Local false discovery rate estimation with competition-based procedures for variable selection

Xiaoya Sun1,2 | Yan Fu1,2

1CEMS, NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
2School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Correspondence
Yan Fu, CEMS, NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
Email: yfu@amss.ac.cn

Funding information
National Key Research and Development Program of China, Grant/Award Numbers: 2022YFA1004801, 2022YFA1304603; National Natural Science Foundation of China, Grant/Award Number: 32070668

Multiple hypothesis testing has been widely applied to problems dealing with high-dimensional data, for example, the selection of important variables or features from a large number of candidates while controlling the error rate. The most prevailing measure of error rate used in multiple hypothesis testing is the false discovery rate (FDR). In recent years, the local false discovery rate (fdr) has drawn much attention, due to its advantage of accessing the confidence of individual hypotheses. However, most methods estimate fdr through \( P \)-values or statistics with known null distributions, which are sometimes unavailable or unreliable. Adopting the innovative methodology of competition-based procedures, for example, the knockoff filter, this paper proposes a new approach, named TDfdr, to fdr estimation, which is free of \( P \)-values or known null distributions. Extensive simulation studies demonstrate that TDfdr can accurately estimate the fdr with two competition-based procedures. We applied the TDfdr method to two real biomedical tasks. One is to identify significantly differentially expressed proteins related to the COVID-19 disease, and the other is to detect mutations in the genotypes of HIV-1 that are associated with drug resistance. Higher discovery power was observed compared to existing popular methods.

KEYWORDS
feature selection, knockoff, local false discovery rate, multiple hypothesis testing, target-decoy, variable selection

1 | INTRODUCTION

Multiple hypothesis testing is widely used in fields where large-scale data are produced, such as genomics,1,2 proteomics,3,4 and neuroimaging.5,6 The aim of multiple hypothesis testing is to make assertions simultaneously for many hypotheses while controlling a certain type of error rate. Formally, suppose one is interested in simultaneously testing \( m \) hypotheses, denoted as \( H_1, H_2, \ldots, H_m \), where \( H_j = 0 \) represents that the null hypothesis holds and \( H_j = 1 \) otherwise. The goal of multiple hypothesis testing is to reject a set of hypotheses \( \{ H_j | j \in R, j = 1, \ldots, m \} \), with the error rate controlled under a predetermined threshold \( q \in (0, 1) \), where \( R \) denotes the index set of rejected hypotheses. Specifically, when the tests are about the importance or relevance of variables or features, multiple hypothesis testing is instanced as the problem of variable or feature selection with error rate control.

False discovery rate (FDR) has become the prevailing error rate used in multiple hypothesis testing since this concept was proposed by Benjamini and Hochberg.7 FDR is defined as
FDR serves as a measure of the expected proportion of falsely rejected hypotheses. It addresses the conservativeness of previous measures, such as the per-family error rate (PFER) and the familywise error rate (FWER), thereby enhancing power.

Since the initial introduction of the Benjamini and Hochberg procedure (referred to as “BH” hereafter) for controlling the FDR, a multitude of procedures have emerged. Modifications have been made to the BH procedure to address the assumption of independence. Storey further developed a Bayesian framework for FDR control, enabling direct control of the FDR. Subsequently, the concept of positive FDR was proposed, accompanied by the introduction of a new metric known as the q-value, which provides a Bayesian interpretation of the FDR.

While numerous procedures have been developed to control the FDR, the majority, if not all, of these procedures rely on $P$-values as the measure of significance for hypothesis tests. However, in situations where $P$-values are unavailable and general statistics or scores are generated, such as those produced by machine learning methods, these conventional procedures become inapplicable. Furthermore, even when $P$-values are provided or derived in some manner, the inherent limitations of $P$-values, such as their inaccuracy in measuring significance or their sensitivity to sample size, can lead to an uncontrolled FDR.

In recent years, a new class of FDR control procedures has gained popularity, offering an alternative approach that does not rely on $P$-values or statistics with known null distributions. These procedures involve the competition between the original variables and their “fake” counterparts. The fake variables possess similar statistical characteristics to the original variables but are unrelated to the response variable. As far as we know, this concept of using competition for FDR control can be traced back to the target-decoy (TD) search strategy employed in mass spectrometry-based proteomics. The TD strategy utilizes competitive decoy peptide sequences to estimate and control the FDR of peptide identifications. The theoretical foundation of the TD search strategy was established by He et al. They further extended the TD approach to two-group studies by introducing competitive decoy permutations of the original samples. Their approach, known as TDFDR, enables FDR control for independent variables.

The knockoff filter is a better-known competition-based method that utilizes fake variables called knockoffs to control the FDR. This method has sparked significant interest and has led to extensive research in FDR control using knockoffs. Initially developed for variable selection in linear regression problems, the knockoff filter achieves FDR control for correlated variables in specific scenarios by constructing knockoffs that capture the inherent correlation structure of the design matrix. Subsequently, a series of methods have been proposed to relax the limitations of the original model settings, including sample size and linearity, enabling the knockoff method to address more complex situations.

Both the TDFDR and knockoff filter methods employ a competition procedure to divide variables into two groups. They utilize the number (+1) of winning “fake” variables to estimate the number of winning null variables and ultimately achieve FDR control. The inclusion of the “+1” correction is crucial for achieving FDR control. Notably, Rajchert and Keich have proposed that incorporating the “+t” term, when $t = 1$, is effectively optimal under a certain condition that holds true for a majority of practical scenarios. However, the “+1” also introduces a drawback in these competition-based FDR control procedures. Specifically, when a low FDR control level is employed, these methods may become overly conservative, leading to reduced statistical power.

While FDR has been widely used in multiple hypothesis testing, it primarily measures the overall error rate of a set of hypotheses rather than individual hypotheses. Efron et al. proposed the concept of local false discovery rate (fdr) from a new perspective. The fdr at a specific score $x$ is defined as

$$
\text{fdr}(x) = \frac{\pi_0 f_0(x)}{f(x)} = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)},
$$

where $\pi_0$ and $\pi_1$ are the proportions of null and nonnull hypotheses, respectively, $f_0$ and $f_1$ are the density functions of scores corresponding to the null and nonnull hypotheses, and $f(x)$ is the density of the mixture distribution. Defined in this way, fdr is able to capture the error rates associated with individual hypotheses, providing a more refined measure. Consequently, the focus shifts from the “control” of FDR to the “estimation” of fdr.

