Electronic Journal of Statistics
Vol. 11 (2017) 3251–3271
ISSN: 1935-7524
DOI: 10.1214/17-EJS1284

Oracle P-values and variable screening

Ning Hao and Hao Helen Zhang*

Department of Mathematics, University of Arizona

Abstract: The concept of P-value was proposed by Fisher to measure inconsistency of data with a specified null hypothesis, and it plays a central role in statistical inference. For classical linear regression analysis, it is a standard procedure to calculate P-values for regression coefficients based on least squares estimator (LSE) to determine their significance. However, for high dimensional data when the number of predictors exceeds the sample size, ordinary least squares are no longer proper and there is not a valid definition for P-values based on LSE. It is also challenging to define sensible P-values for other high dimensional regression methods such as penalization and resampling methods. In this paper, we introduce a new concept called *oracle P-value* to generalize traditional P-values based on LSE to high dimensional sparse regression models. Then we propose several estimation procedures to approximate oracle P-values for real data analysis. We show that the oracle P-value framework is useful for developing new and powerful tools to enhance high dimensional data analysis, including variable ranking, variable selection, and screening procedures with false discovery rate (FDR) control. Numerical examples are then presented to demonstrate performance of the proposed methods.

Keywords and phrases: False discovery rate, high dimensional data, inference, P-value, variable selection.

Received December 2014.

1. Introduction

Many contemporary data are featured with high dimensionality. Given a set of observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a predictor and y_i is a response, the dimension p is oftentimes comparable to or much larger than the sample size n. For high dimensional regression models, a large number of methodologies have recently been developed for model selection, coefficient estimation, and prediction; see Fan & Lv (2010) and Bühlmann & van de Geer (2011) for a comprehensive overview.

Let $\mathcal{S}^*$ denote the true model and $\hat{\mathcal{S}}$ denote the selected model by a procedure. In general, $\mathcal{S}^*$ is assumed to be sparse for high dimensional data. Depending on the signal-to-noise ratio level, there are three types of variable selection results. When the signal is strong enough, it is possible to achieve model selection consistency provided some assumptions on the design matrix; see Zhao & Yu (2006); Fan & Lv (2011); Zhang (2010), among others. If the signal is relatively strong but there are too many noise variables, a weaker result known as screening consistency is achievable (Fan & Lv, 2008; Wang, 2009). However, for many data sets in high-throughput sciences, some signals are too weak to be distinguished

*Corresponding author.

from many trivial effects. In this case, it might be more realistic to give up screening consistency and allow some false negatives. A practical approach is to control false positives via the family-wise error rate (FWER), false discovery rate (FDR), or false discovery proportion (FDP). See Table 1 for a list of high-dimensional variable selection results.

TABLE 1
Three types of variable selection results. Here $\mathcal{S}^*$ is the true model, $\hat{\mathcal{S}}$ is the selected model, and n is the sample size.

Model selection consistency	$P(\hat{\mathcal{S}} = \mathcal{S}^*) \to 1$
Screening consistency	$P(\hat{\mathcal{S}} \supset \mathcal{S}^*) \to 1$, $\lvert\hat{\mathcal{S}}\rvert \ll n$
FDR-type error control	control $1 - \frac{\lvert\hat{\mathcal{S}} \cap \mathcal{S}^*\rvert}{\lvert\hat{\mathcal{S}}\rvert}$ or similar quantities

To access FDR-type error rates, it is necessary to assign a significance level for each variable, as in the context of multiple hypothesis testing. For traditional linear models, P-values are well defined based on least squares estimator (LSE) (Bauer et al., 1988; Bunea et al., 2006). In the classical low-dimensional setup, it is known that P-values follow Unif$[0, 1]$ under the null hypothesis. However, it is hard to define valid P-values in high dimensional situations, and the topic has not been studied until recent emerging interests in assigning uncertainties, significance, or confidence to covariate effects. The difficulty is two-fold. First, the ordinary LSE is not well defined for high dimensional models. Second, the distributions of modern penalized least squares estimators such as the LASSO (Tibshirani, 1996) are too complex to access. In the literature, there are some recent proposals of P-values for high dimensional linear regression, including the screen-and-clean approach (Wasserman & Roeder, 2009), the multi-split approach (Meinshausen et al., 2009), and the low dimensional projection (LDP) approach (Zhang & Zhang, 2014; Bühlmann et al., 2013), and a recent work on hypothesis tests for generic penalized M-estimators (Ning & Liu, 2016). In particular, the screen-and-clean approach has three stages: fit a set of candidate models, select one model by cross validation, and then use classical hypothesis testing to eliminate some variables. The procedure involves spitting the data into two disjoint parts, with one used for variable selection and the other used for significance testing. The multi-split approach of Meinshausen et al. (2009) extends the screen-and-clean approach in the following way: randomly split the data into two parts, use one part for variable selection and the other part for fitting LSE and calculating P-values for the selected variables, repeat the splitting-and-fitting step multiple times and calculate P-values based on each split, and aggregate the P-values over multiple splits. The P-values are further adjusted by multiplying a suitable constant to control the FWER. Meinshausen et al. (2009) shows that the final multiplicity-adjusted P-values can asymptotically provide the FWER control. The LDP approach is proposed to construct confidence intervals for regression coefficients in high-dimensional situations, which focuses on a relevant question rather than on directly defining valid P-values. Another seminal work is Fan et al. (2012b) which discusses the issue of FDP control based on marginal regression models instead of joint linear regression models.

One main advantage of their method is that the known covariance structure of P-values can be fully used to improve the accuracy of FDP control. Besides these pioneering works, there have been many breakthroughs in this field. We refer to Dezeure et al. (2015) for a review of recent works. In spite of these developments, it seems that a natural property of P-value is often ignored. That is, P-value under null hypothesis follows a standard uniform distribution. Many "P-values" proposed in recent literatures on high-dimensional inference do not satisfy this fundamental property, which may cause conceptual difficulty to understand the meaning of P-values.

In this paper, we propose a new concept called "oracle P-value" to quantify significance of each predictor in high dimensional regression. Compared to existing works, the oracle P-value framework has both theoretical and computational advantages. Theoretically, the oracle P-values follow Unif$[0, 1]$ under the null hypothesis, which is the same as classical results for low-dimensional situations. Moreover, the covariance structure of oracle P-values are completely known, so the FDR-type quantities can be better controlled. Computationally, the estimation of oracle P-values is simple and fast, as it does not involve multiple data split and model fitting. The basic idea is described as follows. Assume the true model $\mathcal{S}^*$ is sparse and its model size is smaller than n. We define the oracle P-value for the jth variable via LSE by fitting the model $\mathcal{S}^* \cup \{j\}$. The oracle P-value is not available in practice without knowing $\mathcal{S}^*$, but one can obtain an estimated model $\hat{\mathcal{S}}$ by existing state-of-the-art model selection or screening procedures. We point out that, as long as $\hat{\mathcal{S}}$ is a reasonable estimator of $\mathcal{S}^*$, oracle P-values can be estimated well. We propose several ways to access significance of predictors based on their oracle P-values in practice. Furthermore, we illustrate how the oracle P-value can be useful to enhance variable ranking and screening with FDR control, which makes it a valuable tool for high dimensional modeling and inference.

The paper is organized as follows. Section 2 introduces the concept of the oracle P-value. Section 3 illustrates several practical procedures to mimic the oracle P-value. Section 4 discusses the applications of oracle P-values to variable ranking and screening with FDR control. Numerical studies are demonstrated in Section 5. The paper ends with a short discussion.

2. Oracle P-value

Consider a linear regression model

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

where $\mathbf{Y} = (Y_1, ..., Y_n)^T$ is an n-vector of responses, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a p-vector of parameters, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$ is an n-vector of independent and identically distributed (*i.i.d.*) random noises with mean 0 and variance σ^2. In model (1), a subset of variables are assumed to be relevant to the response, denoted by $\mathcal{S}^* = \{j : \beta_j \neq 0\}$. We call these variables in $\mathcal{S}^*$ *important* variables and the rest *noise* variables. For any subset

$\mathcal{M} \subset \{1, 2, ..., p\}$, we denote by $\mathbf{X}_{\mathcal{M}}$, depending on the context, the submatrix of $\mathbf{X}$ consisting of columns indexed by $\mathcal{M}$ or a submodel with variables in $\mathcal{M}$.

