

IPAD: Stable Interpretable Forecasting with Knockoffs Inference

Yingying Fan^a, Jinchi Lv^a, Mahrad Sharifvaghefi^b, and Yoshimasa Uematsu^a

^aUniversity of Southern California, Data Sciences and Operations Department, Los Angeles, CA; ^bUniversity of Southern California, Department of Economics, Los Angeles, CA; ^cTohoku University, Department of Economics and Management, Sendai, Japan

ABSTRACT

Interpretability and stability are two important features that are desired in many contemporary big data applications arising in statistics, economics, and finance. While the former is enjoyed to some extent by many existing forecasting approaches, the latter in the sense of controlling the fraction of wrongly discovered features which can enhance greatly the interpretability is still largely underdeveloped. To this end, in this article, we exploit the general framework of model-X knockoffs introduced recently in Candès, Fan, Janson and Lv [(2018), “Panning for Gold: ‘model X’ Knockoffs for High Dimensional Controlled Variable Selection,” *Journal of the Royal Statistical Society, Series B*, 80, 551–577], which is nonconventional for reproducible large-scale inference in that the framework is completely free of the use of p -values for significance testing, and suggest a new method of intertwined probabilistic factors decoupling (IPAD) for stable interpretable forecasting with knockoffs inference in high-dimensional models. The recipe of the method is constructing the knockoff variables by assuming a latent factor model that is exploited widely in economics and finance for the association structure of covariates. Our method and work are distinct from the existing literature in which we estimate the covariate distribution from data instead of assuming that it is known when constructing the knockoff variables, our procedure does not require any sample splitting, we provide theoretical justifications on the asymptotic false discovery rate control, and the theory for the power analysis is also established. Several simulation examples and the real data analysis further demonstrate that the newly suggested method has appealing finite-sample performance with desired interpretability and stability compared to some popularly used forecasting methods. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received February 2019
Revised June 2019

KEYWORDS

Large-scale inference and FDR; Latent factors; Model-X knockoffs; Power; Reproducibility; Stability

1. Introduction

Forecasting is a fundamental problem that arises in statistics, economics, and finance. With the availability of big data, many machine learning algorithms such as the Lasso and random forest can be resorted to for such a purpose by exploring a large pool of potential features. Many of these existing procedures provide a certain measure of feature importance which can then be utilized to judge the relative importance of selected features for the goal of interpretability. Yet, the issue of stability in the sense of controlling the fraction of wrongly discovered features is still largely underdeveloped. As argued in De Mol, Giannone, and Reichlin (2008) in the econometric settings, it is difficult to obtain interpretability and stability simultaneously even in simple Lasso forecasting. A natural question is how to ensure both interpretability and stability for flexible forecasting.

Naturally, stability is related to statistical inference. The recent years have witnessed a growing body of work on high-dimensional inference in the statistics and econometrics literature; see, for example, Wooldridge and Zhu (2018), Stucky and van de Geer (2018), Zhang and Cheng (2017), Chernozhukov et al. (2018), Chernozhukov, Newey, and Robins (2018), Chernozhukov et al. (2018), Shah and Bühlmann

(2018), and Guo et al. (2018). Most existing work on high-dimensional inference for interpretable models has focused primarily on the aspects of post-selection inference known as selective inference and debiasing for regularization and machine learning methods. In real applications, one is often interested in conducting *global* inference relative to the full model as opposed to *local* inference conditional on the selected model. Moreover, many statistical inferences are based on p -values from significance testing. However, oftentimes obtaining valid p -values even for the Lasso in relatively complicated high-dimensional nonlinear models also remains largely unresolved, not to mention for the case of more complicated model fitting procedures such as random forest. Indeed, high-dimensional inference is intrinsically challenging even in the parametric settings Fan, Demirkaya, and Lv (2019).

The desired property of stability for interpretable forecasting in this article concentrates on *global* inference by controlling precisely the fraction of wrongly discovered features in high-dimensional models, which is also known as reproducible large-scale inference. Such a problem involves testing the joint significance of a large number of features simultaneously, which is known widely as the problem of multiple testing in statistical inference. For this problem, the null hypothesis for each feature states that the feature is unimportant in the joint model

CONTACT Yoshimasa Uematsu  uematsu0911@gmail.com  Tohoku University, Department of Economics and Management, 27-1 Kawauchi, Aobaku, Sendai 980-8576, Japan.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

which can be understood as the property that this individual feature and the response are *independent* conditional on all the remaining features, while the corresponding alternative hypothesis states the opposite. Conventionally, p -values from the hypothesis testing are used to decide whether or not to reject each null hypothesis with a significance level to control the probability of false discovery in a single hypothesis test, meaning rejecting the null hypothesis when it is true. When performing multiple hypothesis tests, the probability of making at least one false discovery which is known as the family-wise error rate can be inflated compared to that for the case of a single hypothesis test. The work on controlling such an error rate for multiple testing dates back to Bonferroni (1935), where a simple, useful idea is lowering the significance level for each individual test as the target level divided by the total number of tests to be performed. The Bonferroni correction procedure is, however, well known to be conservative with relatively low power. Later on, Holm (1979) proposed a step-down procedure which is less conservative than the Bonferroni procedure. More recently, Romano and Wolf (2005) suggested a procedure in which the critical values of individual tests are constructed sequentially.

A more powerful and extremely popular approach to multiple testing is the Benjamini–Hochberg (BH) procedure for controlling the false discovery rate (FDR) which was originated in Benjamini and Hochberg (1995), where the FDR is defined as the expectation of the fraction of falsely rejected null hypotheses known as the false discovery proportion. Given the p -values from the multiple hypothesis tests, this procedure sorts the p -values from low to high and chooses a simple, intuitive cut-off point, which can be viewed as an adaptive extension of the Bonferroni correction for multiple comparisons, of the p -values for rejecting the null hypotheses. The BH procedure was shown to be capable of controlling the FDR at the desired level for independent test statistics in Benjamini and Hochberg (1995) and for positive regression dependency among the test statistics in Benjamini and Yekutieli (2001), where it was shown that a simple modification of the procedure can control the FDR under other forms of dependency but such a modification is generally conservative. There is a huge literature on the theory, applications, and various extensions of the original BH procedure for FDR control; see, for instance, Benjamini (2010), Fan, Han, and Gu (2012), Belloni et al. (2018), and Chudik, Kapetanios, and Pesaran (2018).

The aforementioned econometric and statistical inference methods including the BH-type procedures for FDR control are all rooted on the availability and validity of computable p -values for evaluating variable importance. As mentioned before, such a prerequisite can become a luxury that is largely unclear how to obtain in high dimensions even for the case of Lasso in general nonlinear models and random forest. In contrast, Barber and Candès (2015) proposed a novel procedure named the knockoff filter for FDR control that bypasses the use of p -values in the Gaussian linear model with deterministic design matrix, where the dimensionality is no larger than the sample size, and Barber and Candès (2016) generalized the method to high-dimensional linear models as a two-step procedure based on sample splitting, where a feature screening approach is used to reduce the dimen-

sionality to below sample size (see, e.g., Fan and Fan (2008) and Fan and Lv (2008)) and then the knockoff filter is applied to the set of selected features after the screening step for selective inference. The key ingredient of the knockoff filter is constructing the so-called knockoff variables in a geometrical way that mimic perfectly the correlation structure among the original covariates and can be used as control variables to evaluate the importance of original variables. Recently, Candès et al. (2018) extended the work of Barber and Candès (2015) by introducing the framework of model-X knockoffs for FDR control in general high-dimensional nonlinear models. A crucial distinction is that the knockoff variables are constructed in a probabilistic fashion such that the joint dependency structure of the original variables and their knockoff copies is invariant to the swapping of any set of original variables and their knockoff counterparts, which enables us to go beyond linear models and handle high dimensionality. As a result, model-X knockoffs enjoys exact finite-sample FDR control at the target level. However, a major assumption in Candès et al. (2018) is that the joint distribution of all the covariates needs to be *known* for the valid FDR control.

Motivated by applications in economics and finance, in this article we model the association structure of the covariates using the latent factor model, which reduces effectively the dimensionality and enables reliable estimation of the *unknown* joint distribution of all the covariates. By taking into account, the latent factor model structure, we first estimate the association structure of covariates and then construct *empirical* knockoffs matrix using the estimated dependency structure. Our empirical knockoffs matrix can be regarded as an approximation to the *oracle* knockoffs matrix in Candès et al. (2018) that requires the knowledge of the true covariate distribution. Exploiting the general framework of model-X knockoffs in Candès et al. (2018), we suggest the new method of intertwined probabilistic factors decoupling (IPAD) for stable interpretable forecasting with knockoffs inference in high-dimensional models. The innovations of our method and work are fourfold. First, we estimate the covariate distribution from data instead of assuming that it is known when constructing the knockoff variables. Second, our procedure does not require any sample splitting and is thus more practical when the sample size is limited. Third, we provide theoretical justifications on the asymptotic FDR control when the estimated dependency structure is employed. Fourth, the theory for power analysis is also established which reveals that there can be asymptotically no power loss in applying the knockoffs procedure compared to the underlying variable selection method. Therefore, FDR control by knockoffs can be a pure gain. Compared to earlier work, an additional challenge of our study is that knowing the true underlying distribution does *not* lead to the most efficient construction of the oracle knockoffs matrix due to the presence of latent factors.

The rest of the article is organized as follows. Section 2 introduces the model setting and presents the new IPAD procedure. We establish the asymptotic properties of IPAD in Section 3. Sections 4 and 5 present several simulation and real data examples to showcase the finite-sample performance and the advantages of our newly suggested procedure compared to some popularly used ones. We discuss some implications and extensions of our work in Section 6. The proofs of the

main results are relegated to appendix. Additional technical details and numerical results are provided in the supplementary material.

2. Intertwined Probabilistic Factors Decoupling

To facilitate the technical presentation, we will introduce the model setting for the high-dimensional FDR control problem in Section 2.1 and present the new IPAD procedure in Section 2.2.

2.1. Model Setting

Consider the high-dimensional linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the random matrix of a large number of potential regressors, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$ is the regression coefficient vector, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the vector of model errors, and n and p denote the sample size and dimensionality, respectively. Here without loss of generality, we assume that both the response and the covariates are centered with mean zero and thus there is no intercept. Motivated by many applications in economics and finance, we further assume that the design matrix \mathbf{X} follows the *exact* factor model

$$\mathbf{X} = \mathbf{F}^0 \boldsymbol{\Lambda}^0 + \mathbf{E} = \mathbf{C}^0 + \mathbf{E}, \tag{2}$$

where $\mathbf{F}^0 = (\mathbf{f}_1^0, \dots, \mathbf{f}_n^0)' \in \mathbb{R}^{n \times r}$ is a random matrix of latent factors, $\boldsymbol{\Lambda}^0 = (\boldsymbol{\lambda}_1^0, \dots, \boldsymbol{\lambda}_p^0)' \in \mathbb{R}^{p \times r}$ is a matrix of deterministic factor loadings, and error term $\mathbf{E} \in \mathbb{R}^{n \times p}$ captures the remaining variation that cannot be explained by these latent factors. We assume that the number of factors r is fixed but *unknown* and the components of \mathbf{E} are independent and identically distributed (iid) from some parametric distribution with cumulative distribution function $G(\cdot; \boldsymbol{\eta}^0)$, where $\boldsymbol{\eta}^0 \in \mathbb{R}^m$ is a finite-dimensional unknown parameter vector. For technical simplicity, models (1) and (2) are assumed to have no endogeneity and satisfy that \mathbf{F}^0 has iid rows and is independent of \mathbf{E} .

In this article, we focus on the high-dimensional scenario when the dimensionality p can be much larger than sample size n . Therefore, to ensure model identifiability we impose the sparsity assumption that the true regression coefficient vector $\boldsymbol{\beta}$ has only a small portion of nonzeros; specifically, $\boldsymbol{\beta}$ takes nonzero values only on some (unknown) index set $\mathcal{S}^0 \subset \{1, \dots, p\}$ and $\beta_j = 0$ for all $j \in \mathcal{S}^1 := \{1, \dots, p\} \setminus \mathcal{S}^0$. Denote by $s = |\mathcal{S}^0|$ the size of \mathcal{S}^0 . We assume that $s = o(n)$ throughout the article.