Numerous approaches have been developed to estimate the fdr. While these methods exhibit various innovations and effectiveness, a notable challenge arises from their reliance on $P$-values. Firstly, due to the potential
ambiguity of \( P \)-values in hypothesis testing, particularly in scenarios involving high-dimensional data, these methods offer insufficient reliability in fdr estimation. Secondly, these approaches become inapplicable when \( P \)-values are unavailable.

Several approaches have emerged that do not restrict the inputs to \( P \)-values, each making different assumptions. Efron and Tibshirani\(^{27}\) proposed an empirical Bayes approach called locfdr, which utilizes maximum likelihood and the optional central matching method to estimate fdr. Another approach, proposed by Robin et al\(^{35}\) and Guedj et al,\(^{36}\) is called “kerfdr” and is based on a semi-parametric method for fdr estimation. Assuming knowledge of the null distribution, kerfdr iteratively estimates the nonnull distribution and fdr. Stephens\(^{37}\) introduced the ash method, which employs empirical Bayes estimation under the unimodal assumption, using the estimated effect size and standard error (SE) as inputs. Jeong et al\(^{38}\) proposed a semi-parametric mixture method for fdr estimation, combining Efron’s methodology of empirical null and log-concave density estimation for the nonnull distribution. Bickel and Rahal\(^{39}\) proposed the CFDR method, which estimates fdr by transforming an estimated FDR. These methods, whether parametric or semi-parametric, rely on specific assumptions regarding the inputs, which introduces the risk of misspecification of the prior distribution. When these assumptions are not met, parametric or semi-parametric methods may fail or exhibit inferior performance. Moreover, as the complexity of the data generating mechanism increases in practical applications, the limitations of these methods become more apparent. To the best of our knowledge, the use of competitive fake variables for fdr estimation has not yet been explored. Although the PeptideProphet algorithm\(^{40, 41}\) estimates the fdr of peptide identifications based on the target-decoy competition of peptide sequences, it has not been extended to problems beyond proteomics, such as variable selection.

In approaches to FDR control and fdr estimation, the proportion of true null hypotheses (\( \pi_0 \)) plays a crucial role. The original BH procedure treats \( \pi_0 \) as unknown and allows it to be as large as one, which can significantly reduce power when the true value of \( \pi_0 \) is small. Various methods have been proposed to estimate \( \pi_0 \), many of which rely on the analysis of \( P \)-values. Storey\(^{11}\) estimates \( \pi_0 \) by leveraging the property that null \( P \)-values follow a uniform distribution between 0 and 1. Langaa et al estimate \( \pi_0 \) through nonparametric maximum likelihood estimation of the \( P \)-value density.\(^{42}\) Efron simultaneously estimates \( \pi_0 \) and the null density function (\( f_0 \)) using \( z \)-values instead of \( P \)-values, employing several optional methods that assume the normality of \( f_0 \).\(^{43}\) Despite these advancements, achieving more accurate estimation of \( \pi_0 \) remains a primary objective in current FDR research.\(^{44, 45}\)

In this paper, we propose a novel method called TDfdr for estimating fdr in error-controlled variable selection, utilizing competition-based procedures. Unlike traditional approaches that rely on \( P \)-values, TDfdr is designed to handle general scores or test statistics with or without known distributions. By treating \( \pi_0 \) as a special form of FDR (i.e., the case where all the variables are rejected), we first exploit the competition procedure to obtain an estimator of \( \pi_0 \). Subsequently, we estimate the null distribution \( f_0 \) by utilizing the competitive “fake” variables through the kernel density method. Finally, we employ a semi-parametric method to simultaneously estimate the nonnull distribution \( f_1 \) and fdr. Simulation studies demonstrate the superior performance of TDfdr in terms of fdr estimation accuracy, FDR control and power, for both the TDFDR and knockoff competition procedures. The performance of TDfdr was also evaluated on two real biomedical datasets.

The remainder of the paper is organized as follows. Section 2 introduces the competition-based procedures used for FDR control, which form the basis of the TDfdr method. Section 3 describes the assumptions TDfdr relies on, the fdr estimation procedure and implementation algorithm of TDfdr. Section 4 presents the validation of assumptions and results of comparisons on simulation data. Section 5 displays the results of real data analysis. Finally, Section 6 concludes the paper.

## 2 | COMPETITION-BASED PROCEDURES FOR FDR CONTROL

Among the existing methods for FDR control, the competition-based procedures are a new class of methods that apply to general scores other than \( P \)-values. Since TDfdr is built on the competition-based procedures, here we first introduce two competition-based procedures, TDFDR\(^{14, 15}\) and the knockoff filter\(^ {13}\). TDFDR is an approach to selecting variables that have significant differences between two groups. The knockoff filter is used for variable selection in the regression model. Both methods use the competitive fake variables, which are called the “decoys” in TDFDR, and “knockoffs” in the knockoff filter, to estimate the number of null variables which are regarded significant.
In the language of multiple hypothesis testing, each null hypothesis corresponds to the irrelevance of a variable to the outcome. In the remainder of the paper, we do not distinguish the terminologies “hypothesis” and “variable”, and “rejecting hypothesis j” corresponds to the claim of “the relevance of variable j”.

2.1 Two-group study with decoy permutations

In the case-control two-group study, suppose we have m variables \(X_1, X_2, \ldots, X_m\) observed on n subjects, \(n_1\) of which are the control samples and \(n_2 (= n - n_1)\) are the case samples, constituting the data matrix \(X \in \mathbb{R}^{n \times m}\), with \(x_{ij}\) representing the \(i\)th sample of the \(j\)th variable. Aiming at discovering the variables that are significantly different between the two groups, TDFDR tests the following \(m\) null hypotheses, \(H_j\): the joint distribution of \(X_{1j}, X_{2j}, \ldots, X_{nj}\) is symmetric, \(j = 1, 2, \ldots, m\).

That is, for any random permutation \(\pi\), the density function of \(X_{1j}, X_{2j}, \ldots, X_{nj}\) satisfies \(f_{X_j}(x_{1j}, x_{2j}, \ldots, x_{nj}) = f_{X_j}(x_{1j}, x_{2j}, \ldots, x_{nj}) = f_{X_j}(x_{1j}, x_{2j}, \ldots, x_{nj})\).