A significance level, or P-value, for each variable can be assigned through the testing hypotheses

$$H_j^0 : \beta_j = 0 \qquad \text{versus} \qquad H_j^a : \beta_j \neq 0, \qquad j = 1, ..., p. \tag{2}$$

When $p < n$, a natural testing procedure is based on LSE. For each variable $\mathbf{X}_j$, one calculates its least squares estimator $\hat{\beta}_j$, the standard error $\text{se}(\hat{\beta}_j)$, t-statistic $t_j = |\hat{\beta}_j|/\text{se}(\hat{\beta}_j)$ and the P-value defined based on the cumulative distribution function (CDF) of Student's t or normal distribution. The P-value under null hypothesis follows a uniform distribution exactly or asymptotically due to the central limit theorem, depending on the error distribution. For high dimensional models with $p > n$, this machinery does not work any more. Instead we introduce a concept of oracle P-value for a high dimensional sparse linear model. Assume $s = |\mathcal{S}^*| < n$.

Definition 1. For each variable $\mathbf{X}_j$, we define its oracle P-value, denoted by p_j^o, as

$$\begin{cases} \text{the P-value based on LSE via linear model } \mathbf{Y} \sim \mathbf{X}_{\mathcal{S}^*}, & \text{if } j \in \mathcal{S}^*; \\ \text{the P-value based on LSE via linear model } \mathbf{Y} \sim \mathbf{X}_{\mathcal{S}^* \cup \{j\}}, & \text{if } j \notin \mathcal{S}^*. \end{cases}$$

It is straightforward to show that $p_j^o \sim \text{Unif}[0, 1]$ for $j \notin \mathcal{S}^*$. In practice, since $\mathcal{S}^*$ is unknown, one can not obtain the exact oracle P-value. But this concept offers us an intuitive and simple way to assign significance levels for variables. Its clear interpretation can serve as a benchmark when evaluating other methods. In the next, we propose several procedures to mimic oracle P-values.

3. Mimicking oracle P-values

In practice, since $\mathcal{S}^*$ is unknown, we need to mimic the oracle procedure. The key idea is to identify a set of variables as a good *approximation* of $\mathcal{S}^*$ in the oracle procedure. For each variable $\mathbf{X}_j$, we propose to choose a set $\mathcal{M}_j$ such that the P-value of $\mathbf{X}_j$, p_j, mimics its oracle P-value. Based on the definition of the oracle P-value, an ideal approximation set $\mathcal{M}_j$ should be chosen such that the P-value p_j based on LSE via linear model $\mathbf{Y} \sim \mathbf{X}_{\mathcal{M}_j}$ is (approximately) from Unif$[0, 1]$ if $\mathbf{X}_j$ is a noise variable, and p_j is close to 0 if $\mathbf{X}_j$ is important. The set $\mathcal{M}_j$ can be of form $\mathcal{M} \cup \{j\}$ for a common $\mathcal{M}$ or chosen data adaptively for each j. After we obtain the sets $\{\mathcal{M}_j\}_{j=1}^p$, a set of P-values $\{p_j\}_{j=1}^p$ mimicking the oracle one can be easily calculated. In particular, for the oracle P-value, $\mathcal{M}_j = \mathcal{S}^* \cup \{j\}$ is used.

Before we discuss how to choose $\mathcal{M}_j$'s in practice, we first make the following observations. The first observation is that the testing procedure is tolerant with some noise variables in $\mathcal{M}_j$. In other words, the P-value is valid as long as $\mathcal{M}_j \supset \mathcal{S}^*$, $|\mathcal{M}_j| < n$. The second observation is that absence of an important variable from $\mathcal{M}_j$ would not affect much on the procedure if, conditional on

$\mathbf{X}_{\mathcal{M}_j \setminus \{j\}}$, the missing important variable is independent of $\mathbf{X}_j$. For each variable $\mathbf{X}_j$, we define $\mathcal{S}_j \subset \mathcal{S}^*$ such that $\mathbf{X}_{\mathcal{S}^* \setminus \mathcal{S}_j} \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{X}_{\mathcal{S}_j}$. The following proposition gives a sufficient condition for the P-value of a noise variable to be uniformly distributed on $[0, 1]$.

Proposition 1. *If $\mathcal{M}_j \supset \mathcal{S}_j$, the P-value p_j for a noise variable $\mathbf{X}_j$ based on the LSE via linear model $\mathbf{Y} \sim \mathbf{X}_{\mathcal{M}_j}$ satisfies $p_j \sim$ Unif $[0, 1]$.*

Proof. First, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{X}_{\mathcal{S}^*}$ as $\mathbf{X}_j$ is a noise variable. By definition of $\mathcal{S}_j$, $\mathbf{X}_{\mathcal{S}^* \setminus \mathcal{S}_j} \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{X}_{\mathcal{S}_j}$, which implies $\mathbf{X}_{\mathcal{S}^*} \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{X}_{\mathcal{S}_j}$, and hence $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{X}_{\mathcal{S}_j}$. Finally, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}_j \mid \mathbf{X}_{\mathcal{M}_j \setminus \{j\}}$ as $\mathcal{M}_j \supset \mathcal{S}_j$, so we conclude $p_j \sim$ Unif$[0, 1]$.

Moreover, when $\mathcal{M}_j \supset \mathcal{S}^*$, the joint distribution of these P-values can be accessed by the corresponding LSEs. Recall we calculate, for each j, the LSE $\hat{\beta}_j|_{\mathcal{M}_j}$, or $\hat{\beta}_j$ for short when $\mathcal{M}_j$ is clearly defined from the context. Write $\mathbf{X}_j = \mathbf{X}_j^{\perp} + \mathbf{X}_j^{\shortparallel}$ where $\mathbf{X}_j^{\shortparallel}$ is the projection of $\mathbf{X}_j$ to the column space of $\mathbf{X}_{\mathcal{M}_j \setminus \{j\}}$. Then

$$\hat{\beta}_j = (\mathbf{X}_j^{\perp})^{\top} \mathbf{Y} / (\mathbf{X}_j^{\perp})^{\top} \mathbf{X}_j \tag{3}$$

$$\operatorname{cov}(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 (\mathbf{X}_j^{\perp})^{\top} \mathbf{X}_k^{\perp} / ||\mathbf{X}_j^{\perp}||^2 ||\mathbf{X}_k^{\perp}||^2 \tag{4}$$

Then we have the following result:

Proposition 2. *Under linear model (1) with Gaussian noise. If $\mathcal{M}_j \supset \mathcal{S}^*$, the conditional LSE $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^{\top}$ defined above is jointly normal $N(\boldsymbol{\beta}, \boldsymbol{\Sigma}^*)$ where*

$$\boldsymbol{\Sigma}^*_{jk} = \sigma^2 (\mathbf{X}_j^{\perp})^{\top} \mathbf{X}_k^{\perp} / ||\mathbf{X}_j^{\perp}||^2 ||\mathbf{X}_k^{\perp}||^2.$$

The proof is straightforward and hence omitted here. This result indicates that the covariance information of the P-values is completely determined by the design matrix $\mathbf{X}$. With this important information, we can rank and screen variables by P-values and control the FDR in an accurate way. See next section for more details.

In order to construct $\mathcal{M}_j$, we start with a subset $\hat{\mathcal{S}}$, which is obtained from existing selection or screening procedures. One preferred choice is a procedure which has sure screening properties. Another important factor is computational cost, especially when p is large. Based on the above concerns, we propose to mimic the oracle P-values using the sure independence screening (SIS) procedures (Fan & Lv, 2008) and the LASSO (Tibshirani, 1996). Next, we describe three practical procedures to construct $\mathcal{M}_j$.

Procedure 1. Use sure independence screening (SIS) to obtain $\hat{\mathcal{S}}$, i.e., $\hat{\mathcal{S}} = \{k : |\operatorname{corr}(\mathbf{Y}, \mathbf{X}_k)| > c\}$ for some threshold value c. Define $\mathcal{M}_j = \hat{\mathcal{S}} \cup \{j\}$.

Procedure 2. Use sure independence screening (SIS) to identify $\hat{\mathcal{S}} = \{k : |\operatorname{corr}(\mathbf{Y}, \mathbf{X}_k)| > c\}$. Let $\mathcal{H}_j = \{k : |\operatorname{corr}(\mathbf{X}_j, \mathbf{X}_k)| > c_j\}$, which includes those variables which are highly correlated to $\mathbf{X}_j$. Define $\mathcal{M}_j = \{k : |\operatorname{corr}(\mathbf{Y}, \mathbf{X}_k)| > c\} \cup \{k : |\operatorname{corr}(\mathbf{X}_j, \mathbf{X}_k)| > c_j\}$. Here the threshold values c and c_j are both pre-specified.

Procedure 3. Use LASSO to obtain $\hat{\mathcal{S}}$. Define $\mathcal{M}_j = \hat{\mathcal{S}} \cup \{j\}$.