We are interested in identifying the index set \mathcal{S}^0 with a theoretically guaranteed error rate. To be more precise, we try to select variables in \mathcal{S}^0 while keeping the FDR under some prespecified desired level $q \in (0, 1)$, where the FDR is defined as

$$\text{FDR} := \mathbb{E}[\text{FDP}] \quad \text{with} \quad \text{FDP} := \frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^1|}{|\widehat{\mathcal{S}} \vee \mathcal{S}^1|}. \tag{3}$$

Here, the FDP stands for the false discovery proportion and $\widehat{\mathcal{S}}$ represents the set of variables selected by some procedure

using observed data (\mathbf{X}, \mathbf{y}) . A slightly modified version of FDR is defined as

$$\text{mFDR} := \mathbb{E} \left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^1|}{|\widehat{\mathcal{S}}| + q^{-1}} \right]. \tag{4}$$

Clearly, FDR is more conservative than mFDR in that the latter is always under control if the former is.

It is easy to see that FDR is a measurement of Type I error for variable selection. The other important aspect of variable selection is power, which is defined as

$$\text{Power} := \mathbb{E} \left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^0|}{|\mathcal{S}^0|} \right] = \mathbb{E} \left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^0|}{s} \right]. \tag{5}$$

It is well known that FDR and power are two sides of the same coin. We aim at developing a variable selection procedure with theoretically guaranteed FDR control and meanwhile achieving high power.

2.2. IPAD

The key ingredient of the model-X knockoffs framework introduced originally in Candès et al. (2018) is the construction of the so-called model-X knockoff variables defined as follows.

Definition 1 (Model-X knockoff variables Candès et al. (2018)).

For a set of random variables $\mathbf{x} = (X_1, \dots, X_p)$, a new set of random variables $\tilde{\mathbf{x}} = (\tilde{X}_1, \dots, \tilde{X}_p)$ is called a set of model-X knockoff variables if it satisfies the following properties:

- 1) For any subset $\mathcal{S} \subset \{1, \dots, p\}$, we have $[\mathbf{x}, \tilde{\mathbf{x}}]_{\text{swap}(\mathcal{S})} \stackrel{d}{=} [\mathbf{x}, \tilde{\mathbf{x}}]$, where $\stackrel{d}{=}$ denotes equal in distribution and the vector $[\mathbf{x}, \tilde{\mathbf{x}}]_{\text{swap}(\mathcal{S})}$ is obtained by swapping X_j and \tilde{X}_j for each $j \in \mathcal{S}$.
- 2) Conditional on \mathbf{x} , the knockoffs vector $\tilde{\mathbf{x}}$ is independent of response Y .

See Section B of the supplementary material for a brief review of the model-X knockoffs framework. In theory, if the distribution of \mathbf{C}^0 and the value of $\boldsymbol{\eta}^0$ are known, the SCIP algorithm proposed in Candès et al. (2018) can be used to construct the knockoff variables. However, the computational cost can be high depending on the exact distributions. Instead, we introduce a more efficient and practically implementable approach for constructing the knockoff variables below.

We start with introducing the *knockoff generating function*—for each given augmented parameter vector $\boldsymbol{\theta} = \text{vec}(\text{vec}(\mathbf{C}, \boldsymbol{\eta}))$, define

$$\tilde{\mathbf{X}}(\boldsymbol{\theta}) = \mathbf{C} + \mathbf{E}_\eta, \tag{6}$$

where \mathbf{E}_η is a matrix composed of iid random samples from distribution $G(\cdot; \boldsymbol{\eta})$. To gain some insights, let us first consider the ideal situation where the factor model structure (2) is fully available; that is, we know the realization \mathbf{C}^0 and the true distribution $G(\cdot; \boldsymbol{\eta}^0)$ for the error matrix \mathbf{E} . In such case, the *oracle (ideal)* knockoffs matrix $\tilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ can be constructed as

$$\tilde{\mathbf{X}}(\boldsymbol{\theta}^0) = \mathbf{C}^0 + \mathbf{E}_{\eta^0}, \tag{7}$$

where \mathbf{E}_{η^0} is an iid copy of \mathbf{E} . Note that \mathbf{E}_{η^0} itself is not a function of $\boldsymbol{\eta}^0$, but we slightly abuse the notation to emphasize

the dependence of the distribution function on parameter η^0 . In practice, θ_0 is unknown and needs to be estimated. Letting $\hat{\theta}$ denote an estimator (obtained using data \mathbf{X}) of θ^0 , we name $\tilde{\mathbf{X}}(\hat{\theta})$ as the empirical knockoffs matrix

$$\tilde{\mathbf{X}}(\hat{\theta}) = \widehat{\mathbf{C}} + \mathbf{E}_{\hat{\eta}}, \tag{8}$$

where $\widehat{\mathbf{C}}$ is an empirical estimate of \mathbf{C}^0 and $\mathbf{E}_{\hat{\eta}} \in \mathbb{R}^{n \times p}$ is composed of iid random variables from the plug-in estimate of the distribution function, $G(\cdot; \hat{\eta})$, and is independent of (\mathbf{X}, \mathbf{y}) conditional on $\hat{\eta}$. The following proposition justifies the validity of the oracle knockoffs matrix.

Proposition 1. Under model setting (2), the oracle knockoffs matrix defined in (7) satisfies Definition 1.

However, the empirical knockoffs matrix given in (8) generally does not satisfy the exchangeability property because of the dependence of $\hat{\theta}$ on the training data \mathbf{X} . Although the oracle knockoffs matrix is generally unavailable, it plays an important role in our theoretical developments as a proxy of the empirical knockoffs matrix. We remark that in the construction above, we slightly misuse the concept and call \mathbf{C}^0 a parameter. This is because although \mathbf{C}^0 is a random matrix, for the construction of valid knockoff variables it is the particular realization \mathbf{C}^0 leading to the observed data matrix \mathbf{X} that matters. In other words, a valid construction of knockoff variables requires the knowledge of the specific realization \mathbf{C}^0 instead of the distribution of \mathbf{C}^0 . To understand this, consider the scenario where the underlying parameter η^0 and the exact distribution of \mathbf{C}^0 are fully available. If we independently generate random variables from this known distribution and form a new data matrix \mathbf{X}_1 , because of the independence between \mathbf{X}_1 and \mathbf{X} , the exchangeability assumption in Definition 1 will be violated and thus \mathbf{X}_1 cannot be a valid knockoffs matrix. On the other hand, as long as we know the realization \mathbf{C}^0 and parameter η^0 , a valid knockoffs matrix $\tilde{\mathbf{X}}(\theta^0)$ can be constructed using (7) regardless of whether the exact distribution of \mathbf{C}^0 is available or not.

In practice, however, θ^0 is unavailable and consequently, $\tilde{\mathbf{X}}(\theta^0)$ is inaccessible. To overcome this difficulty, we next introduce our new method IPAD. With the aid of empirical knockoffs matrix, we suggest the following IPAD procedure for FDR control with knockoffs inference.

- Procedure 1 (IPAD).**
- (Estimation of parameters) Estimate the unknown parameters in θ^0 using the design matrix \mathbf{X} . Denote by $\hat{\theta} = (\widehat{\mathbf{C}}, \hat{\eta})$ the resulting estimated parameter vector.
 - (Construction of empirical knockoffs matrix) Construct the empirical knockoffs matrix (8) by applying the knockoff generating function in (6) to the estimated parameter $\hat{\theta}$.
 - (Application of knockoffs inference) Calculate knockoff statistics $W_j(\hat{\theta})$ using data $([\mathbf{X}, \tilde{\mathbf{X}}(\hat{\theta})], \mathbf{y})$ and then construct $\hat{\mathcal{S}}$ by applying knockoffs inference to $W_j(\hat{\theta})$.

Intuitively, the accuracy of the estimator $\hat{\theta}$ in Step 1 will affect the performance of our IPAD procedure. In fact, as shown later in our Theorem 1 in Section 3, the consistency rate of $\hat{\theta}$ is indeed reflected in the asymptotic FDR control of the IPAD procedure.

There are various ways to construct estimator $\hat{\theta}$. A natural and popularly used one is the principal component (PC) estimator $(\widehat{\mathbf{F}}, \widehat{\mathbf{\Lambda}})$ studied in Bai (2003). Specifically, we first estimate the number of factors r , denoted as \hat{r} , using some method such as the information criterion in Bai and Ng (2002) or the approach in Ahn and Horenstein (2013), and then set $\widehat{\mathbf{C}} = \widehat{\mathbf{F}}\widehat{\mathbf{\Lambda}}'$, where $\widehat{\mathbf{F}}$ is $T^{1/2}$ times the eigenvectors corresponding to the top \hat{r} largest eigenvalues of $\mathbf{X}\mathbf{X}'$, and $\widehat{\mathbf{\Lambda}} = \mathbf{X}'\widehat{\mathbf{F}}/T$. As for the estimation of η^0 , existing methods such as the method of moments can be used based on the residual matrix $\widehat{\mathbf{E}} = (\hat{e}_{ij})$. As a concrete example, consider the case where \mathbf{E} has iid $\mathcal{N}(0, \sigma^2)$ entries. Then the unknown population parameter is $\eta^0 = \sigma^2$ and can be estimated naturally as $(np)^{-1} \sum_{i=1}^n \sum_{j=1}^p \hat{e}_{ij}^2$.

In Step 3, various methods can be used to construct knockoff statistics. For the illustration purpose, we use the Lasso coefficient difference (LCD) statistic as in Candès et al. (2018). Specifically, with \mathbf{y} being the response vector and $([\mathbf{X}, \tilde{\mathbf{X}}(\hat{\theta})])$ the augmented design matrix, we consider the variable selection procedure Lasso Tibshirani (1996) which solves the following optimization problem

$$\hat{\boldsymbol{\beta}}^{\text{aug}}(\hat{\theta}; \lambda) = \arg \min_{\mathbf{b} \in \mathbb{R}^{2p}} \left\{ \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}(\hat{\theta})]\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}, \tag{9}$$

where $\lambda \geq 0$ is the regularization parameter and $\|\cdot\|_m$ with $m \geq 1$ denotes the vector ℓ_m -norm. Then, for each variable \mathbf{x}_j , the knockoff statistic can be constructed as

$$W_j(\hat{\theta}; \lambda) = |\hat{\beta}_j^{\text{aug}}(\hat{\theta}; \lambda)| - |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta}; \lambda)|, \tag{10}$$

where $\hat{\beta}_\ell^{\text{aug}}(\hat{\theta}; \lambda)$ is the ℓ th component of the Lasso regression coefficient vector $\hat{\boldsymbol{\beta}}^{\text{aug}}(\hat{\theta}; \lambda)$. It is seen that intuitively the LCD knockoff statistics evaluate the relative importance of the j th original variable by comparing its Lasso coefficient $\hat{\beta}_j^{\text{aug}}(\hat{\theta}; \lambda)$ with that of its knockoff copy $\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta}; \lambda)$. In the ideal case when the oracle knockoffs matrix $\tilde{\mathbf{X}}(\theta^0)$ is used instead of $\tilde{\mathbf{X}}(\hat{\theta})$ in (9), it is easy to verify that the LCD is a valid construction of knockoff statistics and satisfies the sign-flip property in (A.2) of Supplementary Material. Consequently, the general theory in Candès et al. (2018) can be applied to show that the FDR is controlled in finite sample. We next show that even with the empirical knockoffs matrix employed in (9), the FDR can still be asymptotically controlled with delicate technical analyses.