To measure the difference of a variable between two groups, a scoring function \(s\) satisfying \(s(x_{ij}, \ldots, x_{nj}) = s(\pi(x_{ij}, \ldots, x_{nj}), \pi(x_{n1+j}, \ldots, x_{nj}))\) is needed. Without loss of generality, we assume that larger values of scores represent larger differences between groups.

For each variable \(j\), TDFDR first calculates an “original score” \(S_j^o = s(x_{ij}, \ldots, x_{nj})\) for the original sample and \(N\) “permutation scores” \(S_j^p = s(\pi(x_{ij}, \ldots, x_{nj})), k = 1, 2, \ldots, N\) on the permuted samples and then sorts the \(N + 1\) scores in descending order.

Next, TDFDR assigns each variable a label of “target” or “decoy,” by comparing the original score with a statistic of the \(N\) permutation scores, where the statistic can be maximum, median, or other statistics representing the population characteristic of the permutation scores. Here we utilize the “median” rule to label variable \(j\) as

\[
L_j = \begin{cases} 
T, & R_j < (N + 2)/2 \\
T/D, & R_j = (N + 2)/2 \\
D, & R_j > (N + 2)/2,
\end{cases}
\]

where \(R_j\) is the rank of the original score in the \(N + 1\) scores, \(T\) and \(D\) represent “target” and “decoy,” respectively. Note that if \(N\) is an even number, there exists the possibility of \(R_j = (N + 2)/2\), and we label the corresponding variable randomly as “target” or “decoy” with an equal probability, that is, \(P(L_j = T) = P(L_j = D) = 1/2\). This labeling creates a division of the variables. We define \(T := \{j = 1, 2, \ldots, m : L_j = T\}\) and \(D := \{j = 1, 2, \ldots, m : L_j = D\}\).

After each variable is assigned a label, a final score is determined for it:

\[
S_j = \begin{cases} 
S_j^o, & j \in T \\
\tilde{S}_j^o, & j \in D,
\end{cases}
\]

where \(\tilde{S}_j\) is the sorted \(N + 1\) scores. When the label is target, we set the original score as the final score directly; when the label is decoy, the final score is set as the permutation score ranking at the symmetric position of \(R_j\) about the median score.

There exist other ways to label and score the variables, which could enhance power.\(^{14,16,46}\)

The variables are then sorted according to their final scores decreasingly, and the label of the \(j\)th variable in the sorted list is denoted by \(L_{(j)}\). With this notation, a higher-scored variable with \(L_{(j)} = T\) potentially has a less possibility of being a true null.

Finally, those variables with \(L_{(j)} = T\) and \(j \leq K_{ad}\) are selected, where \(K_{ad}\) is determined by

\[
K_{ad} = \max \left\{ k = 1, \ldots, m : \frac{\# \{ j \leq k : L_{(j)} = D \} + 1}{\# \{ j \leq k : L_{(j)} = T \}} \leq q \right\}.
\]

The selected variables are considered significantly different between the case and control groups. It can be proven that the TDFDR procedure controls the FDR under level \(q.15\)
Note that the proof of FDR control of the TDFDR method relies on the assumption of independence between variables, though practically the TDFDR method shows good control of FDR for data with dependency.

### 2.2 Variable selection with knockoffs

As mentioned before, the knockoff filter\(^{13}\) is an influential competition-based method for FDR control, mainly used in the context of linear regression model: \(y = X\beta + z\), where \(y \in \mathbb{R}^n\) is a vector of responses, \(X = (X_1, \ldots, X_m) \in \mathbb{R}^{n \times m}\) is the design matrix, \(\beta \in \mathbb{R}^m\) is an unknown vector of coefficients and \(z \sim N(0, \sigma^2 I)\) is Gaussian noise. The regression model aims to search for the variables whose coefficients are nonzero.

The knockoff method first constructs a “knockoff” matrix \(\hat{X}\) so that it exhibits the same covariance structure as the original design matrix, but in addition, the correlations between distinct original and knockoff variables are the same as those between the distinct variables in the original matrix. Let \(\Sigma = X^T X\) be the Gram matrix. By requiring the “knockoff” matrix obey that

\[
\hat{X}^T \hat{X} = \Sigma, \quad X^T X = \Sigma - \text{diag}(s),
\]

the knockoff matrix can be solved as \(\hat{X} = X(I - \Sigma^{-1} \text{diag}(s)) + UC\), where \(s\) is an \(m\)-dimensional nonnegative vector, \(U\) is an \(n \times m\) orthonormal matrix that is orthogonal to the span of the design matrix \(X\), and \(C^T C = 2 \text{diag}(s) - \text{diag}(s) \Sigma^{-1} \text{diag}(s) \geq 0\) is a Cholesky decomposition. By maximizing the diagonal entries in \(s\), the knockoff filter makes the correlations between the knockoff variables and the true signals as small as possible. Note that the construction above is only suited for the situation where \(n \geq 2m\). The knockoff filter can also be extended to \(m < n < 2m\) with certain settings.

A statistic \(Z_j\) can be computed to measure the relevance of the original variable \(X_j\) to the response variable, and similarly, \(\hat{Z}_j\) for the knockoff variable. For instance, in the Lasso model, the statistics (\(Z_1, \ldots, Z_m, \hat{Z}_1, \ldots, \hat{Z}_m\)) can be computed by solving the optimization \(\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \frac{1}{2} \|y - [X \hat{X}]\beta\|^2 + \lambda \|\beta\| \right\}\), with \(Z_j (\hat{Z}_j)\) representing the largest value of penalty tuning parameter \(\lambda\) when variable \(X_j (\hat{X}_j)\) enters the Lasso path. Note that the design matrix in the general Lasso model is replaced by \([X \hat{X}]\) (the columnwise concatenation of \(X\) and \(\hat{X}\)) to achieve competition between variables, and the length of \(\beta\) is also doubled.