The strategy of Procedure 2 is particularly useful if there is some priori knowledge on the covariance structures of the covariates, such as a block structure or a tapped/fast decayed covariance structure, so $\mathcal{H}_j$ can be chosen according to the covariance structure. Moreover, when $\hat{\mathcal{S}}$ misses some important variables which are also correlated to $\mathbf{X}_j$, the set $\mathcal{H}_j$ may make up these variables.

4. Applications to variable screening and ranking

4.1. Variable screening

Variable selection is a central topic in high dimensional data analysis, because it can effectively achieve dimension reduction by identifying important predictors and screening out noise. From the viewpoint of hypothesis testing, the problem of variable selection can be treated as p separate testing problems (2), as the rejected hypotheses can naturally result an estimator $\hat{\mathcal{S}}$ for $\mathcal{S}^*$. When $p < n$, a testing procedure based on LSE for model selection is described as follows. We calculate the P-values based on LSE and then choose a threshold p^* to control the FWER or FDR. This approach has been shown to achieve consistency in model selection (Bauer et al., 1988; Bunea et al., 2006). However, it requires that the data dimensionality p is fixed or much smaller than n. When $p \gg n$, this strategy does not work any more. In the following, we propose an FDR approach based on oracle P-values for high dimensional sparse linear regression.

We first describe an oracle procedure for controlling FDR-type error rates for the ideal situation when oracle P-values were available.

Oracle Procedure: (ideal case)

1. Calculate p_j^o for each $j = 1, \ldots, p$.
2. Choose a threshold p^* and reject H_j^0 if $p_j^o < p^*$.

In practice, the true important set $\mathcal{S}^*$ is generally unknown, and therefore oracle P-values are not available. In the following, we propose an oracle proxy procedure which first approximates oracle P-values and then controls the FDR based on the approximated P-values

Oracle Proxy Procedure:

Stage 1: For each $j = 1, \ldots, p$, find $\mathcal{M}_j$ using the procedures proposed in Section 3 and calculate p_j based on LSE $\hat{\beta}_j$ via model $\mathbf{Y} \sim \mathbf{X}_{\mathcal{M}_j}$.

Stage 2: Choose a threshold p^* and reject H_j^0 if $p_j < p^*$.

For implementation practice, we suggest first splitting the data into two parts and then using one half to obtain $\{\mathcal{M}_j\}_{j=1}^p$ and the other half to compute p_j. After $\{p_j\}_{j=1}^p$ are obtained, one can apply any FDR-control method to determine the threshold p^*. In the following, we discuss three procedures to control FDR-type at the target level $0 < \alpha < 1$.

Benjamini-Hochberg 1 (BH1) method: Let $p_{(1)} \leq p_{(2)} \leq \cdots p_{(p)}$ be the ordered P-values. Define $k = \max\{j : p_{(j)} \leq \frac{j}{p}\alpha\}$ and reject $H_{(1)}^0$,..., $H_{(k)}^0$. This has become a standard procedure to control FDR since it was proposed in the

seminal work of Benjamini & Hochberg (1995). The BH1 is expected to work well when the test statistics are independent or in some special scenarios of dependence (Benjamini & Yekutieli, 2001).

Benjamini-Hochberg 2 (BH2) method: Let $p_{(1)} \leq p_{(2)} \leq \cdots p_{(p)}$ be the ordered P-values. Define $k = \max\{j : p_{(j)} \leq \frac{j}{p}\frac{\alpha}{\sum_{i=1}^{p} i^{-1}}\}$ and reject $H^0_{(1)}$,..., $H^0_{(k)}$. In theory, this procedure can control the FDR under arbitrary dependence structures (Benjamini & Yekutieli, 2001) but it can be conservative.

Principle Factor Approximator (PFA) method: This procedure, proposed by Fan et al. (2012b), is designed to make use of the covariance information of the test statistics to improve the control of FDP, which is a realized FDR and arguable more relevant quantity to control. In particular, for a multivariate normal test statistic with known covariance, they expressed the test statistic by an approximate factor model with weakly dependent random errors, and derived an explicit formula for the FDP in large-scale simultaneous tests. Note that, in Proposition 2, we showed that the covariance structure of test statistic $\hat{\boldsymbol{\beta}}$ is completely known except for the variance parameter σ^2, which can estimated by refitted cross-validation (Fan et al., 2012a). Moreover, the following proposition indicates that, when $n \ll p$, the leading eigenvalues of Σ^* dominates its trace, which is a key condition to make PFA work. Therefore, the PFA procedure can be applied directly to our proposed procedures for FDR control.

Proposition 3. *The covariance matrix* $\boldsymbol{\Sigma}^*$ *defined in Proposition 2 has rank at most* n*, so PFA theory can be applied when* $n \ll p$*.*

Proof: Define $\mathbf{Z} = (\mathbf{X}_1^{\perp}, \cdots, \mathbf{X}_p^{\perp})$ and $\mathbf{D} = diag(\|\mathbf{X}_1^{\perp}\|^2, \cdots, \|\mathbf{X}_p^{\perp}\|^2)$. Then $\boldsymbol{\Sigma}^* = \mathbf{D}^{-1}\mathbf{Z}^{\top}\mathbf{Z}\mathbf{D}^{-1}$. So the rank of $\boldsymbol{\Sigma}^*$ is at most n.

The first two procedures, i.e., BH1 and BH2, are easy to implement. However, the obtained P-values in this paper are usually highly correlated so BH1 can not control the FDR to the targeted level. BH2 can control the FDR under the targeted level but is often too conservative. On the other hand, PFA can fully take the advantage of the known covariance structure and control the FDP accurately, which is also confirmed by our numerical studies. We implement and compare performance of these three procedures to control FDR based on the oracle P-values in Section 5.

4.2. Variable ranking

Variable ranking is of practical importance since an informative and accurate ranking helps to understand relative importance of the predictors to the response. In practice, it is desired to rank the predictors in the decreasing order of their relevance to the output, i.e., more relevant variables are ranked at the top while noise variables are at the bottom. A reliable list of variable ranking can be used as a variable selection procedure by only retaining variables ranked at the top of the list in the model, or equivalently, discarding those ranked at the bottom. A critical question is how to determine the boundary between "top" and "bottom" variables.

Based on Definition 1, the magnitudes of oracle P-values reflect relative importance of *important* variables and *noise* variables. In general, oracle P-values of important variables are expected to be much smaller than those of noise variables. Therefore, we can use oracle P-values or their approximations to rank variables into two clusters. In Section 5, we use two examples to illustrate the ranking performance of oracle P-values. As pointed out by one reviewer, our P-value based ranking measures the relative importance of a variable in a joint model with other variables adjusted. It is not the same as marginal ranking utilized in sure screening approaches (Fan & Lv, 2008), which rank variables based on the marginal correlation between each individual predictor with the response, without the presence of other variables in the model. Therefore, our rank is more informative by taking into account contributions of other variables to the model.

5. Numerical results

We use several numerical examples to illustrate the properties of oracle P-values and their approximations. Four different regression model settings are considered, including both independent predictors and correlated predictors. To assess robustness of the methods in relatively weak signal settings, Examples 3 and 4 are designed to include predictors with very small regression coefficients, which is difficult to select at the model selection step. In each setting, we run $M = 100$ Monte Carlo simulations and summarize the results. Recall that the full model index set is $\{1, \ldots, p\}$ and the true model set is $\mathcal{S}^*$, where $|\mathcal{S}^*| = s < n$.

We compare the true oracle procedure (referred to as "Oracle") and three estimation procedures: *Procedure 1* (referred to as "SIS-1"), *Procedure 2* (referred to as "SIS-2"), and *Procedure 3* (referred to as "LASSO"). The oracle procedure serves as the gold standard, and it is generally not available in practice. Both SIS-1 and SIS-2 methods implement the SIS procedure of Fan & Lv (2008) to obtain $\hat{\mathcal{S}}$ and $\mathcal{M}_j$. For SIS-1 and SIS-2, we select a model with fixed model size $\lceil n/\log n \rceil$; for LASSO, we select the tuning parameter (λ) by cross validation. Furthermore, to compare with existing P-value procedures widely used in the literature, we also include two methods of Dezeure et al. (2015): the hdi (multiple split) method and the ridge projection method, both implemented using their R package **hdi**.

Example 1 (independent predictors)**.** Let $(n, p, q) = (200, 1000, 5)$. Generate $\mathbf{X}$'s *i.i.d.* from $MVN(\mathbf{0}, I_p)$, and the response from the linear model $Y_i = \beta_0 + \mathbf{X}_i^T\boldsymbol{\beta} + \epsilon_i, i = 1, \cdots, n$, with the error $\epsilon \sim N(0, \sigma^2)$. The true $\boldsymbol{\beta} = (1, 1, 2, 0, 2, 0, 0, -1, \mathbf{0}_{992})$. The error variance σ^2 is chosen such that $R^2 = 0.6$. The true model set $\mathcal{S}^* = \{1, 2, 3, 5, 8\}$.