3. Asymptotic Properties of IPAD

We now provide theoretical justifications for our IPAD procedure suggested in Section 2 with the LCD knockoff statistics $W_j(\hat{\theta}; \lambda) = w_j([\mathbf{X}, \tilde{\mathbf{X}}(\hat{\theta})], \mathbf{y}; \lambda)$ defined in (10). We will first present some technical conditions in Section 3.1, then prove in Section 3.2 that the FDR is asymptotically under control at a desired target level q , and finally in Section 3.3 show that asymptotically IPAD has no power loss compared to the Lasso under some regularity conditions.

3.1. Technical Conditions

We first introduce some notation and definitions which will be used later on. We use $X \sim \text{subG}(C_x^2)$ to denote that X is

a sub-Gaussian random variable with *variance proxy* $C_x^2 > 0$ if $\mathbb{E}[X] = 0$ and its tail probability satisfies $\mathbb{P}(|X| > u) \leq 2 \exp(-u^2/C_x^2)$ for each $u \geq 0$. In all technical assumptions below, we use $M > 1$ to denote a large enough generic constant. Throughout the article, for any vector $\mathbf{v} = (v_i)$ let us denote by $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_{\max}$ the ℓ_1 -norm, ℓ_2 -norm, and max-norm defined as $\|\mathbf{v}\|_1 = \sum_i |v_i|$, $\|\mathbf{v}\|_2 = (\sum_i v_i^2)^{1/2}$, and $\|\mathbf{v}\|_{\max} = \max_i |v_i|$, respectively. For any matrix $\mathbf{M} = (m_{ij})$, we denote by $\|\mathbf{M}\|_F$, $\|\mathbf{M}\|_1$, $\|\mathbf{M}\|_2$, and $\|\mathbf{M}\|_{\max}$ the Frobenius norm, entrywise ℓ_1 -norm, spectral norm, and entrywise ℓ_∞ -norm defined as $\|\mathbf{M}\|_F = \|\text{vec}(\mathbf{M})\|_2$, $\|\mathbf{M}\|_1 = \|\text{vec}(\mathbf{M})\|_1$, $\|\mathbf{M}\|_2 = \sup_{\mathbf{v} \neq \mathbf{0}} \|\mathbf{M}\mathbf{v}\|_2 / \|\mathbf{v}\|_2$, and $\|\mathbf{M}\|_{\max} = \|\text{vec}(\mathbf{M})\|_{\max}$, respectively, where $\text{vec}(\mathbf{M})$ represents the vectorization of \mathbf{M} . For a symmetric matrix \mathbf{M} , $\text{vech}(\mathbf{M})$ stands for the vectorization of the lower triangular part.

Condition 1 (Regression errors). The model error vector $\boldsymbol{\varepsilon}$ has iid components from $\text{subG}(C_\varepsilon^2)$.

Condition 2 (Latent factors). The rows of \mathbf{F}^0 consist of mean zero iid random vectors $\mathbf{f}_i^0 \in \mathbb{R}^r$ such that $\|\mathbf{F}^0\|_{\max} \leq M$ almost surely (a.s.) and $\|\boldsymbol{\Sigma}_f\|_2 + \|\boldsymbol{\Sigma}_f^{-1}\|_2 \leq M$, where $\boldsymbol{\Sigma}_f := \mathbb{E}[\mathbf{f}_i^0 \mathbf{f}_i^{0'}]$.

Condition 3 (Factor loadings). The rows of $\mathbf{\Lambda}^0$ consist of deterministic vectors $\boldsymbol{\lambda}_j^0 \in \mathbb{R}^r$ such that $\|\mathbf{\Lambda}^0\|_{\max} \leq M$ and $\|p^{-1} \mathbf{\Lambda}^0 \mathbf{\Lambda}^0\|_2 + \|(p^{-1} \mathbf{\Lambda}^0 \mathbf{\Lambda}^0)^{-1}\|_2 \leq M$.

Condition 4 (Factor errors). The entries of matrix \mathbf{E}_{η^0} are iid copies of $e_{\eta^0} \sim \text{subG}(C_e^2)$ with continuous distribution function $G(\cdot; \boldsymbol{\eta}^0)$. For each $1 \leq \ell \leq m$, the ℓ th element of $\boldsymbol{\eta}^0$ is specified as $\eta_\ell^0 = h_\ell(\mathbb{E}[e_{\eta^0}], \dots, \mathbb{E}[e_{\eta^0}^m])$ with $h_\ell: \mathbb{R}^m \rightarrow \mathbb{R}$ some local Lipschitz continuous function in the sense that

$$\begin{aligned} & \left| h_\ell(t_1, \dots, t_m) - h_\ell(\mathbb{E}[e_{\eta^0}], \dots, \mathbb{E}[e_{\eta^0}^m]) \right| \\ & \leq M \max_{k \in \{1, \dots, m\}} \left| t_k - \mathbb{E}[e_{\eta^0}^k] \right| \end{aligned}$$

for each $t_k \in \{t : |t - \mathbb{E}[e_{\eta^0}^k]| \leq M c_{np}\}$ and $1 \leq k \leq m$, where $c_{np} := (p^{-1} \log n)^{1/2} + (n^{-1} \log p)^{1/2}$. Moreover, there exists some stochastic process $(e_\eta)_\eta$ such that

- (i) for each $\boldsymbol{\eta} \in \{\boldsymbol{\eta} \in \mathbb{R}^m : \|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|_{\max} \leq M c_{np}\}$, the entries of \mathbf{E}_η in (6) have identical distribution to e_η ,
- (ii) for some sub-Gaussian random variable $Z \sim \text{subG}(c_e^2)$ with some positive constant c_e ,

$$\sup_{\boldsymbol{\eta}: \|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|_{\max} \leq M c_{np}} |e_\eta - e_{\eta^0}| \leq M^{1/2} c_{np}^{1/2} |Z|. \quad (11)$$

Condition 5 (Eigenseparation). The r eigenvalues of $p^{-1} \mathbf{\Lambda}^0 \mathbf{\Lambda}^0 \boldsymbol{\Sigma}_f$ are distinct for all p .

The number of factors r is assumed to be known for developing the theory with simplification, but in practice it can be estimated consistently using methods such as information criteria (Bai and Ng 2002) and test statistics (Ahn and Horenstein 2013). The sub-Gaussian assumptions in Conditions 1 and 4 can be replaced with some other tail conditions as long as similar concentration inequalities hold. Condition 3 is standard in the

analysis of factor models. Stochastic loadings can be assumed in Condition 3 with some appropriate distributional assumption, such as sub-Gaussianity, at the cost of much more tedious technical arguments. The boundedness of the eigenvalues of $\boldsymbol{\Sigma}_f$ in Condition 2 is standard while the iid assumption and boundedness of \mathbf{f}_i^0 are stronger compared to the existing literature (e.g., Bai and Ng (2002) and Bai (2003)). However, these conditions are imposed mostly for technical simplicity. In fact, the boundedness condition on \mathbf{f}_i^0 can be replaced with (unbounded) sub-Gaussian or other heavier-tail assumption whenever concentration inequalities are available at the cost of slower convergence rates and stronger sample size requirement. Our theory on FDR control is based on that in Candès et al. (2018), which applies only to the case of iid rows of design matrix \mathbf{X} . This is the main reason for imposing the iid assumption on ε_i and \mathbf{f}_i in Conditions 1 and 2. However, we conjecture that similar results can also hold in the presence of some sufficiently weak serial dependence in ε_i and \mathbf{f}_i . Condition 4 introduces a *sub-Gaussian process* e_η with respect to $\boldsymbol{\eta}$. The norm in (11) can be replaced with any other norm since $\boldsymbol{\eta}$ is finite dimensional. In the specific case when the components of \mathbf{E} have Gaussian distribution such that $\boldsymbol{\eta}$ is a scalar parameter representing variance, by the reflection principle for the Wiener process (Billingsley (1995), p.511), e_η can be constructed as a Wiener process and the inequality (11) can be satisfied. For more information on sub-Gaussian processes, see, for example, Vizcarra and Viens (2007). To understand why we need Condition 5, note from the proof of Lemma 3 that the PC estimator $(\widehat{\mathbf{F}}, \widehat{\mathbf{\Lambda}})$ is only consistent for $(\mathbf{F}^0 \mathbf{H}, \mathbf{\Lambda}^0 \mathbf{H}^{-1})$, where $\mathbf{H} = (\mathbf{\Lambda}^0 \mathbf{\Lambda}^0 / p) (\mathbf{F}^0 \widehat{\mathbf{F}} / n) \mathbf{V}^{-1}$ with \mathbf{V} an $r \times r$ diagonal matrix of r largest eigenvalues of $\mathbf{X} \mathbf{X}' / (np)$. Condition 5 guarantees that $\widehat{\mathbf{F}} \mathbf{F}^0 / n$ is asymptotically unique and invertible, which have been proved by Bai (2003), and the fact is used in the proof of Lemma 6 in the supplementary material. This ultimately ensures that \mathbf{C}^0 can be estimated well, which in turn guarantees that $\boldsymbol{\eta}^0$ can be estimated accurately.

Recall that in the IPAD procedure, we first obtain the augmented Lasso estimator $\widehat{\boldsymbol{\beta}}^{\text{aug}}(\boldsymbol{\theta}; \lambda) \in \mathbb{R}^{2p}$ by regressing \mathbf{y} on $[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]$. Denote by $\mathcal{A}^{\text{aug}}(\boldsymbol{\theta}; \lambda) = \text{supp}(\widehat{\boldsymbol{\beta}}^{\text{aug}}(\boldsymbol{\theta}; \lambda)) \subset \{1, \dots, 2p\}$ the active set of the augmented Lasso regression coefficient vector. Throughout this section, we content ourselves with sparse estimates satisfying

$$|\mathcal{A}^{\text{aug}}(\boldsymbol{\theta}; \lambda)| \leq k/2 \quad (12)$$

for some positive integer k which may diverge with n at an order to be specified later; see, for example, Fan and Lv (2013) and Lv (2013) for a similar constraint and justifications therein. This can always be achieved since users have the freedom to choose the size of the Lasso model.

3.2. FDR control

To develop the theory for IPAD, we consider the PC estimator $\widehat{\mathbf{C}}$ for the realization \mathbf{C}^0 summarized in Section 2.2. The estimator $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \dots, \widehat{\eta}_m)'$ is constructed as $\widehat{\eta}_\ell = h_\ell(\mathbb{E}_{np} \widehat{e}, \dots, \mathbb{E}_{np} \widehat{e}^m)$ with h_ℓ , $1 \leq \ell \leq m$, introduced in Condition 4 and $\mathbb{E}_{np} \widehat{e}^k = (np)^{-1} \sum_{1 \leq i \leq n, 1 \leq j \leq p} \widehat{e}_{ij}^k$ the empirical moments of \widehat{e}_{ij} . Throughout our theoretical analysis, we consider the regularization parameter fixed at $\lambda = C_0 n^{-1/2} \log p$ with

C_0 some large enough constant for all the Lasso procedures. Therefore, we will drop the dependence of various quantities on λ whenever there is no confusion. For example, we will write $\mathcal{A}^{\text{aug}}(\theta; \lambda)$ and $\hat{\beta}^{\text{aug}}(\theta; \lambda)$ as $\mathcal{A}^{\text{aug}}(\theta)$ and $\hat{\beta}^{\text{aug}}(\theta)$, respectively.

Denote by $\mathbf{U}(\theta) := n^{-1}[\mathbf{X}, \tilde{\mathbf{X}}(\theta)]'[\mathbf{X}, \tilde{\mathbf{X}}(\theta)]$ and $\mathbf{v}(\theta) := n^{-1}[\mathbf{X}, \tilde{\mathbf{X}}(\theta)]'\mathbf{y}$ and define $\mathbf{T}(\theta) := \text{vec}(\text{vech } \mathbf{U}(\theta), \mathbf{v}(\theta)) \in \mathbb{R}^P$ with $P := p(2p+3)$. The following lemma states that the statistic $\mathbf{T}(\theta)$ plays a crucial role in our procedure.