In order to tease apart those variables that are in the regression model (ie, \(\beta_j \neq 0\)) from those that are not (ie, \(\beta_j = 0\)), test statistics \(W_j\)’s are constructed so that large positive values are evidence against the null hypothesis \(\beta_j = 0\). As long as a statistic satisfies the sufficiency property (ie, \(W_j\) depends only on the Gram matrix and variable-response inner products) and the antisymmetry property (ie, swapping \(X_j\) and \(\hat{X}_j\) has the effect of switching the sign of \(W_j\)), it can be chosen as a proper test statistic for the knockoff method. For instance,

\[
W_j = Z_j \vee \hat{Z}_j = \begin{cases} +1, & Z_j > \hat{Z}_j, \\ -1, & Z_j < \hat{Z}_j, \\ 0, & Z_j = \hat{Z}_j. \end{cases}
\]

Other forms of statistics that satisfy the two properties can also be chosen depending on the circumstances. In fact, the principle in the antisymmetry property is that \(W_j\) is yielded by competition between \(X_j\) and \(\hat{X}_j\).

Finally, variables are selected with \(W_j > T_{ko}\), where \(T_{ko}\) is determined as

\[
T_{ko} = \min \left\{ t \in W : \frac{\# \{ j : W_j \leq -t \}}{\# \{ j : W_j \geq t \}} + 1 \leq q \right\},
\]

where \(q\) is the FDR control level and \(W = \{ |W_j| : j = 1, \ldots, m \} \setminus \{0\}\) is the set of unique nonzero values of \(|W_j|\)’s.

It is proven that, with the exchangeability property of statistics \(W_j\)’s, the knockoff method is able to control FDR at the given threshold \(q\) under arbitrary variable dependency.\(^{13}\)
2.3 The connection of TDFDR and the knockoff filter

Note that the symmetric rule of $W_j$'s in the knockoff filter is equivalent to giving the variables a division, by the signs of $W_j$'s. Therefore, the knockoff procedure can be described in the target-decoy framework, by replacing the signs of statistics $W_j$'s with labels of “target” or “decoy” and defining the absolute values of $W_j$’s as the final scores, and vice versa.

On the whole, both TDFDR and the knockoff filter are competition-based procedures. First, they both create new “fake” variables which are called decoys or knockoffs. Second, the “fake” variables compete with their corresponding original variables to produce antisymmetric ranking statistics. That is, for true nulls, their statistics have an equal probability of being target (positive) or decoy (negative). Third, they use the same formula to compute the rejection region, that is, Equations (1) and (3). Note that there is a “+1” term in both equations, which is essential for FDR control, and was first proposed in the context of mass spectrometry based proteomics. 19,20

3 LOCAL FALSE DISCOVERY RATE ESTIMATION

Since the TDfdr method is developed from the competition-based procedure TDFDR, we continue to use the notations of TDFDR to describe TDfdr, although the framework of the knockoff filter can be applied similarly to TDfdr.

Assume that we have $m$ simultaneous tests, each of which has the null hypothesis $H_j$, $j = 1, \ldots, m$. For simplicity, we write $H_j = 0$ when the null hypothesis $H_j$ holds and $H_j = 1$ otherwise. In the situation of variable selection, the purpose of tests becomes judgment of the significance of variables, and we will describe our TDfdr method in terms of variable selection as in the TDFDR and knockoff methods. TDfdr aims to estimate the fdr of individual variables.

As in the labeling step of the TDFDR method described in Section 2, the variables are separated into the target ($T$) and decoy ($D$) groups. The final scores, $S_j$’s, of the variables in the two groups are called “target scores” and “decoy scores,” respectively. Further, we divide the variables into several subsets according to their labels and significance (Table 1). For example, $T_0 := \{ j = 1, 2, \ldots, m : L_j = T, H_j = 0 \}$, where $L_j$ is the label of variable $j$ obtained from the competition-based procedure. Note that only $T$ and $D$ are observable, and the four sets $T_0$, $T_1$, $D_0$, $D_1$ are unobservable in practice.

3.1 Assumptions

Now we introduce the assumptions that TDfdr relies on.

- Assumption 1: The probabilities of a null variable being labeled as target or decoy are equal.
- Assumption 2: The probability of a nonnull variable being labelled as decoy is negligibly small and vanishes as the sample size increases.
- Assumption 3: The final scores of decoy variables and null target variables have the same probability distribution.

Assumption 1 can be framed as $Pr(L_j = T|H_j = 0) = Pr(L_j = D|H_j = 0)$. This equality is guaranteed by the properties of the competition procedures, that is, theorem 2.1 in TDFDR 15 and Lemma 1 in the knockoff filter. 13 It is satisfied due to the fair competition between the original and fake variables among the null ones, which exhibit no difference in comparison with each other. Assumption 1 indicates that the numbers of the null target and null decoy variables follow a binomial distribution with equal selection probability.

| TABLE 1 Division of variables. |
|-------------------------------|
| Target variables               |
| Decoy variables                |
| Null variables                 | $T_0$ | $D_0$ |
| Nonnull variables              | $T_1$ | $D_1$ |
| All                           | $T$   | $D$   |
Assumption 2 is about the probability of having nonnull decoy variables, which can be expressed as $Pr\{L_j = D | H_j = 1\}$. Generally, this probability will be extremely small, if the signals of all the nonnull variables are sufficiently large. Otherwise, there may exist a positive number of nonnull decoy variables. This is more likely to happen when the sample size is small, which leads to the generation of low-quality fake variables, and thus an inadequate competition between the original and fake variables. However, as the sample size increases, the probability of having nonnull decoy variables is expected to vanish. We here give some theoretical demonstrations in a simplified two-group scenario where the permutation time $N = 1$ and the $t$-statistic is used for scoring. Following the notations introduced in Section 2.1 where the variable $j$ observed from the two groups has the sample size $n = n_1 + n_2$, we are to demonstrate that the probability of labeling a nonnull variable as decoy tends to 0 as the sample sizes of two groups $n_1$, $n_2$ go to infinity. With a sufficiently large sample size, the permutation yields fake variables which have the same means between the two groups and therefore can be treated as null variables. Denoting the $t$-statistics of the original and permutation samples of variable $j$ as $S^n_j, j = 1, \ldots, m$ and $S^p_j, j = 1, \ldots, m$, respectively, we use the absolute value of the $t$-statistic as the scoring function. In the general case where the control and case data are not normal (generally any distribution with finite variance), we can prove through the triangular array central limit theorem and the Lindeberg-Feller theorem that the $t$-statistic converges to $N(0, 1)$ in distribution under the null. Under the nonnull, it can be proved that the absolute value of the $t$-statistic almost surely diverges to infinity as the sample sizes go to infinity. Applying these to the original and permutation scores, we have $S^n_j \overset{d}{\to} N(0, 1)$ and $|S^p_j| \overset{\text{a.s.}}{\to} +\infty$. Thus we have

$$
\lim_{n_1 \to \infty} \lim_{n_2 \to \infty} Pr\{|S^n_j| > |S^p_j| \mid H_j = 1\} = 1,
$$

and according to the labeling rule in the TDFDR method, we have

$$
\lim_{n_1 \to \infty} \lim_{n_2 \to \infty} Pr\{L_{j,n} = D \mid H_j = 1\} = 0, \quad j = 1, \ldots, m.
$$

where $L_{j,n}$ denotes the label of variable $j$ with the sample size $n = n_1 + n_2$.