Example 2 (AR correlation)**.** Consider the same setup as Example 1, except that $\mathbf{X}$ follows MVN with mean $\mathbf{0}$ and $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$ for $1 \le j, k \le p$.

Example 3 (weak signal, independent predictors)**.** Consider the same set up as Example 1, except the true $\boldsymbol{\beta} = (1, 0.5, 2, 0, 2, 0, 0, -1, \mathbf{0}_{992})$. The true model

set $\mathcal{S}^* = \{1, 2, 3, 5, 8\}$. In this example, X_2 has a small coefficient, so it is a relatively weaker predictor.

Example 4 (weak signal, AR correlation)**.** Consider the same setup as Example 3, except that $\mathbf{X}$ follows MVN with mean $\mathbf{0}$ and $\text{Cov}(X_j, X_k) = 0.5^{|j-k|}$ for $1 \leq j, k \leq p$.

5.1. Distribution of oracle P-values

Proposition 1 states that for any $j \notin \mathcal{S}^*$, its oracle P-value has a marginal distribution Unif$[0, 1]$. We illustrate this by showing the distribution of oracle P-values and their estimates given by SIS-1, SIS-2, LASSO, hdi, and ridge projection. To start with, we use the quantile-quantile (q-q) plot to compare quantiles of oracle P-values against those of the null distribution Unif$[0, 1]$. Then, for a more rigorous analysis, we conduct the Kolmogorov-Smirnov (K-S) test for the empirical distribution functions of oracle P-values compared with the reference probability distribution Unif$[0, 1]$. The K-S test statistic measures the distance between the empirical distribution function of oracle P-values and the cumulative distribution function of the reference distribution Unif$[0, 1]$. We compute the K-S statistics and their associated P-values over 100 Monte Carlo samples and report their average values.

Figure 1 shows Q-Q plots for oracle P-values against the quantiles of Unif$[0, 1]$ for Example 1. Due to the space constraint, we only show the plots for the first six variables. Recall that X_1, X_2, X_3, X_5 are important variables, while X_4 and X_6 are noise variables. The first row is for the true oracle P-values, calculated based on Definition 1. The next three rows are for the approximately oracle P-values given by SIS-1, SIS-2, and LASSO, respectively. We observe that, the Q-Q plots of noise variables follow a straight line, and those for important variables seriously deviate from a straight line. For the remaining 994 variables, we observe the similar pattern, showing that the oracle P-values of noise variables follow a Unif$[0, 1]$. The last row displays the Q-Q plots produced by the hdi method, suggesting that the distribution of the P-values for X_4 and X_6 does not follow Unif(0,1) under the null hypothesis. We also implement the ridge projection method and observe the same pattern. The reason is that the hdi-type procedures intentionally set P-values for unimportant variables to be one. Due to the space constraint, we only present the Q-Q plots for the hdi method in Figure 1. Example 2 considers correlated predictors and we observe the similar pattern as in Example 1. In particular, the last row in Figure 2 show the Q-Q plots for the ridge projection method, also suggesting that the null distribution of the P-values for X_4 and X_6 does not follow Unif(0,1). The same pattern is also observed for the hdi method. Examples 3 and 4 contain a weak signal X_2 which has a small regression coefficient. In Figures 3 and 4, we observe that its Q-Q plot deviates from a straight line, suggesting that its oracle P-value does not follow Unif$[0, 1]$.

We summarize the K-S tests results in Tables 2 to 5. Each table consists of two parts: the top part shows the K-S test statistics, and the bottom part

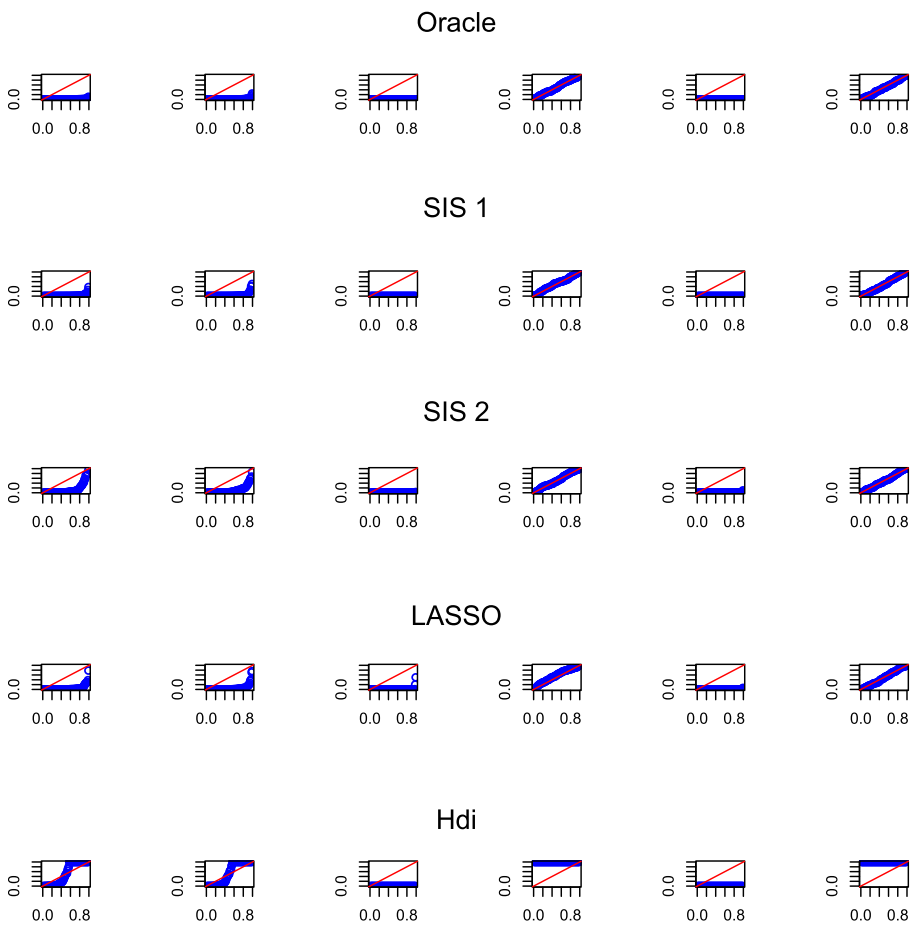

FIG 1. *The Q-Q plots of P-values estimated in Example 1 by various procedures: the proposed oracle P-values, SIS-1, SIS-2, LASSO, and hdi. In each row, each column (from left to right) corresponds to the Q-Q plot for one variable in* $\{X_1, \ldots, X_6\}$*, respectively. Recall that* X_1, X_2, X_3, X_5 *are important variables, and* X_4 *and* X_6 *are noise variables. In each plot, x-axis represents the theoretical quantiles of Unif*$(0, 1)$*, and y-axis represents the observed quantiles for the calculated P-values.*

reports the P-values for the corresponding K-S tests. Based on Tables 2 and 3, we observe that the test results are significant for important variables, implying the rejection of the null Unif$[0, 1]$ distribution. By contrast, the K-S P-values for noise variables are mostly larger than 0.05, suggesting that their distribution is from Unif$[0, 1]$. The K-S test results for the other noise variables are not significant, though they are not reported here due to the space constraint. For weak signal cases in Examples 3 and 4, the K-S P-value of X_2 is close to zero, suggesting that its oracle P-values do not follow Unif$[0, 1]$. Therefore, the proposed oracle P-values and their approximations still work well for weak signal