Lemma 1. The set of variables $\hat{\mathcal{S}}$ selected by Procedure 1 depends only on $\mathbf{T}(\theta)$.

For any given θ , define the active set $\mathcal{A}^*(\theta) := \mathcal{A}_1^{\text{aug}}(\theta) \cup \mathcal{A}_2^{\text{aug}}(\theta) \subset \{1, \dots, p\}$, where $\mathcal{A}_1^{\text{aug}}(\theta) := \{j : j \in \{1, \dots, p\} \cap \mathcal{A}^{\text{aug}}(\theta)\}$ and $\mathcal{A}_2^{\text{aug}}(\theta) := \{j-p : j \in \{p+1, \dots, 2p\} \cap \mathcal{A}^{\text{aug}}(\theta)\}$. That is, $\mathcal{A}^*(\theta)$ is equal to the support of knockoff statistics $(W_1(\theta), \dots, W_p(\theta))'$ if there are no ties on the magnitudes of the augmented Lasso coefficient vector $\hat{\beta}^{\text{aug}}(\theta)$.

We next focus on the low-dimensional structure of $\mathbf{T}(\theta)$ inherited from the augmented Lasso because it will be made clear that this is the key to controlling the FDR without sample splitting. For any subset $\mathcal{A} \subset \{1, \dots, p\}$, define a lower-dimensional expression of the vector as $\mathbf{T}_{\mathcal{A}}(\theta) := \text{vec}(\text{vech } \mathbf{U}_{\mathcal{A}}(\theta), \mathbf{v}_{\mathcal{A}}(\theta))$ with $\mathbf{U}_{\mathcal{A}}(\theta)$ the principle submatrix of $\mathbf{U}(\theta)$ formed by columns and rows in \mathcal{A} and $\mathbf{v}_{\mathcal{A}}(\theta)$ the subvector of $\mathbf{v}(\theta)$ formed by components in \mathcal{A} . Then it is easy to see that $\mathbf{U}_{\mathcal{A}}(\theta) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\theta)]'[\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\theta)]$ and $\mathbf{v}_{\mathcal{A}}(\theta) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\theta)]'\mathbf{y}$. Motivated by Lemma 1, we define a family of mappings indexed by \mathcal{A} that describes the selection algorithm of Procedure 1 with given dataset $([\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\theta)], \mathbf{y})$ that forms $\mathbf{T}_{\mathcal{A}}(\theta)$. Formally, define a mapping $S_{\mathcal{A}} : \mathbb{R}^{|\mathcal{A}|(2|\mathcal{A}|+3)} \rightarrow 2^{\mathcal{A}}$ as $\mathbf{t}_{\mathcal{A}} \mapsto S_{\mathcal{A}}(\mathbf{t}_{\mathcal{A}})$ for given $\mathbf{T}_{\mathcal{A}}(\theta) = \mathbf{t}_{\mathcal{A}}$, where $2^{\mathcal{A}}$ refers to the power set of \mathcal{A} . That is, $S_{\mathcal{A}}(\mathbf{t}_{\mathcal{A}})$ represents the outcome of first restricting ourselves to the smaller set of variables \mathcal{A} and then applying IPAD to $\mathbf{T}_{\mathcal{A}}(\theta) = \mathbf{t}_{\mathcal{A}}$ to further select variables from set \mathcal{A} .

Lemma 2. Under Conditions 1–4, for any subset $\mathcal{A} \supset \mathcal{A}^*(\theta)$ we have $S_{\{1, \dots, p\}}(\mathbf{T}(\theta)) = S_{\mathcal{A}}(\mathbf{T}_{\mathcal{A}}(\theta))$.

When restricting on set \mathcal{A} , we can apply Procedure 1 to a lower-dimensional dataset $([\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\theta)], \mathbf{y})$ that forms $\mathbf{T}_{\mathcal{A}}(\theta)$ to further select variables from \mathcal{A} . The previous two lemmas ensure that this gives us a subset of \mathcal{A} that is identical to $S_{\{1, \dots, p\}}(\mathbf{T}(\theta))$. Note that the lower-dimensional problem based on $\mathbf{T}_{\mathcal{A}}(\theta)$ can be easier compared to the original one. We also would like to emphasize that the dimensionality reduction to a smaller model \mathcal{A} is only for assisting the theoretical analysis and our Procedure 1 does not need any knowledge of such set \mathcal{A} .

It is convenient to define $\mathbf{t}_0 = \mathbb{E} \mathbf{T}(\theta^0) \in \mathbb{R}^P$. Denote by

$$\mathbb{I} := \{ \mathbf{t} \in \mathbb{R}^P : \|\mathbf{t} - \mathbf{t}_0\|_{\max} \leq a_{np} := C_1(k^{1/2} + s^{3/2})\tilde{c}_{np} \}, \tag{13}$$

where C_1 is some positive constant and $\tilde{c}_{np} = p^{-1/2} \log n + n^{-1/2} \log p$. For any subset $\mathcal{A} \subset \{1, \dots, p\}$, let $\mathbb{I}_{\mathcal{A}}$ be the subspace of \mathbb{I} when taking out the coordinates corresponding to $\mathbb{E} \mathbf{T}_{\mathcal{A}}(\theta^0)$. Thus, $\mathbb{I}_{\mathcal{A}} \subset \mathbb{R}^{|\mathcal{A}|(2|\mathcal{A}|+3)}$. In addition to Conditions 1–5, we need an assumption on the algorithmic stability of Procedure 1.

Condition 6 (Algorithmic stability). For any subset $\mathcal{A} \subset \{1, \dots, p\}$ that satisfies $|\mathcal{A}| \leq k \leq n \wedge p$, there exists a positive sequence $\rho_{np} \rightarrow 0$ as $n \wedge p \rightarrow \infty$ such that

$$\sup_{|\mathcal{A}| \leq k} \sup_{\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{I}_{\mathcal{A}}} \frac{|S_{\mathcal{A}}(\mathbf{t}_2) \Delta S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_1)| \wedge |S_{\mathcal{A}}(\mathbf{t}_2)|} = O(\rho_{np}),$$

where Δ stands for the symmetric difference between two sets.

Intuitively the above condition assumes that the knockoffs procedure is stable with respect to a small perturbation to the input \mathbf{t} in any lower-dimensional subspace $\mathbb{I}_{\mathcal{A}}$. Under these regularity conditions, the asymptotic FDR control of our IPAD procedure can be established.

Theorem 1 (Robust FDR control). Assume that Conditions 1–6 hold. Fix an arbitrary positive constant ν . If (s, k, n, p) satisfies $s \vee k \leq n \wedge p$, $c_{np} \leq c/[r^2 M^2 C(\nu + 2)]^{1/2}$, and $(k^{1/2} + s^{3/2})\tilde{c}_{np} \rightarrow 0$ as $n \wedge p \rightarrow \infty$ with c and C some positive constants defined in Lemma 7 in the supplementary material, then the set of variables $\hat{\mathcal{S}}$ obtained by Procedure 1 (IPAD) with the LCD knockoff statistics controls the FDR (3) to be no larger than $q + O(\rho_{np} + n^{-\nu} + p^{-\nu})$.

Recall that by definition, the FDR is a function of $\mathbf{T}(\hat{\theta})$ and can be written as $\mathbb{E} \text{FDP}(\mathbf{T}(\hat{\theta}))$ while the FDR computed with the oracle knockoffs, $\mathbb{E} \text{FDP}(\mathbf{T}(\theta^0))$, is perfectly controlled to be no larger than q . This observation motivates us to first establish asymptotic equivalence of $\mathbf{T}(\hat{\theta})$ and $\mathbf{T}(\theta^0)$ with large probability. Then a natural idea is to show that $\mathbb{E} \text{FDP}(\mathbf{T}(\hat{\theta}))$ converges to $\mathbb{E} \text{FDP}(\mathbf{T}(\theta^0))$ in probability, which turns out to be highly non-trivial because of the discontinuity of $\text{FDP}(\cdot)$ (the convergence would be straightforward via the Portmanteau lemma if $\text{FDP}(\cdot)$ were continuous). Condition 6 above provides a remedy to this issue by imposing the algorithmic stability assumption.

3.3. Power Analysis

We have established the asymptotic FDR control for our IPAD procedure in Section 3.2. We now look at the other side of the coin—the power (5). Recall that in IPAD, we apply the knockoffs inference procedure to the knockoff statistics LCD, which are constructed using the augmented Lasso in (9). Therefore, the final set of variables selected by IPAD is a subset of variables picked by the augmented Lasso. For this reason, the power of IPAD is always upper bounded by that of Lasso. We will show in this section that there is in fact no power loss relative to the augmented Lasso in the asymptotic sense.

Condition 7 (Signal strength I). For any subset $\mathcal{A} \subset \mathcal{S}^0$ that satisfies $|\mathcal{A}|/s > 1 - \gamma$ for some $\gamma \in (0, 1]$, it holds that $\|\beta_{\mathcal{A}}\|_1 > b_{np} s n^{-1/2} \log p$ for some positive sequence $b_{np} \rightarrow \infty$.

Condition 8 (Signal strength II). There exists some constant $C_2 \in (2(qs)^{-1}, 1)$ such that $|\mathcal{S}_2| \geq C_2 s$ with $\mathcal{S}_2 = \{j : |\beta_j| \gg (s/n)^{1/2} \log p\}$.

Condition 7 requires that the overall signal is not too weak, but is weaker than the conventional beta-min condition

$\min_{j \in S^0} |\beta_j| \gg n^{-1/2} \log p$. Under [Condition 8](#), we can show that $|\widehat{S}| \geq C_2 s$ with probability at least $1 - O(p^{-\nu} + n^{-\nu})$ using similar techniques to those of [Lemma 6](#) in [Fan et al. \(2019\)](#). The intuition is that given $s \rightarrow \infty$, for a variable selection procedure to have high power it should select at least a reasonably large number of variables. The result $|\widehat{S}| \geq C_2 s$ will be used to derive the asymptotic order of threshold T , which is in turn crucial to establish the theorem below on power.

Theorem 2 (Power guarantee). Assume that [Conditions 1–5](#) and [7–8](#) hold. Fix an arbitrary positive constant ν . If (s, k, n, p) satisfies $2s \leq k \leq n \wedge p$, $c_{np} \leq c/(r^2 M^2 C(\nu + 2))^{1/2}$, and $sk^{1/2} \tilde{c}_{np} \rightarrow 0$ as $n \wedge p \rightarrow \infty$ with c and C some positive constants defined in [Lemma 7](#), then both the Lasso procedure based on (\mathbf{X}, \mathbf{y}) and our IPAD procedure ([Procedure 1](#)) have power bounded from below by $\gamma - o(1)$ as $n \wedge p \rightarrow \infty$. In particular, if $\gamma = 1$ IPAD has no power loss compared to Lasso asymptotically.

4. Simulation Studies

We have shown in [Section 3](#) that IPAD can asymptotically control the FDR in high-dimensional setting and there can be no power loss in applying the procedure. We next move on to numerically investigate the finite-sample performance of IPAD using synthetic datasets. We will compare IPAD with the knockoff filter in [Barber and Candès \(2015\)](#) (BCKnockoff) and the high-dimensional knockoff filter in [Barber and Candès \(2016\)](#) (HD-BCKnockoff). In what follows, we will first explain in detail the model setups and simulation settings, then discuss the implementation of the aforementioned methods, and finally summarize the comparison results.