Although the above analysis is for the $t$-statistic, we expect that other effective scores should be also able to discriminate a relevant variable from its fake counterpart as long as the sample size is sufficiently large.

Assumption 2 is essential for the formulation in Section 3.2, where the fdrestimation is transferred from all the variables to the subset of only “target” ones, and leaves the decoy ones as insignificant directly. Later in the estimation step, decoy variables will play the role of approximating the proportion and score distribution of null target variables.

Assumption 3 is based on Assumptions 1 and 2. With Assumption 2, there are no nonnull decoy variables and we only need to compare the score distributions of (null) decoy and null target variables. With Assumption 1, the labeling process can be viewed as a Bernoulli sampling from the $m_0$ null variables with the equal probability (0.5) of being or not included in the sample. Then Assumption 3 is to mean that the score distributions of the included (target) variables and the excluded (decoy) variables are the same.

The score distribution of the null variables is mainly determined by the scoring function and the sample data. First, we consider a special case where the $t$-statistic is used as the scoring function. When the null hypothesis holds, the distribution of the $t$-statistics is asymptotically standard normal, regardless of the distribution of the sample data. Then the target and decoy null variables have the same score distribution in an asymptotic sense. Hence the score distributions of the target and decoy null variables are similar, as long as $n$ is large. Further, if a more general scoring function than the $t$-statistic is used, there may not exist a common asymptotic distribution of scores for the target and decoy variables. This might happen when, for instance, the definition of the scoring function is not only related to the significance of the variables, but also related to the distribution of the sample data. Then even for all the null variables, their score distributions may be different. However, if the score distribution of null variables is a mixture of $k$ subdistributions ($k > 1$) and $k \ll m_0$, where $m_0 = \pi_0 m$ is the number of null variables, then Assumption 3 can approximately hold. The condition $k \ll m_0$ can be generally satisfied for high-dimensional data, for which $m_0$ is usually very large and $k$ can be relatively small if the scores are well-normalized. In practice, we recommend standardizing the sample data before scoring in order to produce scores with similar distributions. Although Assumption 3 cannot be rigorously proven in theory, we believe that it can generally hold in reality.

Additionally, empirical validation of the assumptions using simulation data can be found in Section 4.2.
3.2 The TDfdr method

Now, according to Assumption 2, the aim of TDfdr becomes the estimation of the fdr of target variables, which we call “target fdr,”

\[
fdr_t(S_j) = \frac{\pi_{0t} f_{0t}(S_j)}{\pi_{0t} f_{0t}(S_j) + \pi_{1t} f_{1t}(S_j)} , \quad j \in T ,
\]

where \( \pi_{0t} \) represents the proportion of true nulls in target variables and \( \pi_{1t} = 1 - \pi_{0t} \), \( f_{0t} \) is the density function of scores of the null target variables, and \( f_{1t} \) is the density of scores of the nonnull target variables. The extra subscript “t” (eg, “\( f_{0t} \)” compared to the original “\( f_0 \)”) represents that the proportion or density is defined in terms of the target set, which is chosen during each run of the algorithm and essentially dependent on the competition between the original and fake variables.

Regarding the unknown quantities in Equation (4), TDfdr first estimates \( \pi_{0t} \) and \( f_{0t} \), in the following Sections 3.2.1 and 3.2.2. With these two quantities, TDfdr then adopts the framework of a semi-parametric method\(^{35,36} \) to iterate \( f_{1t} \), which will be explained in Section 3.2.3. In the iteration procedure, the desired target fdr values can be calculated simultaneously.

3.2.1 \( \pi_{0t} \) estimation

TDfdr employs the competition procedure to estimate \( \pi_{0t} \). The real value of \( \pi_{0t} \) is

\[
\pi_{0t} = \frac{|T_0|}{|T|} = \frac{\#\{j = 1, 2, \ldots, m : L_j = T, H_j = 0\}}{\#\{j = 1, 2, \ldots, m : L_j = T\}}.
\]

(5)

where \( T_0 \) is unknown and is of our interest.

Relying on Assumption 1, we are able to estimate the number of null target variables \( |T_0| \) using the number of decoy variables \( |D| \). So we can estimate \( \pi_{0t} \) as

\[
\hat{\pi}_{0t} = \frac{|D|}{|T|} = \frac{\#\{j = 1, 2, \ldots, m : L_j = D\}}{\#\{j = 1, 2, \ldots, m : L_j = T\}}.
\]

If the estimated \( \hat{\pi}_{0t} \) is larger than 1, it is set as 1, that is, \( \hat{\pi}_{0t} = \min \{ \hat{\pi}_{0t}, 1 \} \).

Besides, \( \pi_0 \) can also be estimated using the similar idea,

\[
\hat{\pi}_0 = \frac{2|D|}{m} = \frac{2\#\{j = 1, 2, \ldots, m : L_j = D, H_j = 0\}}{m}.
\]

(6)

Note that if the probabilities in Assumption 1 are not equal, as long as they are constants, say \( r \) for decoy and \( 1 - r \) for target, we can estimate the \( |T_0| \) as \( \frac{1-r}{r}|D| \).

3.2.2 \( f_{0t} \) estimation

The second part is the estimation of \( f_{0t} \), that is, the probability density function of the scores of true null target variables. According to Assumption 3, we can use the distribution of decoy scores to estimate \( f_{0t} \). Here we use the kernel density estimation for implementation:

\[
\hat{f}_{0t}(S|h_0) = \frac{1}{|D|h_0} \sum_{j \in D} K \left( \frac{S - S_j}{h_0} \right),
\]

(7)

where the function \( K(\cdot) \) represents the kernel function, which we choose to be the Gaussian kernel. \( h_0 \) is the bandwidth of the kernel density estimation and we use the decoy scores to select an optimal bandwidth through 10-fold...
unbiased cross-validation (UCV). Specifically, we use the “ucv()” function in the “MASS” library in R to implement 10-fold cross-validation bandwidth selection.