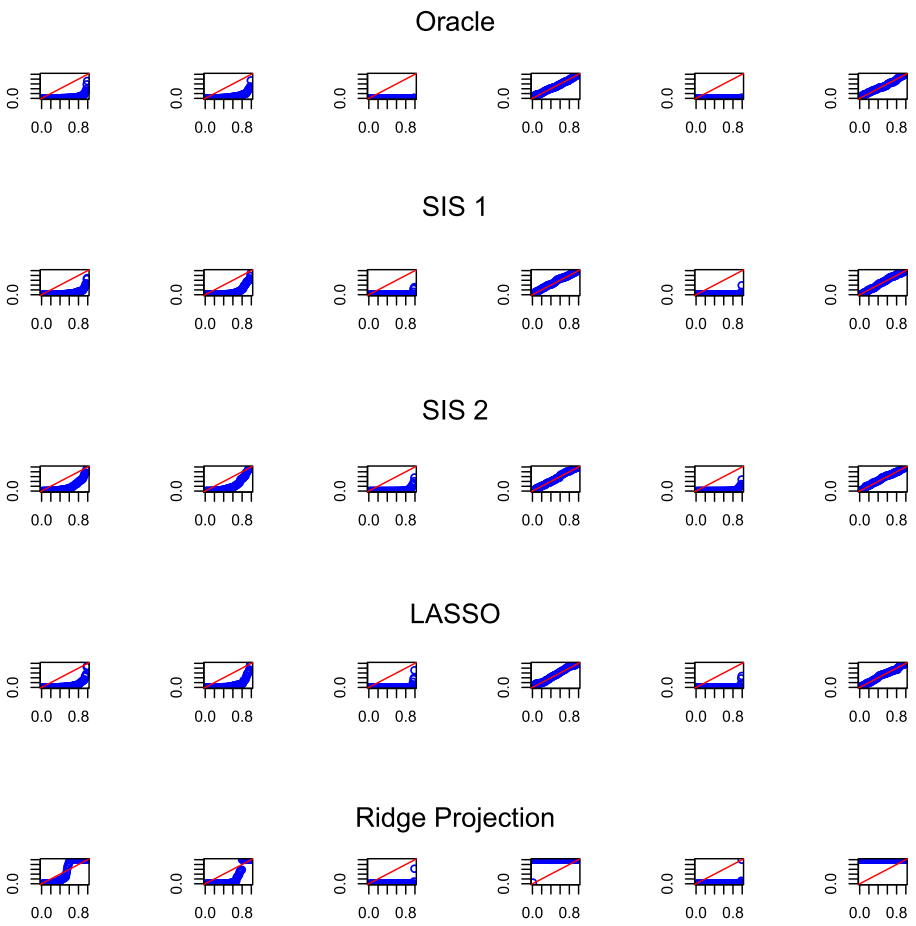

FIG 2. *The Q-Q plots of P-values estimated in Example 2 by various procedures: the proposed oracle P-values, SIS-1, SIS-2, LASSO, and ridge projection. In each row, each column (from left to right) corresponds to the Q-Q plot for one variable in* $\{X_1, \ldots, X_6\}$*, respectively. Recall that* X_1, X_2, X_3, X_5 *are important variables, and* X_4 *and* X_6 *are noise variables. In each plot, the x-axis represents the theoretical quantiles of Unif*$(0, 1)$*, and the y-axis represents the observed quantiles for the calculated P-values.*

settings. The last two rows in Tables 2 to 5 suggest that, for both hdi and ridge projection methods, the estimated P-valuves do not follow Unif$[0, 1]$ under the null hypothesis for noise variables $\{X_4, X_6, X_7\}$.

5.2. Variable ranking

We show that oracle P-values provide an informative guide on the relative importance of variables. In Figures 5–8, we depict the box plots for the ranks of all the 1000 variables. Due to the space limit, only the ranks of first eight

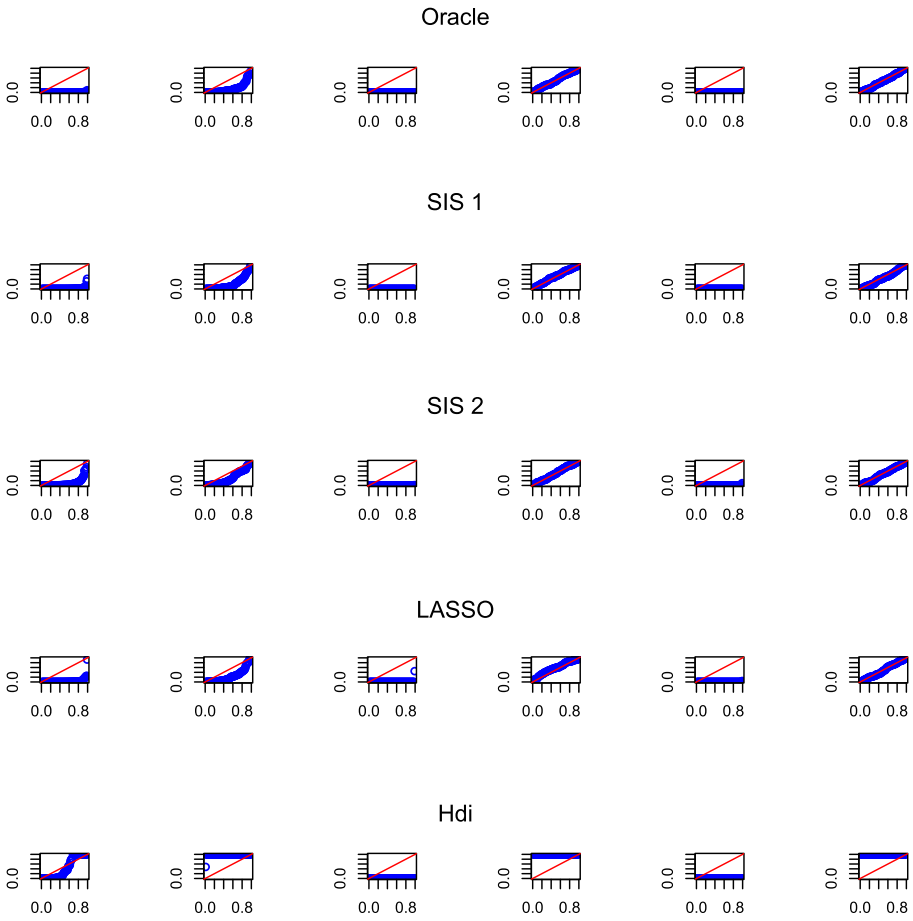

FIG 3. *The Q-Q plots of P-values estimated in Example 3 by various procedures: the proposed oracle P-values, SIS-1, SIS-2, LASSO, and hdi. In each row, each column (from left to right) corresponds to the Q-Q plot for one variable in* $\{X_1, \ldots, X_6\}$*, respectively. Recall that* X_1, X_2, X_3, X_5 *are important variables, and* X_4 *and* X_6 *are noise variables. In each plot, the x-axis represents the theoretical quantiles of Unif*$(0, 1)$*, and the y-axis represents the observed quantiles for the calculated P-values.*

variables are reported here. In each figure, there are four panels, which correspond the procedure of oracle, SIS 1, SIS 2 and LASSO. Overall, we observe that the ranks of important variables in $\mathcal{S}^*$ are much smaller than those of noise variables. Therefore oracle P-values perform very well in ranking variables, by putting important variables ahead of noise variables. In Examples 2 and 3, the rank of the weak signal X_2 is relatively larger than strong signals, but it is still in average smaller those of noise variables, as shown in Figures 7 and 8. This suggests the robust performance of the oracle P-values.

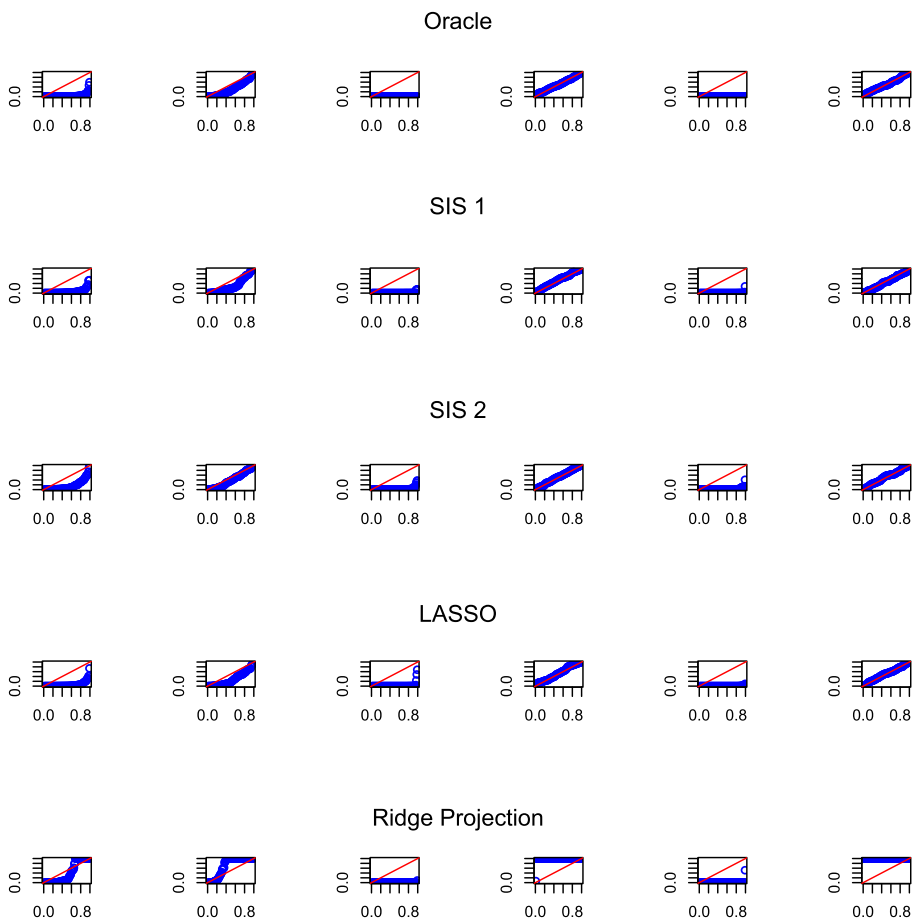

FIG 4. *The Q-Q plots of P-values estimated in Example 4 by various procedures: the proposed oracle P-values, SIS-1, SIS-2, LASSO, and ridge projection. In each row, each column (from left to right) corresponds to the Q-Q plot for one variable in* $\{X_1, \ldots, X_6\}$*, respectively. Recall that* X_1, X_2, X_3, X_5 *are important variables, and* X_4 *and* X_6 *are noise variables. In each plot, the x-axis represents the theoretical quantiles of Unif*$(0, 1)$*, and the y-axis represents the observed quantiles for the calculated P-values.*