4.1. Simulation Designs and Settings

In all simulations, the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is generated from the factor model

$$\mathbf{X} = \mathbf{F}^0 (\mathbf{\Lambda}^0)' + \sqrt{r\theta} \mathbf{E} = \mathbf{C}^0 + \sqrt{r\theta} \mathbf{E}, \quad (14)$$

where $\mathbf{F}^0 = (\mathbf{f}_1^0, \dots, \mathbf{f}_n^0)' \in \mathbb{R}^{n \times r}$ is the matrix of latent factors, $\mathbf{\Lambda}^0 = (\boldsymbol{\lambda}_1^0, \dots, \boldsymbol{\lambda}_p^0)' \in \mathbb{R}^{p \times r}$ is the matrix of factor loadings, $\mathbf{E} \in \mathbb{R}^{n \times p}$ is the matrix of model errors, and θ is a constant controlling the signal-to-noise ratio. The term \sqrt{r} is used to single out the effect of the number of factors in calculating the signal-to-noise ratio in factor model (14). We then rescale each column of \mathbf{X} to have ℓ_2 -norm one and simulate the response vector $\mathbf{y} = (y_1, \dots, y_n)'$ from the following model

$$y_i = f(\mathbf{x}_i) + \sqrt{c} \varepsilon_i, \quad i = 1, \dots, n, \quad (15)$$

where $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is the link function which can be linear or nonlinear, $c > 0$ is a constant controlling the signal-to-noise ratio, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ is the vector of model error. We next explain the four different designs of our simulation studies.

4.1.1. Design 1: Linear Model With Normal Factor Design Matrix

The elements of \mathbf{F}^0 , $\mathbf{\Lambda}^0$, \mathbf{E} , and $\boldsymbol{\varepsilon}$ are drawn independently from $\mathcal{N}(0, 1)$. The link function takes a linear form, that is, $\mathbf{y} =$

$\mathbf{X}\boldsymbol{\beta} + \sqrt{c}\boldsymbol{\varepsilon}$, where the coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$ is generated by first choosing s random locations for the true signals and then setting β_j at each location to be either A or $-A$ randomly with A some positive value. The remaining $p - s$ components of $\boldsymbol{\beta}$ are set to zero.

4.1.2. Design 2: Linear Model With Fat-tail Factor Matrix and Serial Dependence

The elements of \mathbf{E} are generated as

$$e_{ij} = \left(\frac{\nu - 2}{\chi_{\nu, j}^2} \right) u_{ij}, \quad (16)$$

where $u_{ij} \sim \text{iid } \mathcal{N}(0, 1)$ for all $i = 1, \dots, n$ and $j = 1, \dots, p$, and $\chi_{\nu, j}^2$, $j = 1, \dots, p$ are iid random variables from chi-square distribution with $\nu = 8$ degrees of freedom. The rest of the design is the same as in [Design 1](#). It is worth mentioning that in this case, the entries of matrix \mathbf{E} have fat-tail distribution with serial dependence in each column because of the common factor $\chi_{\nu, j}^2$. This design is used to check the robustness of IPAD method with respect to the serial dependence and the fat-tail distribution of \mathbf{E} .

4.1.3. Design 3: Linear Model With Misspecified Design Matrix

To evaluate the robustness of IPAD procedure to the misspecification of the factor model structure (14), we set $\mathbf{\Lambda} = \mathbf{0}$, $r\theta = 1$ and simulate the rows of matrix \mathbf{E} independently from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{ij})$, $\sigma_{ij} = \rho^{|i-j|}$ for $i, j \leq p$. The remaining design is the same as in [Design 1](#). It is seen that our assumption on the independence of the entries of \mathbf{E} is violated. This design is used to test the robustness of IPAD to misspecification of the factor model structure of \mathbf{X} .

4.1.4. Design 4: Nonlinear Model With Normal Factor Design Matrix

Our last design is used to evaluate the performance of IPAD method when the link function f is nonlinear. To be more specific, we assume the following nonlinear model between the response and covariates

$$\mathbf{y} = \sin(\mathbf{X}\boldsymbol{\beta}) \exp(\mathbf{X}\boldsymbol{\beta}) + \sqrt{c}\boldsymbol{\varepsilon},$$

where the coefficient vector $\boldsymbol{\beta}$, design matrix \mathbf{X} , and model error $\boldsymbol{\varepsilon}$ are generated similarly as in [Design 1](#).

4.1.5. Simulation Settings

The target FDR level is set to be $q = 0.2$ in all simulations. For [Design 1](#) and [Design 2](#), we set $n = 2000$, $p = 2000$, $A = 4$, $s = 50$, $c = 0.2$, $r = 3$, and $\theta = 1$. To evaluate the sensitivity of our method to the dimensionality p and the model sparsity s , we also explore the settings of $p = 1000, 3000$ and $s = 100, 150$. In [Design 3](#), we set $r = 0$ and $\rho = 0, 0.5$. In [Design 4](#), since the model is nonlinear, we use the nonparametric method of random forest [Breiman \(2001\)](#) to fit the model and consider lower-dimensional settings of $p = 50, 250$, and 500 . We also decrease the number of observations to $n = 1000$ and number of true variables to $s = 10$. Moreover, we set $\theta = 1, 2$ and $c = 0.1, 0.2, 0.3$ to test the effects of signal-to-noise ratio on the performance of IPAD procedure in [Design](#)

4. The implementation details for the estimation procedure of IPAD are provided in Section E.1 of the supplementary material.

4.2. Simulation Results

For each method, we use 100 simulated datasets to calculate its empirical FDR and power, which are the average FDP and TDP (true discovery proportion as in (5)) over 100 repetitions, respectively. Two different thresholds, knockoff and knockoff+ (T_1 and T_2 in Result 1 in the supplementary material, respectively), are used in the knockoffs inference implementation. It is worth mentioning that as shown in Candès et al. (2018) and summarized in Result 1, knockoff+ controls FDR (3) exactly while knockoff controls only the modified FDR (4).

Tables 1 and 2 summarize the results from Designs 1 and 2, respectively. As shown in Table 1, all approaches can control empirical FDR at the target level ($q = 0.2$) and knockoff+, which is more conservative, reduces power negligibly. It is worth mentioning that even for Design 2, in which the design matrix X is drawn from fat-tail distribution with serial dependence, we still have FDR under control with decent level of power. This suggests that the no serial correlation assumption in our theoretical analysis could just be technical. Compared to the results by BCKnockoff and HD-BCKnockoff, we see that using the extra information from the factor structure in constructing knockoff variables can help with both FDR and power. Table 2 also shows the effects of model sparsity on the performance of various approaches. It can be seen that when the number of true signals is increased from 50 to 150, the FDR is still under control and the empirical power of IPAD remains steady.

Table 3 is devoted to the case of Design 3, where the rows of matrix X are generated independently from multivariate normal distribution with AR(1) correlation structure. This is a setting where the factor model structure in X is misspecified. Since BCKnockoff and HD-BCKnockoff make no use of the factor structure in generating knockoff variables, in both low- and high-dimensional examples both methods control FDR exactly at the target level. IPAD based methods have empirical FDR slightly over the target level, which may be caused by the misspecification of the factor structure. On the other hand, IPAD-based approaches have much higher empirical power than comparison methods.

Table 4 corresponds to Design 4 in which response y is related to X nonlinearly. Since BCKnockoff and HD-BCKnockoff are designed for linear models, only the results from IPAD method are reported. It can be seen from Table 4 that IPAD approach can control FDR with reasonably high power even in the nonlinear setting. We also observe that in nonlinear setting, the power of IPAD deteriorates faster as dimensionality p increases compared to the linear setting due to the use of the fully nonparametric approach for estimation.

5. Empirical Analysis

Our simulation results in Section 4 suggest that IPAD is a powerful approach with asymptotic FDR control. We further examine the application of IPAD to the quarterly data on 109 macroeconomic variables from the third quarter of year 1960 (1960Q3) to the fourth quarter of year 2008 (2008Q4) in the United States discussed in Stock and Watson (2012). These variables are

Table 1. Simulation results for Designs 1 and 2 of Section 4.1 with different values of dimensionality p .

	Design 1					Design 2				
	FDR	Power	FDR ₊	Power ₊	R^2	FDR	Power	FDR ₊	Power ₊	R^2
$p = 1000$										
IPAD	0.195	0.991	0.180	0.990	0.659	0.199	0.961	0.180	0.960	0.652
BCKnockoff	0.207	0.942	0.192	0.938	0.659	0.172	0.887	0.152	0.885	0.653
$p = 2000$										
IPAD	0.194	0.979	0.179	0.979	0.649	0.199	0.935	0.183	0.933	0.656
HD-BCKnockoff	0.142	0.706	0.127	0.691	0.649	0.136	0.607	0.113	0.581	0.644
$p = 3000$										
IPAD	0.191	0.964	0.176	0.963	0.652	0.188	0.913	0.171	0.911	0.658
HD-BCKnockoff	0.172	0.668	0.149	0.658	0.652	0.125	0.559	0.099	0.524	0.651

Note that FDR₊ and Power₊ are the values of FDR and Power corresponding to the knockoff+ threshold T_2 .

Table 2. Simulation results for Designs 1 and 2 of Section 4.1 with different sparsity levels s .

	Design 1					Design 2				
	FDR	Power	FDR ₊	Power ₊	R^2	FDR	Power	FDR ₊	Power ₊	R^2
$s = 50$										
IPAD	0.194	0.979	0.179	0.979	0.649	0.199	0.935	0.183	0.933	0.656
HD-BCKnockoff	0.142	0.706	0.127	0.691	0.649	0.136	0.607	0.113	0.581	0.644
$s = 100$										
IPAD	0.191	0.978	0.183	0.977	0.783	0.181	0.937	0.174	0.936	0.789
HD-BCKnockoff	0.152	0.703	0.140	0.698	0.787	0.106	0.583	0.097	0.573	0.778
$s = 150$										
IPAD	0.183	0.973	0.178	0.972	0.842	0.188	0.935	0.182	0.935	0.848
HD-BCKnockoff	0.139	0.660	0.130	0.654	0.858	0.115	0.578	0.106	0.570	0.843

Table 3. Simulation results for Design 3 of Section 4.1.

	$\rho = 0$					$\rho = 0.5$				
	FDR	Power	FDR ₊	Power ₊	R^2	FDR	Power	FDR ₊	Power ₊	R^2
$\rho = 1000$										
IPAD	0.204	0.995	0.189	0.995	0.444	0.226	0.984	0.216	0.984	0.446
BCKnockoff	0.188	0.919	0.172	0.917	0.444	0.137	0.827	0.117	0.821	0.445
$\rho = 2000$										
IPAD	0.203	0.993	0.189	0.993	0.447	0.220	0.982	0.202	0.980	0.445
HD-BCKnockoff	0.151	0.630	0.126	0.603	0.449	0.115	0.522	0.090	0.467	0.442
$\rho = 3000$										
IPAD	0.225	0.988	0.205	0.987	0.445	0.219	0.979	0.206	0.978	0.443
HD-BCKnockoff	0.150	0.589	0.126	0.560	0.446	0.092	0.439	0.064	0.381	0.447

Table 4. Simulation results for Design 4 of Section 4.1.

	$\theta = 1$					$\theta = 2$				
	FDR	Power	FDR ₊	Power ₊	R^2	FDR	Power	FDR ₊	Power ₊	R^2
$\rho = 50$										
$c = 0.1$	0.109	0.839	0.081	0.720	0.707	0.110	0.943	0.061	0.858	0.707
$c = 0.2$	0.137	0.847	0.068	0.726	0.547	0.097	0.920	0.061	0.837	0.547
$c = 0.3$	0.137	0.765	0.091	0.582	0.451	0.123	0.907	0.076	0.774	0.451
$\rho = 250$										
$c = 0.1$	0.189	0.740	0.104	0.504	0.702	0.174	0.876	0.139	0.788	0.702
$c = 0.2$	0.218	0.666	0.131	0.522	0.552	0.209	0.831	0.118	0.660	0.552
$c = 0.3$	0.200	0.569	0.101	0.361	0.451	0.224	0.766	0.141	0.599	0.451
$\rho = 500$										
$c = 0.1$	0.243	0.661	0.169	0.497	0.702	0.223	0.831	0.173	0.740	0.702
$c = 0.2$	0.204	0.507	0.111	0.266	0.543	0.216	0.749	0.126	0.594	0.543
$c = 0.3$	0.247	0.478	0.128	0.299	0.451	0.241	0.691	0.156	0.550	0.451

transformed by taking logarithms and/or differencing following Stock and Watson (2012). Our real data analysis consists of two parts. In the first part, we focus on the performance of IPAD method in terms of empirical FDR and power. To save space, the numerical results for the real data-based simulation study are presented in Section E.2 of the supplementary material. In the second part, the forecasting performance of IPAD method will be evaluated.