3.2.3 \( f_{1t} \) and \( fdr_t \) estimation

\( f_{1t} \) is the density of scores of nonnull target variables. We use the target scores to estimate it following the iterative framework of kerfdr. First, for simplicity, we define a quantity \( p_j \) as

\[
p_j = \frac{\pi_1 f_{1t}(S_j)}{\pi_0 f_{0t}(S_j) + \pi_1 f_{1t}(S_j)}, \quad j \in \mathcal{T},
\]

which represents the probability of correctly rejecting hypothesis \( j \). Obviously, the \( fdr \) of variable \( j \) is

\[
fdr_t(S_j) = 1 - p_j, \quad j \in \mathcal{T}.
\]

In the iteration process, the aim is to estimate \( f_{1t} \) and \( p_j \)’s simultaneously. The process is described below:

1. Initiation.
   For the variable \( j^* \) with the highest score, set \( \hat{p}_j^{(0)} \) to 1, and set \( \hat{p}_j^{(0)} \) \( j \in \mathcal{T} \setminus \{j^*\} \) to 0. When there are ties in the highest scores, set all of the corresponding \( \hat{p}_j^{(0)} \)’s to 1.

2. Iteration
   (2.1) Estimation of \( f_{1t} \).
   We estimate \( f_{1t} \) using the kernel density estimation method as well. The estimate in the \( l \)th iteration is

\[
\hat{f}_{1t}^{(l)}(S|h_1) = \frac{\sum_{j \in \mathcal{T}} \hat{p}_j^{(l-1)} K\left(\frac{S - S_j}{h_1}\right)}{h_1 \sum_{j \in \mathcal{T}} \hat{p}_j^{(l-1)}}, \quad l \geq 1,
\]

where the function \( K(\cdot) \) is still chosen as Gaussian, and \( h_1 \) is optimized similarly as the way of obtaining \( h_0 \) through 10-fold UCV; instead, the scores used are target ones.

(2.2) Updating \( \hat{p}_j \)’s.
   Having \( \hat{x}_{0t}, \hat{f}_{0t} \) and \( \hat{f}_{1t} \), we update \( \hat{p}_j \) for variable \( j \) as

\[
\hat{p}_j^{(l)} = \frac{\hat{x}_{1t} \hat{f}_{1t}^{(l)}(S_j)}{\hat{x}_{0t} \hat{f}_{0t}(S_j) + \hat{x}_{1t} \hat{f}_{1t}^{(l)}(S_j)}, \quad j \in \mathcal{T}, \quad l \geq 1.
\]

3. Stopping criterion.
   Stop the iteration if \( l \geq l_{\text{max}} \) or \( \max_j \left| \frac{\hat{p}_j^{(l)} - \hat{p}_j^{(l-1)}}{\hat{p}_j^{(l-1)}} \right| < \epsilon, \quad j \in \mathcal{T} \), where \( l_{\text{max}} \) is the maximal number of iterations and \( \epsilon \) is a minor value which we choose as a threshold. Otherwise, go back to the “Iteration” step.

4. \( fdr \) estimation.
   Finally, with the optimal \( \hat{p}_j \)’s estimated, the objective \( fdr \) can be estimated as

\[
\hat{fdr}_j^{(l)} = 1 - \hat{p}_j^{(l)}, \quad j \in \mathcal{T}.
\]

3.2.4 | The TDfdr algorithm

The whole TDfdr algorithm is described in Algorithm 1. It takes the labels and scores of variables as the input and outputs the \( fdr \) estimates for target variables.
Algorithm 1. TDfdr algorithm for fdr estimation

**Input:** Labels $L_j$ and scores $S_j$ of variables, $j = 1, 2, \ldots, m$.  
**Output:** fdr estimates for target variables

1. $\pi_0$ estimation: $\hat{\pi}_0 = \frac{|D|}{|T|} = \frac{\#\{j=1,\ldots,m; L_j=0\}}{\#\{j=1,\ldots,m; L_j=D\}}$, $\hat{\pi}_0 = \min \{\hat{\pi}_0, 1\}$, $\hat{\pi}_1 = 1 - \hat{\pi}_0$

2. $f_0$ estimation: $\hat{f}_0(S|h_0) = \hat{f}_0(S|h_0) = \frac{1}{|D|h_0} \sum_{j \in D} K\left(\frac{S_j - \mu_{h_0}}{\sigma_{h_0}}\right)$

3. Initiation: $\hat{p}_j^{(0)} = 1$, for $j^* = \arg \max J_j$ and $\hat{p}_j^{(0)} = 0$, $j \in T \setminus \{j^*\}, l = 1$

4. repeat

5. estimate $f_1$: $\hat{f}_1^{(l)}(S|h_1) = \frac{\sum_{j \in T} \hat{p}_j^{(0)l} K\left(\frac{S_j - L_j}{\hat{\mu}_{l}(h_1)}\right)}{\sum_{j \in T} \hat{p}_j^{(0)l}}$, $l \geq 1$

6. update $\hat{p}_j$: $\hat{p}_j^{(l+1)} = \frac{\hat{p}_j^{(0)l} / \hat{f}_1^{(l)}(S|h_1)}{\sum_{j \in T} \hat{p}_j^{(0)l} / \hat{f}_1^{(l)}(S|h_1)}$, $j \in T, l \geq 1$

7. $l = l + 1$

8. until $\max_j \left|\frac{\hat{p}_j^{(l+1)} - \hat{p}_j^{(l)}}{\hat{p}_j^{(l)}}\right| < \epsilon, j \in T$ or $l \geq l_{\text{max}}$

9. fdr estimation: $\hat{fdr}_j^{(l)} = 1 - \hat{p}_j^{(l)}, j \in T$

10. return $\hat{fdr}_j^{(l)}, j \in T$

*h_0, h_1$ are determined by 10-fold unbiased cross-validation.\(^{47}\)

