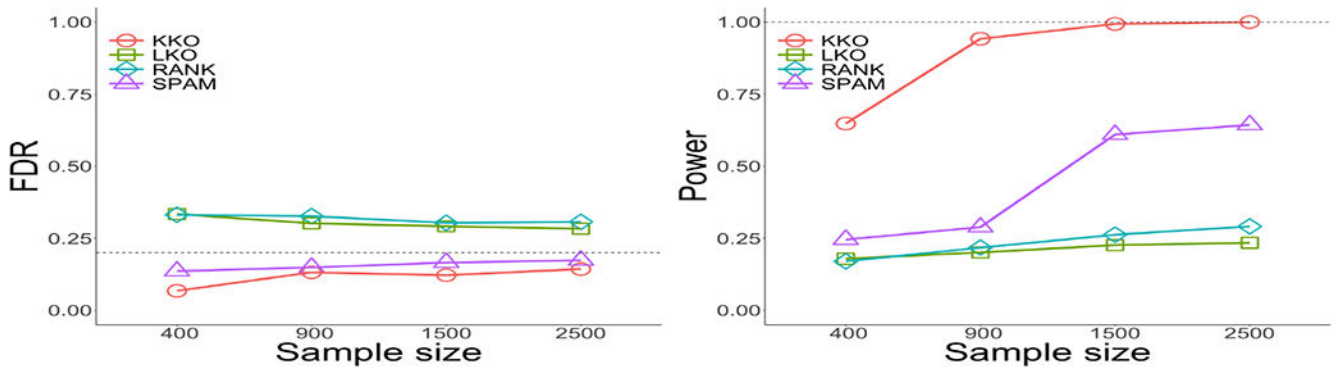

Figure 4:

Empirical performance and comparison in terms of FDR and power with the varying sample size *n*. The same four methods as in Figure 2 are compared.

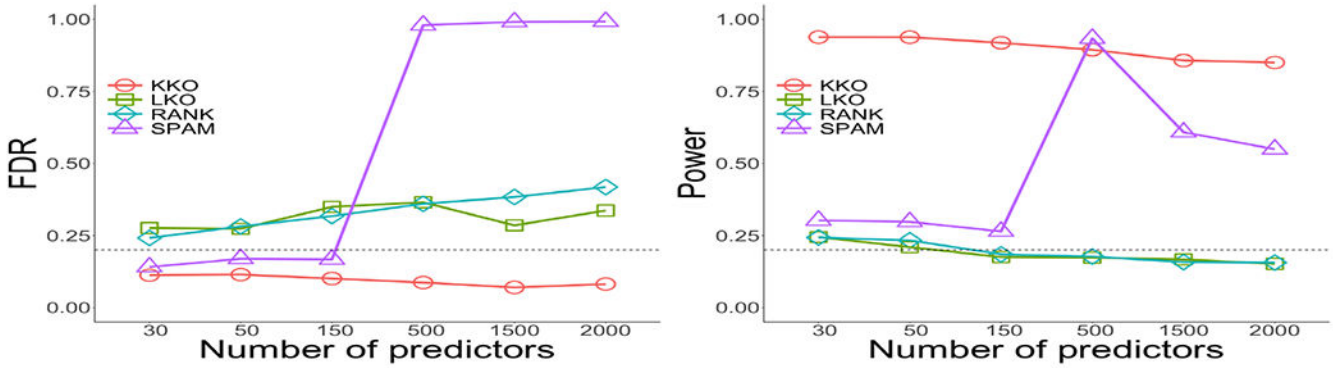

Figure 5:

Empirical performance and comparison in terms of FDR and power with the varying number of predictors p in both regimes where $p < n$ and $p > n$ with $n = 900$. The same four methods as in Figure 2 are compared.

Table 1:

Brain regions identified by the kernel knockoffs selection procedure. “l-” stands for the left hemisphere, and “r-” stands for the right hemisphere.

l-middletemporal	l-superiorparietal	r-fusiform	r-inferiorparietal	l-entorhinal
l-fusiform	l-inferiorparietal	l-precuneus	l-superiortemporal	r-entorhinal