We now apply the IPAD approach to the real economic dataset for forecasting. One-step ahead prediction is conducted using rolling window of size 120. More specifically, one of the 109 variables is chosen as the response and the remaining 108 variables are treated as predictors. For each quarter between 1990Q3 and 2008Q4, we use the previous 120 periods for model fitting and then one-step ahead prediction is conducted based on the fitted model. We compare IPAD with the competing methods of autoregression of order one (AR(1)), factor augmented AR(1) (FAR), and Lasso, where each method is implemented in the same way as IPAD for one-step ahead prediction; see Section E.3 of the supplementary material for the implementation details of all the methods.

The number of factors \hat{r} is chosen by the PC_{p1} criterion in Bai and Ng (2002). For the Lasso and IPAD, the regularization parameter λ is selected with the tenfold cross-validation. Table 5 shows the root-mean-squared prediction error (RMSE) of these methods. As can be seen, the RMSE of IPAD is very close to those of comparison methods. To statistically compare the relative prediction accuracy of IPAD versus other approaches, we have used the Diebold–Mariano test Diebold and Mariano

Table 5. Root-mean-squared error of one-period ahead forecast of various macroeconomic variables.

	AR	FAR	Lasso	IPAD
RGDP	2.245	1.929	2.070	2.106
CPI-ALL	1.526	1.552	1.579	1.571
Imports	7.549	5.871	6.595	6.993
IP: cons dble	9.683	8.353	8.424	9.175
Emp: TTU	1.112	0.989	1.167	1.100
U: mean duration	0.573	0.487	0.502	0.494
HStarts: South	0.074	0.071	0.076	0.074
NAPM new ordrs	4.800	4.378	4.659	4.673
PCED-NDUR-ENERGY	31.927	32.121	33.546	32.164
Emp. Hours	2.102	1.899	2.080	1.944
FedFunds	0.421	0.396	0.406	0.392
Cons credit	2.573	2.537	2.648	2.580
EX rate: Canada	10.132	10.139	10.122	10.113
DJIA	23.117	23.997	24.585	23.398
Consumer expect	6.496	6.888	6.681	6.661

(1995), where the square of one-step ahead prediction error is used as the loss function. Table 6 reports the test results. The results indicate that one-step ahead prediction accuracy of IPAD is comparable to other approaches.

It is worth mentioning that one main advantage of IPAD is its interpretability and stability. Using IPAD for forecasting, we not only enjoy the same level of accuracy as other methods but also obtain the information on variable importance with stability. Recall that for each one-step ahead prediction, we apply IPAD 100 times and obtain 100 sets of selected variables. Thus, we can calculate the selection frequency of each variable in each one-step ahead prediction. Figure 1 depicts the frequencies of top five

Table 6. Diebold–Mariano test for comparing prediction accuracy of IPAD against other procedures.

	IPAD vs. AR	IPAD vs. FAR	IPAD vs. Lasso
RGDP	-0.780	1.160	0.462
CPI-ALL	0.521	0.394	-0.218
Imports	-0.976	2.631**	1.464
IP: cons dble	-1.026	1.567	2.487*
Emp: TTU	-0.140	1.692	-1.845
U: mean duration	-3.383***	0.672	-0.505
HStarts: South	0.096	0.821	-0.766
NAPM new ordrs	-0.517	1.814	0.076
PCED-NDUR-ENERGY	0.753	0.049	-1.759
Emp. Hours	-1.200	0.297	-2.063*
FedFunds	-0.971	-0.134	-0.625
Cons credit	0.207	0.359	-0.661
EX rate: Canada	-0.466	-0.138	-0.037
DJIA	0.585	-0.959	-1.428
Consumer expect	1.212	-1.038	-0.277

selected variables in predicting real GDP growth before and after year 2000, where the variable importance is ranked according to the aggregated frequencies over the entire time period before or after 2000. We have experimented with different cut-off years around year 2000, and the top five-ranked variables stay the same so only the results corresponding to cut-off year 2000 are reported. Changes in index of help wanted advertising in newspapers, percentages of changes in real personal consumption of services, and percentage of changes in real gross private domestic investment in residential sector were the top three important variables in predicting real GDP growth during the whole period. It is interesting to see that percentage of changes in residential price index was among top five important variables in predicting GDP growth during the 1990s, and then starting from year 2000 it was replaced by changes in index of consumer expectations about stability of economy. Moreover, it is also seen that the percentage of changes in industrial production of fuels was of great importance for predicting real GDP growth during some periods but not the others.

As a comparison, it is very difficult to interpret the results of FAR. As for the Lasso-based method, there is no theoretical guarantee on FDR control and in addition, Lasso usually gives us models with much larger size. For instance, in predicting real GDP growth, IPAD on average selects 5.42 macroeconomic variables while Lasso on average selects 13.32 variables. To summarize, our real data analysis indicates that IPAD is an applicable approach for controlling FDR with competitive prediction power and high interpretability and stability.

6. Discussions

We have suggested in this article a new procedure IPAD for feature selection in high-dimensional linear models that achieves asymptotic FDR control while retaining high power. Our model setting involves a latent factor model that is motivated by applications in economics and finance. Our method falls into the general model-X knockoffs framework in Candès et al. (2018), but allows the unknown covariate distribution for the knockoff variable construction. With the LCD knockoff statistics, we have shown that the FDR of IPAD can be asymptotically under control while the power can be asymptotically the same as that of Lasso. Our simulation study and empirical analysis also suggest

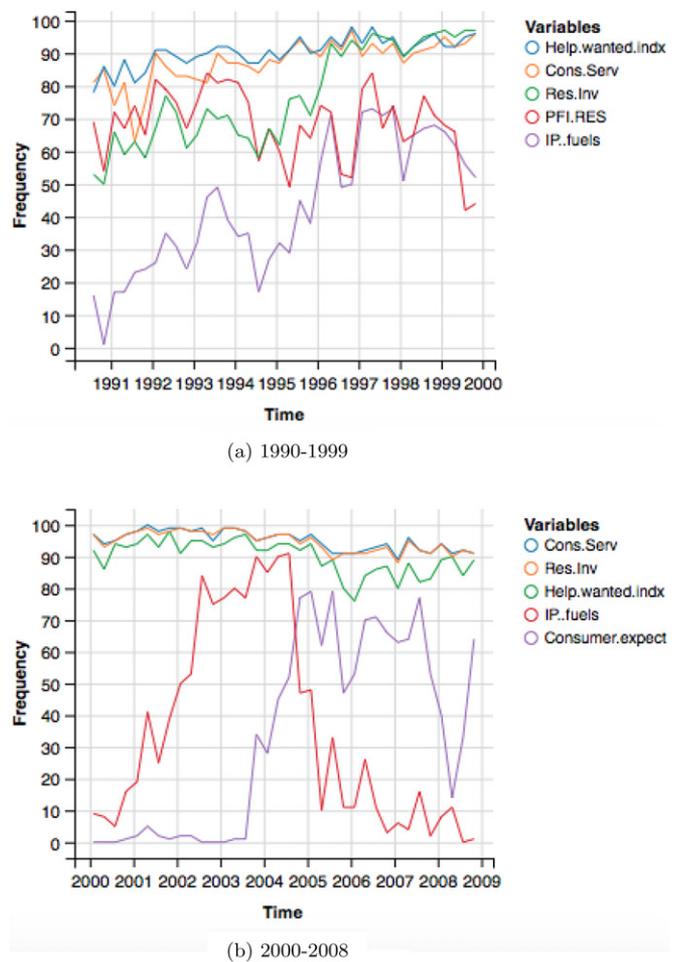


Figure 1. Frequencies of top selected variables in predicting real GDP growth. The set of selected variables are index of help-wanted advertising in newspapers (Help wanted indx), real personal consumption expenditures - services (Cons-Serv), real gross private domestic investment - residential (Res.Inv), residential price index (PFI-RES), industrial production index - fuels (IP:fuels), and University of Michigan index of consumer expectations (Consumer expect).

that IPAD has highly competitive performance compared to many widely used forecasting methods such as Lasso and FAR, but with much higher interpretability and stability.

Our work has focused on the scenario of static models. It would be interesting to extend the IPAD procedure to high-dimensional dynamic models with time series data. It is also interesting to consider nonlinear models and more flexible machine learning methods for forecasting as well as more refined factor model structures on the covariates for the knockoffs inference with IPAD, and develop theoretical guarantees for the IPAD framework in these more general model settings. These extensions are beyond the scope of the current article and are interesting topics for future research.

Appendix: Proofs of Main Results

We provide the proofs of Theorems 1 and 2 in this appendix. The proofs of Proposition 1 and Lemmas 1–2 and additional technical details are included in the supplementary material.

To ease the technical presentation, let us introduce some notation. We denote by \lesssim the inequality up to some positive constant factor. Restricting the columns of \mathbf{X} and $\tilde{\mathbf{X}}(\hat{\theta})$ to the variables in

index set \mathcal{A} such that $|\mathcal{A}| \leq k$, we obtain the $n \times k$ submatrices $\mathbf{X}_{\mathcal{A}}$ and $\tilde{\mathbf{X}}_{\mathcal{A}}(\hat{\theta})$, respectively. Moreover, we define $\mathbf{T}_{\mathcal{A}}(\hat{\theta}) := \text{vec}(\text{vech } \mathbf{U}_{\mathcal{A}}(\hat{\theta}), \mathbf{v}_{\mathcal{A}}(\hat{\theta})) \in \mathbb{R}^{k(2k+3)}$ with $\mathbf{U}_{\mathcal{A}}(\hat{\theta})$ being the principle submatrix of $\mathbf{U}(\hat{\theta})$ formed by columns and rows in set \mathcal{A} , and $\mathbf{v}_{\mathcal{A}}(\hat{\theta})$ the subvector of $\mathbf{v}(\hat{\theta})$ formed by components in set \mathcal{A} . Then it is easy to see that $\mathbf{U}_{\mathcal{A}}(\hat{\theta}) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\hat{\theta})]'[\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\hat{\theta})]$ and $\mathbf{v}_{\mathcal{A}}(\hat{\theta}) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \tilde{\mathbf{X}}_{\mathcal{A}}(\hat{\theta})]'\mathbf{y}$. For the oracle factor loading matrix Λ^0 , with a slight abuse of notation, we use $\Lambda_{\mathcal{A}}^0$ to denote the row restricted to the variables in \mathcal{A} for notational convenience. Recall that $\nu > 0$ is a fixed positive number, $c_{np} = (p^{-1} \log n)^{1/2} + (n^{-1} \log p)^{1/2}$, and $\tilde{c}_{np} = p^{-1/2} \log n + n^{-1/2} \log p$. We define $\pi_{np} = n^{-\nu} + p^{-\nu}$. Since λ is fixed at $C_0 n^{-1/2} \log p$, in all the proofs we will drop the dependence of various quantities on λ whenever there is no confusion.

Proof of Theorem 1

Recall that for a given θ , $\mathcal{A}^*(\theta)$ is the support of knockoff statistics $(W_1(\theta), \dots, W_p(\theta))'$. Define set $\hat{\mathcal{A}}(\hat{\theta}) := \mathcal{A}^*(\hat{\theta}) \cup \mathcal{A}^*(\theta^0)$. It follows from (12) that the cardinality of $\hat{\mathcal{A}}(\hat{\theta})$ is bounded by k . Hereafter we write $\hat{\mathcal{A}}(\hat{\theta})$ as $\hat{\mathcal{A}}$ for notational simplicity.