Our TDfdr algorithm is inspired by the PeptideProphet algorithm\(^{40,41}\) used in proteomics to estimate the fdr of peptide identifications. Although similar in framework to PeptideProphet, TDfdr has its own innovations and contributions. First, PeptideProphet has been limited by the peptide identification problem in proteomics, and its applicability and potential to general multiple hypothesis testing have not been studied in the past. We extended PeptideProphet from the peptide identification problem to error-controlled variable selection, a hot topic in high-dimensional data analysis. The extension is new and should be valuable to the field of statistics. Second, our TDfdr method takes advantage of the decoy information to estimate $\pi_0$, the null proportion. This is very different from the conventional iterative manner used by PeptideProphet. As a consequence, we offered a new approach to $\pi_0$ estimation, which is still an unsatisfactorily addressed important problem in current FDR and fdr research. Third, we pointed out the assumptions that the competition-based fdr estimation relies on, and provided theoretical and empirical justifications for them. Particularly, Assumptions 2 and 3 are not required by the competition-based FDR control procedures and were not investigated before. Our results should be of interest to the community. Last, considering that competition-based FDR control procedures, for example, the knockoff filter, has launched a revolution in the research of multiple hypothesis testing in recent years, the combination of competition with fdr estimation becomes natural and necessary. To our knowledge, TDfdr is the first algorithm that estimates fdr with and for the competition procedures.

## 4 | SIMULATION

To evaluate the performance of our TDfdr method, we carried out simulations on the two-group study and the regression model. We first describe the simulation design in Section 4.1, which contains the data generating and method parameters in the two scenarios, as well as the criteria of performance evaluation. To validate the assumptions of TDfdr, we carried out simulations on a representative part of the simulation data. The results are shown in Section 4.2. Sections 4.3 and 4.4 display the main results of the simulation in two scenarios, including the comparison of $\pi_0$ accuracy, fdr accuracy, FDR control, and power.

As for the fdr estimation methods, we mainly compared TDfdr with locfdr\(^{27}\) and ash.\(^{37}\) Moreover, TDFDR\(^{15}\) and the knockoff filter\(^{13}\) were also tested when performing comparisons in terms of FDR.

### 4.1 | Simulation design

We carried out repetitive simulations for $M$ times, and compared the results in average. A complete simulation contains (1) generating random samples according to the predefined parameters; (2) computing labels and scores from the generated
samples using a competition procedure; and (3) estimating the FDR and corresponding results of variables based on the computed statistics.

For TDfdr, we input the scoring statistics and estimated the FDR of target variables. For locfdr, we first tested the standard locfdr method by inputting the statistics directly, and estimated the FDR of all variables, which we call the locfdr method. Then, a transformation was made to the statistics to better satisfy the assumption of “normal distribution under null hypothesis”. The locfdr with this kind of transformation is referred to as the locfdr+ method. For ash, we input the effect calculated from the sample and the corresponding SE to the method, and obtained the FDR and FDR (q-value) results from the outputs.

4.1.1 Two-group study

For the two-group simulation, we chose two scenarios, where the data were sampled from normal or gamma distributions.

In the normal scenario, we sampled all the control data from N(0, 1); in contrast, the case data were sampled from N(0, 1) for the null hypotheses and N(±a, 1) for nonnull hypotheses, respectively, where the mean parameter a controls the difference between the two groups, and the nonnull data of N(±a, 1) are sampled randomly with the number of each group being half of the total nonnulls. We simulated different configurations of a = 2, 2.5, 3.

In the gamma scenario, we sampled all the control data from Ga(2, 1); the case data were sampled from Ga(2, 1) for null hypotheses, and the combination of Ga(a, 1) and Ga(1/a, 1) randomly and equally for nonnull hypotheses, respectively. In the simulations, a was set as 6, 7, 8. For other parameters in both normal and gamma scenarios, we chose the number of hypotheses m = 10000, that is, each sample contained 10000 variables; the sample size of each control/case group g = 5; the proportion of null hypotheses π₀ = 0.8, 0.9, 0.95. After the generation of random samples, the absolute value of the t-test statistic was used as the scoring function to characterize the differences between the case and control groups. In this way, the higher a score is, the larger difference it represents between the two groups.

The number of permutations in TDfdr was set as N = 19. The locfdr method was implemented through the “locfdr” package in R. The parameter of estimating null distribution was chosen to be “maximum likelihood” (nulltype = 1, the default). The transformation of the t-statistic S in the two-group study is $S = \phi^{-1}(\rho(S))$, where $\Phi$ and $\psi$ are the cumulative density function of the standard normal distribution and the probability density function of the t distribution, respectively. For ash, the FDR and FDR results were obtained using the “ashr” package in R. Specifically, we used the function ash.workhouse() to implement the FDR estimation with the additional argument “method = ‘fdr’”. The input effect of sample was calculated as the difference in the means of the two groups, and the input SE is the corresponding pooled SE.

Note that for different methods the objectives are different, with $T^{(k)}$ representing the variable set of interest in the kth repetition, which can be set as all variables $\{1, \ldots, m\}$ or the target variables denoted by $T^{(k)}$. Here in the two-group simulation, we set $T^{(k)} = T^{(k)}$ for TDfdr, and $T^{(k)} = \{1, \ldots, m\}$ for locfdr−, locfdr+ and ash, for $k = 1, 2, \ldots, M$.

In the FDR evaluation, we added the TDFDR package for comparison. The permutation time of TDFDR was set as 19, and the labeling rule is the median rule. Details of the TDFDR procedure is described in Section 2.1.

4.1.2 Regression model

We used the regression model described in Section 2.2 and simulation settings of the knockoff filter to test the performance of TDfdr.

First, the design matrix $X = (X_1, \ldots, X_m) \in \mathbb{R}^{n \times m}$ was generated row by row i.i.d. from an N(0, Θ) distribution, where $\Theta_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, m$. Then we centered and normalized the columns of $X$ and calculated the simulation value of $y$ as $y = \beta_1 X_1 + \cdots + \beta_m X_m + z$, where $z \sim N(0, I_n)$ is the Gaussian white noise, $m_1$ is the number of significant variables among all $m$, and $\beta_j$, $j = 1, \ldots, m_1$ are the coefficients of the significant variables. Thus, the null proportion for these $m$ hypotheses is $\pi_0 = 1 - m_1/m$.