5.3. Variable selection under FDR control

We illustrate the performance of oracle P-value for variable selection with FDR-type control for the four examples. Three FDR-control methods are considered: Benjamini-Hochberg 1 (BH1), Benjamini-Hochberg 2 (BH2), and the PFA method by Fan et al. (2012b). The target FDR level is set as 10% for each setting. For each example, we run 100 simulations and report the average number of true positive (TP), false positive (FP), and the average FDR. The results are summarized in Tables 6–9, where TP represents "True Positive" and FP

TABLE 2
Kolmogorov-Smirnov test results for Example 1. (denotes important variables).*

	Method	X_1^*	X_2^*	X_3^*	X_4	X_5^*	X_6	X_7	X_8^*
K-S Statistic	Oracle	0.942	0.905	1.000	0.053	1.000	0.065	0.076	0.914
	SIS-1	0.855	0.847	1.000	0.074	0.993	0.074	0.089	0.836
	SIS-2	0.665	0.647	0.972	0.054	0.945	0.067	0.060	0.703
	LASSO	0.796	0.787	0.970	0.115	0.965	0.098	0.085	0.780
	hdi	0.430	0.440	1.000	1.000	1.000	1.000	1.000	0.410
	ridge projection	0.459	0.393	1.000	1.000	1.000	1.000	1.000	0.453
K-S P-value	Oracle	0.000	0.000	0.000	0.939	0.000	0.796	0.604	0.000
	SIS-1	0.000	0.000	0.000	0.640	0.000	0.642	0.411	0.000
	SIS-2	0.000	0.000	0.000	0.934	0.000	0.760	0.865	0.000
	LASSO	0.000	0.000	0.000	0.140	0.000	0.287	0.463	0.000
	hdi	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	ridge projection	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 3
Kolmogorov-Smirnov test results for Example 2. (denotes important variables).*

	Method	X_1^*	X_2^*	X_3^*	X_4	X_5^*	X_6	X_7	X_8^*
K-S Statistic	Oracle	0.786	0.694	0.982	0.078	0.988	0.074	0.069	0.840
	SIS-1	0.712	0.613	0.932	0.096	0.940	0.068	0.274	0.764
	SIS-2	0.498	0.452	0.827	0.062	0.841	0.080	0.138	0.505
	LASSO	0.685	0.659	0.908	0.073	0.932	0.071	0.109	0.752
	hdi	0.550	0.500	0.945	0.990	0.960	1.000	1.000	0.840
	ridge projection	0.390	0.568	0.976	0.990	0.942	1.000	1.000	0.760
K-S P-value	Oracle	0.000	0.000	0.000	0.575	0.000	0.643	0.730	0.000
	SIS-1	0.000	0.000	0.000	0.315	0.000	0.749	0.000	0.000
	SIS-2	0.000	0.000	0.000	0.833	0.000	0.541	0.044	0.000
	LASSO	0.000	0.000	0.000	0.665	0.000	0.687	0.182	0.000
	hdi	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	ridge projection	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 4
Kolmogorov-Smirnov test results in Example 3. (denotes important variables).*

	Method	X_1^*	X_2^*	X_3^*	X_4	X_5^*	X_6	X_7	X_8^*
K-S Statistic	Oracle	0.953	0.568	1.000	0.053	1.000	0.065	0.076	0.931
	SIS-1	0.881	0.489	1.000	0.081	0.999	0.096	0.058	0.884
	SIS-2	0.670	0.325	0.987	0.060	0.953	0.084	0.069	0.711
	LASSO	0.832	0.419	0.989	0.146	0.974	0.091	0.110	0.822
	hdi	0.354	0.990	1.000	1.000	1.000	1.000	1.000	0.362
	ridge projection	0.512	0.900	1.000	1.000	1.000	1.000	1.000	0.500
K-S P-value	Oracle	0.000	0.000	0.000	0.939	0.000	0.796	0.604	0.000
	SIS-1	0.000	0.000	0.000	0.535	0.000	0.312	0.893	0.000
	SIS-2	0.000	0.000	0.000	0.863	0.000	0.480	0.731	0.000
	LASSO	0.000	0.000	0.000	0.028	0.000	0.385	0.176	0.000
	hdi	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	ridge projection	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

represents "False Positive". Overall speaking, it is observed that all of the procedures perform reasonably well in controlling the FDR. In terms of achieving the nominal FDR value, Fan et al. (2012b) performs slightly better than others.

TABLE 5
Kolmogorov-Smirnov test results in Example 4. (denotes important variables).*

	Method	X_1^*	X_2^*	X_3^*	X_4	X_5^*	X_6	X_7	X_8^*
K-S Statistic	Oracle	0.835	0.279	0.995	0.078	0.989	0.074	0.069	0.873
	SIS-1	0.764	0.366	0.941	0.077	0.955	0.077	0.331	0.792
	SIS-2	0.537	0.188	0.853	0.055	0.857	0.086	0.128	0.555
	LASSO	0.754	0.354	0.936	0.104	0.944	0.057	0.127	0.819
	hdi	0.350	0.670	0.992	0.990	0.982	1.000	1.000	0.780
	ridge projection	0.394	0.590	0.984	0.990	0.977	1.000	1.000	0.630
K-S P-value	Oracle	0.000	0.000	0.000	0.575	0.000	0.643	0.730	0.000
	SIS-1	0.000	0.000	0.000	0.588	0.000	0.595	0.000	0.000
	SIS-2	0.000	0.000	0.000	0.927	0.000	0.450	0.007	0.000
	LASSO	0.000	0.000	0.000	0.227	0.000	0.897	0.008	0.000
	hdi	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	ridge projection	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