By Lemmas 1–2 and the definition of the FDP, we know that $S_{\{1, \dots, p\}}(\mathbf{T}(\hat{\theta})) = S_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta}))$ and thus the resulting FDRs are the same. Therefore, we can restrict ourselves to the smaller model $\hat{\mathcal{A}}$ when studying the FDR of IPAD. The same arguments as above also hold for the oracle knockoffs; that is, the FDR of IPAD applied to $\mathbf{T}(\theta^0)$ is the same as that applied to $\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)$. Note that all the FDRs we discuss here are with respect to the full model $\{1, \dots, p\}$. For this reason, in what follows we will abuse the notation and use $\text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta}))$ and $\text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0))$ to denote the FDR of IPAD based on $\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta})$ and $\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)$, respectively. We want to emphasize that although we put a subscript $\hat{\mathcal{A}}$ in FDRs, their values are still deterministic as argued above. Summarizing the facts, we obtain

$$\begin{aligned} \text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta})) &= \text{FDR}_{\{1, \dots, p\}}(\mathbf{T}(\hat{\theta})), \\ \text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)) &= \text{FDR}_{\{1, \dots, p\}}(\mathbf{T}(\theta^0)). \end{aligned}$$

Meanwhile, by construction $\tilde{\mathbf{X}}(\theta^0)$ satisfies the two properties in Definition 1 and is a valid model-X knockoffs matrix. Therefore, for any value of the regularization parameter, the LCD statistics $W_j(\theta^0)$ based on $([\mathbf{X}, \tilde{\mathbf{X}}(\theta^0)], \mathbf{y})$ together with Result 1 in Supplementary Material ensure the exact FDR control at some target level $q \in (0, 1)$. Summarizing this, we obtain that the FDR of IPAD applied to $\mathbf{T}(\theta^0)$ is controlled at target level q .

Combining the arguments in the previous two paragraphs, we deduce

$$\text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)) = \text{FDR}_{\{1, \dots, p\}}(\mathbf{T}(\theta^0)) \leq q.$$

Thus, the desired results follow automatically if we can prove that $\text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta}))$ is asymptotically close to $\text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0))$. We next proceed to prove it.

Recall the definitions of \mathbb{I} and $\mathbb{I}_{\mathcal{A}}$ as in (13). Define the event

$$\mathcal{E}_{np} = \left\{ \mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta}) \in \mathbb{I}_{\hat{\mathcal{A}}} \right\} \cap \left\{ \mathbf{T}_{\hat{\mathcal{A}}}(\theta^0) \in \mathbb{I}_{\hat{\mathcal{A}}} \right\}.$$

Lemma 3 in Section C.4 establishes $\hat{\theta} \in \Theta_{np}$ with probability at least $1 - O(\pi_{np})$ and $\theta^0 \in \Theta_{np}$. Hence, Lemma 4 in Section C.5 guarantees that

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{np}^c) &\leq 2 \mathbb{P} \left(\sup_{|\mathcal{A}| \leq k, \theta \in \Theta_{np}} \left\| \mathbf{T}_{\mathcal{A}}(\theta) - \mathbb{E}[\mathbf{T}_{\mathcal{A}}(\theta^0)] \right\|_{\max} > a_{np} \right) \\ &= O(\pi_{np}), \end{aligned} \tag{A.1}$$

where $a_{np} = C_1(k^{1/2} + s^{3/2})\tilde{c}_{np}$ for some constant $C_1 > 0$.

For a given deterministic set $\mathcal{A} \subset \{1, \dots, p\}$, let $\text{FDP}_{\mathcal{A}}(\cdot)$ be the FDP function corresponding to $\text{FDR}_{\mathcal{A}}(\cdot)$. By the definition of FDP function, we have for any $\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{R}^{|\mathcal{A}|(2|\mathcal{A}|+3)}$,

$$\begin{aligned} \text{FDP}_{\mathcal{A}}(\mathbf{t}_2) - \text{FDP}_{\mathcal{A}}(\mathbf{t}_1) &= \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)|}{|S_{\mathcal{A}}(\mathbf{t}_2)|} - \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_1)|} \\ &= \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)| \cdot (|S_{\mathcal{A}}(\mathbf{t}_1)| - |S_{\mathcal{A}}(\mathbf{t}_2)|)}{|S_{\mathcal{A}}(\mathbf{t}_1)| \cdot |S_{\mathcal{A}}(\mathbf{t}_2)|} \\ &\quad + \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)| - |\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_1)|}. \end{aligned}$$

Further, note that

$$\begin{aligned} |\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)|/|S_{\mathcal{A}}(\mathbf{t}_2)| &\leq 1, \\ ||S_{\mathcal{A}}(\mathbf{t}_2)| - |S_{\mathcal{A}}(\mathbf{t}_1)|| &\leq |S_{\mathcal{A}}(\mathbf{t}_2) \Delta S_{\mathcal{A}}(\mathbf{t}_1)|, \\ \left| |\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)| - |\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_1)| \right| &\leq \left| \{S_{\mathcal{A}}(\mathbf{t}_2) \Delta S_{\mathcal{A}}(\mathbf{t}_1)\} \cap \mathcal{S}^1 \right|. \end{aligned}$$

Combining the results above yields

$$\begin{aligned} |\text{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \text{FDP}_{\mathcal{A}}(\mathbf{t}_2)| &\leq \frac{||S_{\mathcal{A}}(\mathbf{t}_1)| - |S_{\mathcal{A}}(\mathbf{t}_2)||}{|S_{\mathcal{A}}(\mathbf{t}_1)|} + \frac{|\{S_{\mathcal{A}}(\mathbf{t}_2) \Delta S_{\mathcal{A}}(\mathbf{t}_1)\} \cap \mathcal{S}^1|}{|S_{\mathcal{A}}(\mathbf{t}_1)|} \\ &\leq 2 \frac{|S_{\mathcal{A}}(\mathbf{t}_2) \Delta S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_1)|}. \end{aligned}$$

Similarly, we have

$$|\text{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \text{FDP}_{\mathcal{A}}(\mathbf{t}_2)| \leq 2 \frac{|S_{\mathcal{A}}(\mathbf{t}_2) \Delta S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_2)|}.$$

Thus, it holds that

$$\begin{aligned} \sup_{|\mathcal{A}| \leq k} \sup_{\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{I}_{\mathcal{A}}} |\text{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \text{FDP}_{\mathcal{A}}(\mathbf{t}_2)| &\leq \sup_{|\mathcal{A}| \leq k} \sup_{\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{I}_{\mathcal{A}}} \frac{|S_{\mathcal{A}}(\mathbf{t}_2) \Delta S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_1) \wedge |S_{\mathcal{A}}(\mathbf{t}_2)||} = O(\rho_{np}), \end{aligned} \tag{A.2}$$

where the last two steps are due to Condition 6. Therefore, (A.1) and (A.2) together with the fact that $\text{FDP}(\cdot) \in [0, 1]$ entail that

$$\begin{aligned} &\left| \text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta})) - \text{FDR}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)) \right| \\ &= \left| \mathbb{E} \text{FDP}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta})) - \mathbb{E} \text{FDP}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)) \right| \\ &\leq \mathbb{E} \left| \text{FDP}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta})) - \text{FDP}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)) \right| \\ &\leq \mathbb{E} \left[\left| \text{FDP}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\hat{\theta})) - \text{FDP}_{\hat{\mathcal{A}}}(\mathbf{T}_{\hat{\mathcal{A}}}(\theta^0)) \right| \mid \mathcal{E}_{np} \right] \\ &\quad + \mathbb{P}(\mathcal{E}_{np}^c) + 2 \mathbb{P}(\mathcal{E}_{np}^c) \\ &\leq \sup_{|\mathcal{A}| \leq k} \sup_{\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{I}_{\mathcal{A}}} |\text{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \text{FDP}_{\mathcal{A}}(\mathbf{t}_2)| + O(\pi_{np}) \\ &= O(\rho_{np}) + O(\pi_{np}). \end{aligned}$$

This completes the proof of Theorem 1.

Proof of Theorem 2

By the definition of the LCD statistics, we construct the augmented Lasso estimator for each $\theta \in \Theta_{np}$, which is defined as

$$\hat{\beta}^{\text{aug}}(\theta) = \arg \min_{\mathbf{b} \in \mathbb{R}^{2p}} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}(\theta)]\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1. \tag{A.3}$$

The Lasso estimator of regressing \mathbf{y} on only \mathbf{X} is also given by

$$\hat{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1, \tag{A.4}$$

where $\lambda = O(n^{-1/2} \log p)$. According to the true model \mathcal{S}^0 , the underlying true parameter vector corresponding to $\hat{\beta}^{\text{aug}}(\theta)$ should be given by $\beta^{\text{aug}} := (\beta', \mathbf{0}')' \in \mathbb{R}^{2p}$ with $\beta = (\beta'_{\mathcal{S}^0}, \mathbf{0}')' \in \mathbb{R}^p$ and $|\mathcal{S}^0| = s$ for any $\theta \in \Theta_{np}$. By Lemma 5 in Section C.6, with probability at least $1 - O(\pi_{np})$ the Lasso estimators satisfy

$$\sup_{\theta \in \Theta_{np}} \|\hat{\beta}^{\text{aug}}(\theta) - \beta^{\text{aug}}\|_1 = O(s\lambda), \quad \|\hat{\beta} - \beta\|_1 = O(s\lambda),$$

where $\lambda = O(n^{-1/2} \log p)$.

We now prove that under Condition 7, the power of the augmented Lasso (A.3) is bounded from below by $\gamma \in [0, 1]$; that is,

$$\mathbb{E} \left| \widehat{\mathcal{S}}_{\text{auglasso}} \cap \mathcal{S}^0 \right| / s \geq \gamma, \tag{A.5}$$

where $\widehat{\mathcal{S}}_{\text{auglasso}} = \{j : \hat{\beta}_j^{\text{aug}}(\theta) \neq 0\}$. To this end, we first show that with asymptotic probability one,

$$|\widehat{\mathcal{S}}_{\text{auglasso}}^c \cap \mathcal{S}^0| / s \leq 1 - \gamma. \tag{A.6}$$

The key is to use proof by contradiction. Suppose $|\widehat{\mathcal{S}}_{\text{auglasso}}^c \cap \mathcal{S}^0| / s > 1 - \gamma$. Then we can see that

$$\begin{aligned} \sup_{\theta \in \Theta_{np}} \|\hat{\beta}^{\text{aug}}(\theta) - \beta^{\text{aug}}\|_1 &\geq \sup_{\theta \in \Theta_{np}} \left\| \hat{\beta}_{\widehat{\mathcal{S}}_{\text{auglasso}}^c}^{\text{aug}}(\theta) - \beta_{\widehat{\mathcal{S}}_{\text{auglasso}}^c}^{\text{aug}} \right\|_1 \\ &= \left\| \beta_{\widehat{\mathcal{S}}_{\text{auglasso}}^c}^{\text{aug}} \right\|_1 \geq \left\| \beta_{\widehat{\mathcal{S}}_{\text{auglasso}}^c \cap \mathcal{S}^0}^{\text{aug}} \right\|_1 \\ &> b_{np} s n^{-1/2} \log p, \end{aligned}$$

where the last step is by Condition 7. However, by Lemma 5 with probability at least $1 - O(\pi_{np})$, the left-hand side above is bounded from above by $O(s\lambda)$ with $\lambda = O(n^{-1/2} \log p)$. These two results contradict with each other since $b_{np} \rightarrow \infty$. Hence, (A.6) is proved. Therefore, the result in (A.5) follows immediately since $|\widehat{\mathcal{S}}_{\text{auglasso}} \cap \mathcal{S}^0| = s - |\widehat{\mathcal{S}}_{\text{auglasso}}^c \cap \mathcal{S}^0|$ and

$$\begin{aligned} \mathbb{E} \left| \widehat{\mathcal{S}}_{\text{auglasso}} \cap \mathcal{S}^0 \right| / s &\geq \gamma \mathbb{P} \left(|\widehat{\mathcal{S}}_{\text{auglasso}} \cap \mathcal{S}^0| / s > \gamma \right) \\ &= \gamma \mathbb{P} \left(|\widehat{\mathcal{S}}_{\text{auglasso}}^c \cap \mathcal{S}^0| / s \leq 1 - \gamma \right) = \gamma(1 - O(\pi_{np})). \end{aligned}$$

Let $\widehat{\mathcal{S}}_{\text{lasso}} = \{j : \hat{\beta}_j \neq 0\}$. Using the same argument, we can show that the power of the Lasso (A.4) is also bounded from below by $\gamma(1 - O(\pi_{np}))$ under Condition 7. That is, we have

$$\mathbb{E} \left| \widehat{\mathcal{S}}_{\text{lasso}} \cap \mathcal{S}^0 \right| / s \geq \gamma(1 - O(\pi_{np})).$$

Next, we show that our knockoffs procedure has at least the same power as the augmented Lasso and hence the Lasso itself. Namely we prove

$$\mathbb{E} \left| \widehat{\mathcal{S}} \cap \mathcal{S}^0 \right| / s \geq \gamma \tag{A.7}$$

with threshold T_2 . Note that the same argument is still valid for T_1 . Let $|W_{(1)}| \geq \dots \geq |W_{(p)}|$ and define j^* as $|W_{(j^*)}| = T_2$. Then by the definition of T_2 , it holds that $-T_2 < W_{j^*+1} \leq 0$. Here, we have assumed that there are no ties on the magnitudes of W_j 's which should be a reasonable assumption considering the continuity of the

Lasso solution. As in the proof of Theorem 3 in Fan et al. (2019), it is sufficient to consider the following two cases.