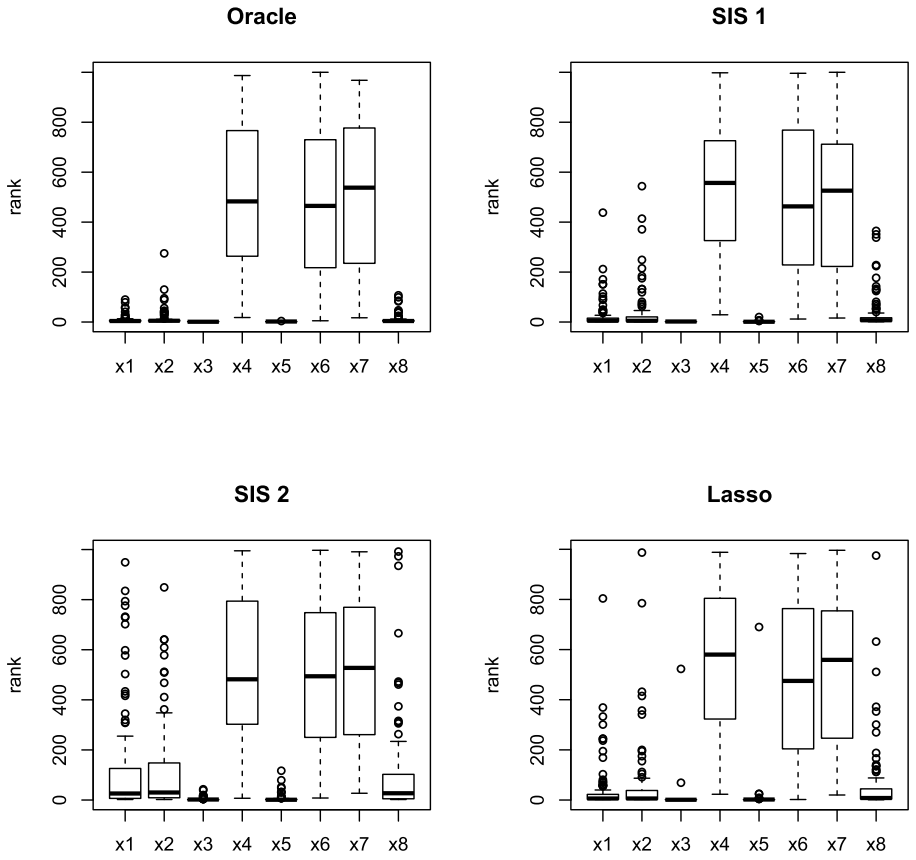

FIG 5. *Boxplots of variable ranks for the first eight variables (Example 1).*

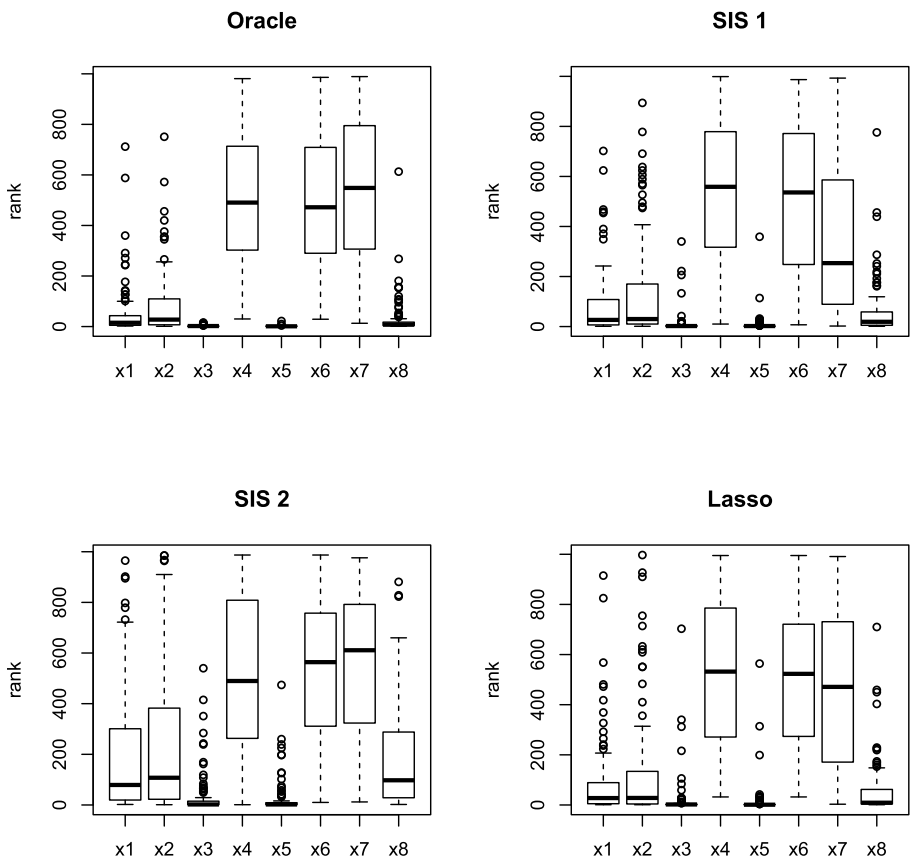

FIG 6. *Boxplots of variable ranks for the first eight variables (Example 2).*

TABLE 6
Variable selection results with FDR control at the level 10% *(Example 1).*

	SIS-1			LASSO			Oracle		
FDR Method	TP	FP	FDR	TP	FP	FDR	TP	FP	FDR
PFA	3.06	0.44	0.09	3.00	0.51	0.09	3.75	0.39	0.07
BH1	3.44	3.49	0.27	3.16	2.01	0.18	3.67	0.45	0.08
BH2	2.74	0.31	0.05	2.44	0.33	0.05	3.11	0.01	0.00

TABLE 7
Variable selection results with FDR control at the level 10% *(Example 2).*

	SIS-1			LASSO			Oracle		
FDR Method	TP	FP	FDR	TP	FP	FDR	TP	FP	FDR
PFA	1.55	0.20	0.06	1.74	0.41	0.11	2.51	0.29	0.06
BH1	1.65	0.78	0.15	1.84	0.77	0.11	2.27	0.28	0.06
BH2	1.00	0.04	0.02	1.29	0.08	0.02	1.83	0.02	0.01

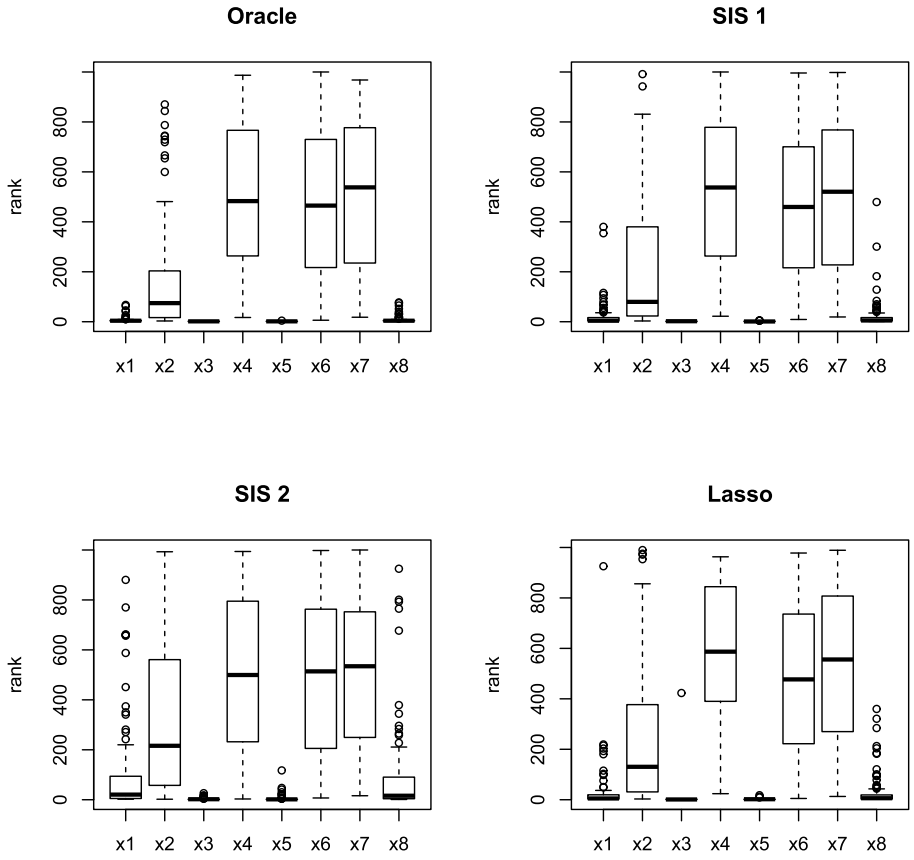

FIG 7. *Boxplots of variable ranks for the first eight variables (Example 3).*

TABLE 8
Variable selection results with FDR control at the level 10% *(Example 3).*

	SIS-1			LASSO			Oracle		
FDR Method	TP	FP	FDR	TP	FP	FDR	TP	FP	FDR
PFA	2.85	0.48	0.10	2.72	0.45	0.09	3.27	0.37	0.08
BH1	3.01	2.14	0.24	2.83	1.91	0.17	3.23	0.39	0.08
BH2	2.59	0.24	0.05	2.38	0.17	0.02	2.91	0.01	0.00

TABLE 9
Variable selection results with FDR control at the level 10% *(Example 4).*

	SIS-1			LASSO			Oracle		
FDR Method	TP	FP	FDR	TP	FP	FDR	TP	FP	FDR
PFA	1.84	0.30	0.08	2.08	0.36	0.08	2.66	0.29	0.07
BH1	1.89	0.84	0.15	1.97	0.78	0.12	2.49	0.29	0.07
BH2	1.32	0.08	0.03	1.51	0.08	0.01	2.04	0.02	0.01

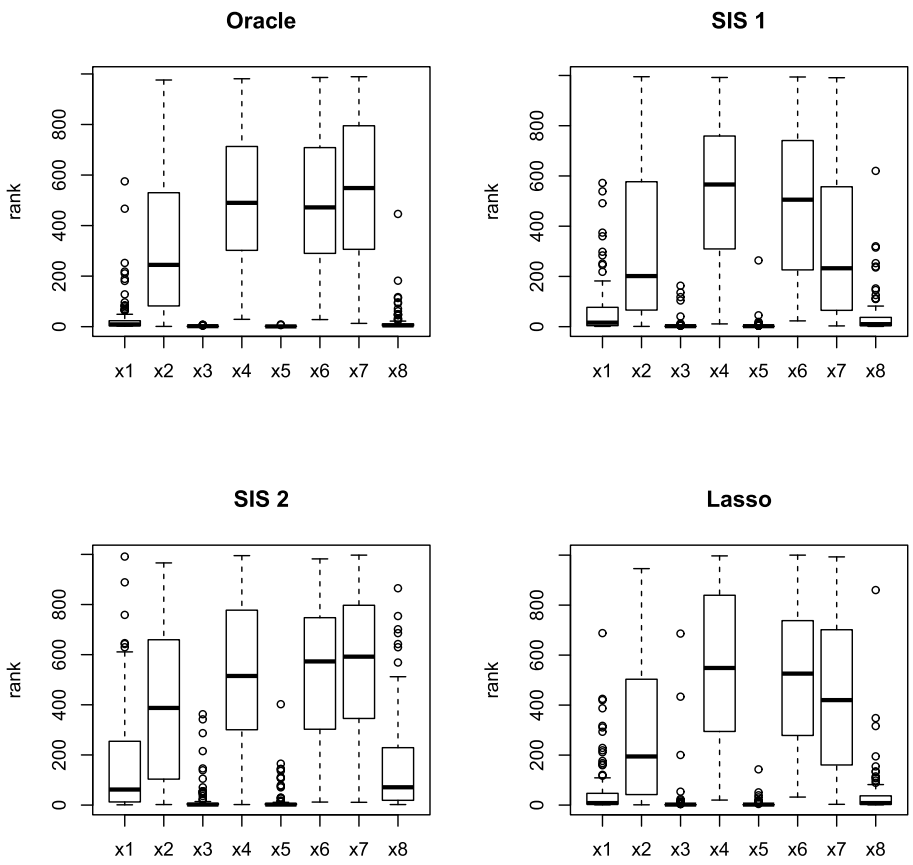

FIG 8. *Boxplots of variable ranks for the first eight variables (Example 4).*

5.4. Real data analysis

We apply the proposed methods to a microarray gene expression data set of Scheetz et al. (2006). The whole experiment aims to under gene regulation in the mammalian eye and to identify genetic variation relevant to human eye disease. The data set we use consists of 120 arrays, and each array contains 18,976 probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array) which exhibited sufficient signal for reliable analysis and at least 2-fold variation in expression. The complete gene expression data is available at Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo`; accession number GSE5680). One primary objective of this study is to identify which gene expressions are related to gene TRIM 32, which is recently found to cause Bardet-Biedl syndrome (Chiang et al., 2006). The probe ID associated with the response, TRIM32, is 1389163 at.

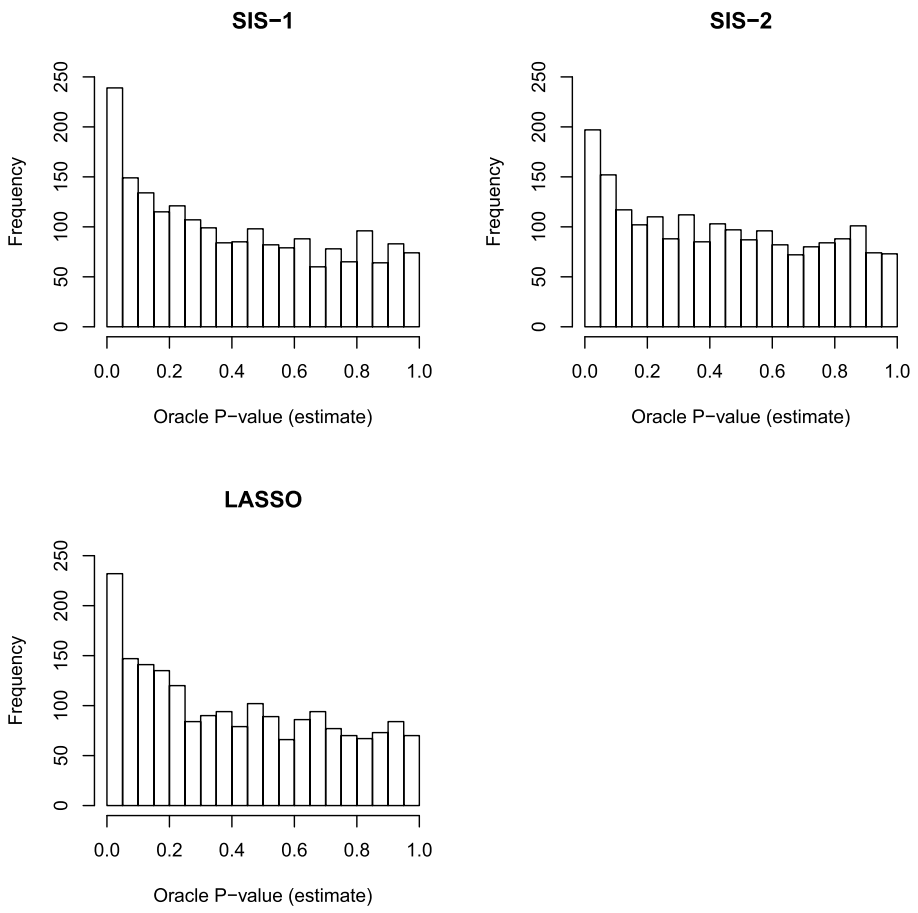

FIG 9. *Histograms of oracle P-values for the rat microarray data set.*

To identify which genes are correlated with TRIM32, we regress TRIM32 on the remaining probes. For illustration, we first identify 2,000 top genes based on their marginal correlation with the response. Then we assign their significance using the proposed oracle P-value measures. In real data analysis, since the true model is unknown, we can only use SIS-1, SIS-2, and LASSO to estimate the oracle P-values. Figure 9 presents histograms of the oracle P-values obtained by the three procedures. It is observed that, except a small portion of genes with very small oracle P-values, the oracle P-values of the rest genes roughly follow Unif$[0, 1]$ distribution. This procedure is useful to provide a short list of "potentially important" genes for scientists to further investigate. For example, the LASSO procedure regards genes with index 180, 1428, 1614, 1769, and 1868 to be among the top-five list. The interpretation of these genes, or whether they may lead to new scientific findings, will rely on scientists' validation experiments.

6. Discussion

For high dimensional linear models, a proper definition of P-value has drawn much attention in literature. In the paper, we propose a new concept of the *oracle P-value* for high dimensional sparse regression models and show that it possesses a uniform distribution for noise variables. For implementation, we propose several practical approaches to mimic the oracle procedure and show their applications to variable ranking and variable screening subject to FDR control. In this work, we illustrate the proposed concept only with simple and fast procedures like SIS and LASSO. It is possible to extend the idea to other procedures.

Acknowledgment

This research is supported in part by National Science Foundations DBI-1261830, DMS-1309507, DMS-1418172, and NSFC-11571009. The authors thank the editors, the associate editor, and the reviewers for their helpful comments and suggestions.

References

Bauer, P., Potscher, B. M. & Hackl, P. (1988). Model selection by multiple test procedures. *Statistics: A Journal of Theoretical and Applied Statistics* **19**, 39–44. MR0921623

Benjamini, Y. & 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. MR1325392

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* **29**, 1165–1188. MR1869245

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer. MR2807761

Bühlmann, P. et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**, 1212–1242. MR3102549

Bunea, F., Wegkamp, M. H. & Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference* **136**, 4349–4364. MR2323420

Chiang, A., Beck, J., Yen, H., Tayeh, M., Scheetz, T., Nishimura, D., Braun, T., Kim, K., Huang, J., Elbedour, K., Carmi, R., Slusarski, D., Casavant, T., Stone, E. & Sheffield, V. (2006). Homozygosity mapping with snp arrays identifies TRIM32, an e3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *Proceeding of National Academy Science, USA* **18**, 6287–92.

Dezeure, R., Buhlmann, P., Meier, L. & Nicolai, M. (2015). High-dimensional inference: confidence intervals, p-values, and R-software hdi. *Statistical Science* **30**, 533–558. MR3432840

Fan, J., Guo, S. & Hao, N. (2012a). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 37–65. MR2885839

Fan, J., Han, X. & Gu, W. (2012b). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association* **107**, 1019–1035. MR3010887

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911. MR2530322

Fan, J. & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148. MR2640659

Fan, J. & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484. MR2849368

Meinshausen, N., Meier, L. & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104**, 1667–1681. MR2750584

Ning, Y. & Liu, H. (2016). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of statistics*, to appear. MR3611489

Scheetz, T. E., Kim, K. Y., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. G., Huang, J., Casavant, T. L., Sheffeld, V. C. & Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* **103**, 14429–14434.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288. MR1379242

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512–1524. MR2750576

Wasserman, L. & Roeder, K. (2009). High dimensional variable selection. *Annals of statistics* **37**, 2178–2201. MR2543689

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942. MR2604701

Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society, Series B.* **76**, 217–242. MR3153940

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563. MR2274449