Case 1. Consider the case of $-T_2 < W_{(j^*+1)} < 0$. In this case, from the definition of threshold T_2 , we have

$$\frac{2 + |\{j : W_{(j)} \leq -T_2\}|}{|\{j : W_{(j)} \geq T_2\}|} > q.$$

Using the same argument as in Lemma 6 of Fan et al. (2019) together with Lemma 5, we can prove from Condition 8 that $|\widehat{\mathcal{S}}| \geq C_2 s$ with probability at least $1 - O(\pi_{np})$. This leads to $|\{j : W_{(j)} \leq -T_2\}| > C_2 q s - 2$ with the same probability. Now from the same argument as in A.5 of Fan et al. (2019), we can obtain $T_2 = O(\lambda)$. On the other hand, Lemma 5 and some algebra establish that

$$\begin{aligned} O(s\lambda) &= \|\hat{\beta}^{\text{aug}}(\hat{\theta}) - \beta^{\text{aug}}\|_1 = \sum_{j=1}^p |\hat{\beta}_j^{\text{aug}}(\hat{\theta}) - \beta_j| + \sum_{j=1}^p |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| \\ &= \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_j^{\text{aug}}(\hat{\theta}) - \beta_j| + \sum_{j \in \mathcal{S}^1} |\hat{\beta}_j^{\text{aug}}(\hat{\theta})| \\ &\quad + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\text{aug}}(\hat{\theta}) - \beta_j| + \sum_{j=1}^p |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})|. \end{aligned} \tag{A.8}$$

We then consider the lower bound of the last term in (A.8). For any $j \in \widehat{\mathcal{S}}^c$, it holds that $|\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| > |\hat{\beta}_j^{\text{aug}}(\hat{\theta})| - T_2$. Hence, we obtain

$$\begin{aligned} \sum_{j=1}^p |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| &\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| \\ &\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\text{aug}}(\hat{\theta})| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0|. \end{aligned} \tag{A.9}$$

Plugging (A.9) into (A.8) and applying the triangle inequality yield

$$\begin{aligned} O(s\lambda) &\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_j^{\text{aug}}(\hat{\theta}) - \beta_j| + \sum_{j \in \mathcal{S}^1} |\hat{\beta}_j^{\text{aug}}(\hat{\theta})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\text{aug}}(\hat{\theta}) - \beta_j| \\ &\quad + \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\text{aug}}(\hat{\theta})| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| \\ &\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_j^{\text{aug}}(\hat{\theta}) - \beta_j| + \sum_{j \in \mathcal{S}^1} |\hat{\beta}_j^{\text{aug}}(\hat{\theta})| \\ &\quad + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\beta_j| + \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\text{aug}}(\hat{\theta})| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| \\ &\geq \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\beta_j| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| = \|\beta_{\widehat{\mathcal{S}}^c \cap \mathcal{S}^0}\|_1 - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0|. \end{aligned}$$

Since $T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| = O(s\lambda)$ for $\lambda = O(n^{-1/2} \log p)$ due to the discussion above, we obtain

$$\|\beta_{\widehat{\mathcal{S}}^c \cap \mathcal{S}^0}\|_1 = O(s n^{-1/2} \log p). \tag{A.10}$$

Suppose $|\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| / s > 1 - \gamma$. Then Condition 7 gives $\|\beta_{\widehat{\mathcal{S}}^c \cap \mathcal{S}^0}\|_1 > b_{np} s n^{-1/2} \log p$ for some positive diverging sequence b_{np} ; this contradicts with (A.10). Thus, we obtain $|\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| / s \leq 1 - \gamma$ with asymptotic probability one, which leads to (A.7) by taking expectation.

Case 2. Consider the case of $W_{(j^*+1)} = 0$. In this case, by the definition of threshold T_2

$$\frac{1 + |\{j : W_{(j)} < 0\}|}{|\{j : W_{(j)} > 0\}|} \leq q. \tag{A.11}$$

If $|\{j : W_{(j)} < 0\}| > C_3 s$ for some constant $C_3 > 0$, then from the same argument as in A.5 of Fan et al. (2019), we can obtain $T_2 = O(\lambda)$, and the rest of the proof is the same as in Case 1. On the other hand, if $|\{j : W_{(j)} < 0\}| \leq o(s)$ we have

$$\begin{aligned} |\{j : W_{(j)} \neq 0\} \cap S^0| &= |\{j : W_{(j)} > 0\} \cap S^0| + |\{j : W_{(j)} < 0\} \cap S^0| \\ &\leq |\widehat{S} \cap S^0| + o(s). \end{aligned}$$

Now note that $|\{j : W_{(j)} \neq 0\}| \geq |\{j : |\hat{\beta}_j^{\text{aug}}| \neq 0, j = 1, \dots, p\}|$. Then we can see that with asymptotic probability one,

$$\begin{aligned} |\{j : W_{(j)} \neq 0\} \cap S^0| &\geq |\{j : \hat{\beta}_j^{\text{aug}} \neq 0, j = 1, \dots, p\} \cap S^0| \\ &= |\widehat{S}_{\text{auglasso}} \cap S^0| \geq \gamma s(1 - o(1)). \end{aligned}$$

Consequently, we obtain $|\widehat{S} \cap S^0|/s \geq \gamma(1 - o(1))$, which leads to (A.7) by taking expectation. Combining these two cases concludes the proof of Theorem 2.

Acknowledgments

The author names are alphabetically ordered. Most of this work was completed while Uematsu visited USC as a JSPS Overseas Research Fellow and Postdoctoral Scholar. The authors sincerely thank the Joint Editor, Associate Editor, and referees for their valuable comments that helped improve the article substantially.

Supplementary Materials

Supplementary materials are available at Journal of the American Statistical Association online.

Funding

This work was supported by NIH (Grant 1R01GM131407-01), NSF (CAREER Award DMS-1150318), a grant from the Simons Foundation, Adobe Data Science Research Award, and a Grant-in-Aid for JSPS Overseas Research Fellowship 29-60.

References

- Ahn, S. C., and Horenstein, A. R. (2013), "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81, 1203–1227. [1825,1826]
- Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [1825,1826]
- Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [1825,1826,1830]
- Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate Via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [1823,1828]
- (2016), "A Knockoff Filter for High-dimensional Selective Inference," arXiv no. 1602.03574. [1823,1828]
- Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018), "High-dimensional Econometrics and Regularized GMM," arXiv no. 1806.01888. [1823]
- Benjamini, Y. (2010), "Discovering the False Discovery Rate," *Journal of the Royal Statistical Society, Series B*, 72, 405–416. [1823]
- 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*, 57, 289–300. [1823]
- Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [1823]
- Billingsley, P. (1995), *Probability and Measure* (3rd ed.), New York: Wiley. [1826]
- Bonferroni, C. E. (1935), "Il Calcolo delle Assicurazioni su Gruppi di Teste," *Studi in Onore del Professore Salvatore Ortu Carboni*, 13–60. [1823]
- Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1828]
- Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: Model X' Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society, Series B*, 80, 551–577. [1823,1824,1825,1826,1829,1831]
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21, C1–C68. [1822]
- Chernozhukov, V., Härdle, W. K., Huang, C., and Wang, W. (2018), "Lasso-driven Inference in Time and Space," arXiv no. 1806.05081. [1822]
- Chernozhukov, V., Newey, W., and Robins, J. (2018), "Double/De-biased Machine Learning Using Regularized Riesz Representers," arXiv no. 1802.08667. [1822]
- Chudik, A., Kapetanios, G., and Pesaran, H. (2018), "A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-dimensional Linear Regression Models," *Econometrica*, 86, 1479–1512. [1823]
- De Mol, C., Giannone, D., and Reichlin, L. (2008), "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?" *Journal of Econometrics*, 146, 318–328. [1822]
- Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 20, 134–144. [1830]
- Fan, J., and Y. Fan (2008), "High-dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637. [1823]
- Fan, J., Han, X., and Gu, W. (2012), "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence" (with discussion), *Journal of American Statistical Association*, 107, 1019–1045. [1823]
- Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society, Series B*, 70, 849–911. [1823]
- Fan, Y., Demirkaya, E., Li, G., and Lv, J. (2019), "RANK: Large-scale Inference With Graphical Nonlinear Knockoffs," *Journal of the American Statistical Association*, to appear. [1828,1833,1834]
- Fan, Y., Demirkaya, E., and Lv, J. (2019), "Nonuniformity of p-values Can Occur Early in Diverging Dimensions," *Journal of Machine Learning Research*, 20, 1–33. [1822]
- Fan, Y., and Lv, J. (2013), "Asymptotic Equivalence of Regularization Methods in Thresholded Parameter Space," *Journal of the American Statistical Association*, 108, 1044–1061. [1826]
- Guo, Z., Kang, H., Cai, T. T., and Small, D. S. (2018), "Confidence Intervals for Causal Effects With Invalid Instruments by Using Two-stage Hard Thresholding With Voting," *Journal of the Royal Statistical Society, Series B*, 80, 793–815. [1822]
- Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70. [1823]
- Lv, J. (2013), "Impacts of High Dimensionality in Finite Samples," *The Annals of Statistics*, 41, 2236–2262. [1826]
- Romano, J. P., and Wolf, M. (2005), "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100, 94–108. [1823]
- Shah, R. D., and Bühlmann, P. (2018), "Goodness-of-fit Tests for High Dimensional Linear Models," *Journal of the Royal Statistical Society, Series B*, 80, 113–135. [1822]
- Stock, J. H., and Watson, M. W. (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business & Economic Statistics*, 30, 481–493. [1829,1830]
- Stucky, B., and van de Geer, S. (2018), "Asymptotic Confidence Regions for High-dimensional Structured Sparsity," *IEEE Transactions on Signal Processing*, 66, 2178–2190. [1822]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of Royal Statistical Society, Series B*, 58, 267–288. [1825]
- Vizcarra, A. B., and Viens, F. G. (2007), "Some Applications of the Malliavin Calculus to Sub-Gaussian and Non-Sub-Gaussian Random Fields," in *Seminar on Stochastic Analysis, Random Fields and Applications V*, eds. R. C. Dalang, M. Dozzi, and F. Russo, Birkhäuser, Basel, pp. 363–395. [1826]
- Wooldridge, J. M., and Zhu, Y. (2018), "Inference in Approximately Sparse Correlated Random Effects Probit Models," *Journal of Business & Economic Statistics*, to appear. [1822]
- Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-dimensional Linear Models," *Journal of the American Statistical Association*, 112, 757–768. [1822]