We simulated $m = 2000$ variables in the regression model, in which $m_1 = 100, 200$. 400 variables were significant, that is, $\pi_0 = 0.95, 0.9, 0.8$. For each variable, 5000 samples were simulated from the multiple normal distribution. Besides, we simulated two cases of variables with and without dependency, corresponding to the correlation coefficient $\rho = 0$ and
$\rho = 0.3$. To vary the difficulty of variable selection, we sampled $\beta_j$, $j = 1, \ldots, m_1$ randomly from $\pm A$ for each of the $m_1$ selected coefficients, where the signal amplitude $A = 2.5, 3.5, 4.5$.

Fitting the data in the Lasso model with the concatenated design matrix from the original and knockoff ones, we computed the statistics $W_j$’s as in the example of Section 2.2. For the knockoff-based simulation, $\text{fdr}$ can be estimated as in the two-group study, yet using the Lasso statistic $W_j$’s. The locfdr method was also used to estimate the $\text{fdr}$ for comparison. Locfdr— took the original Lasso statistics as input, but failed to complete valid fdr estimation due to the violation of the normal assumption of inputs, thus was not displayed as the member of comparison. In carrying out locfdr+, there was also an obstacle when transforming the Lasso statistics, because the theoretical null distribution of them is unknown. Therefore, we leveraged the decoy/knockoff variables to estimate an empirical null, then lent it to locfdr+ to do transformation on the remaining “target” statistics to get the corresponding estimates. Formally, the transformation for a Lasso statistic $S$ in the regression model is $S' = \Phi^{-1}(\Psi(S))$, where $\Phi$ and $\Psi$ are the cumulative density function of the standard normal distribution and the empirical cumulative density function estimated from the decoy variables, respectively. As a result, the $\pi_0$ and $\text{fdr}$ estimations of locfdr+ are in terms of the target variables, that is, $I^{(k)} = T^{(k)}$, $k = 1, 2, \ldots, M$. To be precise, here the $\pi_0$ is in fact $\pi_{0t}$. For TDfdr and ash, the sets of interest are still the target variables and all variables as in the two-group setting, respectively. For the inputs of ash, since it is difficult to derive the SE of the $\lambda$ score in the Lasso model, we fit a linear model and input the estimates of coefficients and the corresponding SEs to ash. The remaining parameters of TDfdr, locfdr+ and ash were set the same as before.

FDR control and power were also evaluated, with the knockoff filter as the benchmark. The scoring function of the knockoff filter method was set the same as Equation (2). Details of the knockoff filter are described in Section 2.1.

### 4.1.3 Performance evaluation

In the simulations of two-group study and regression model, we repeated the simulation for $M = 50$ times and compared the average results, including the estimation accuracies of $\pi_0$ and $\text{fdr}$, FDR control performance, and power. For the accuracy of $\pi_0$ estimation, we compared the estimated $\pi_0$ and real $\pi_0$ by scatter plots, taking the line $x = y$ as a reference. Regarding the accuracy of $\text{fdr}$ estimation, we used the averaged sample RMSE (root mean squared error) of $M$ repetitive simulations as the metric for comparison. The estimation RMSE for the $k$th repetition is

$$\text{RMSE}^{(k)} = \sqrt{\frac{1}{|I^{(k)}|} \sum_{j \in I^{(k)}} (\hat{\text{fdr}}_j^{(k)} - \text{fdr}_j^{(k)})^2},$$

(8)

where $\hat{\text{fdr}}_j^{(k)}$ and $\text{fdr}_j^{(k)}$ are the estimated and real $\text{fdr}$ for the variable $j$ in repetition $k$, respectively; and $I^{(k)}$ is the set of interest in repetition $k$, which was set in the precious two sections for the two-group study and regression model. Then, the average RMSE over all repetitions, that is, $\text{RMSE}_{\text{ave}} = \frac{1}{M} \sum_{k=1}^{M} \text{RMSE}^{(k)}$, was used as the performance metric.

As mentioned before, TDfdr focuses only on the $\text{fdr}$ of target variables, that is, $\text{fdr}_t$. Therefore, we need to know the real values of $\text{fdr}_t$ as given in Equation (4), for comparison with the estimated $\text{fdr}$. With all the data simulated artificially, whether a variable belongs to the null set is known, and thus the real value of $\pi_0$ can be directly computed according to Equation (5). Next, the real $f_{0t}$ and $f_{1t}$ are computed through kernel density estimation using the final scores of target variables with known labels. Specifically, we use the final scores of the null target variables to estimate the real target null density $f_{0t}$ and the final scores of the nonnull target variables to estimate the real target nonnull density $f_{1t}$. That is,

$$f_{0t}(S|h_{0t}) = \frac{1}{|T_0|h_{0t}} \sum_{j \in T_0} K \left( \frac{S - S_j}{h_{0t}} \right),$$

$$f_{1t}(S|h_{1t}) = \frac{1}{|T_1|h_{1t}} \sum_{j \in T_1} K \left( \frac{S - S_j}{h_{1t}} \right),$$

where the explanations of parameters are analogous to those in Equation (7). The values of $\text{fdr}$ are computed similarly with regard to all variables instead of target ones.
In addition, we also evaluated our method in terms of FDR. Based on the connection between \( \text{fdr} \) and FDR:

\[
\text{FDR}(x) = E_f\{\text{fdr}(X)|X \leq x\} \quad (\text{where } E_f\{\cdot\} \text{ represents expectations with respect to the mixture density } f(x))
\]

we can estimate FDR practically from the estimated \( \text{fdr} \) as

\[
\hat{\text{FDR}}(S) = \text{mean}\{\hat{\text{fdr}}_j : S_j > S\}, \quad j \in I
\]

where \( I \) is the set of interest and the setting of it is the same as those described in the previous two sections.

### 4.2 Validation of assumptions

We first carried out simulations to validate the assumptions. For the two-group scenario, we chose the normal data with the difference parameter \( a = 2.5 \) and gamma data with \( \alpha = 7 \), which correspond to the mediate level of the respective three difference levels. We kept the number of variables \( m = 10000 \), and varied the sample size \( g \) from 5 to 15 in order to test the influence of sample size on the assumptions. Similarly, for the regression scenario, the difference level was set as \( A = 3.5 \), the mediate level. The number of variables was \( m = 2000 \) and the sample size \( g \) was set as 5000, 6000, … , 10000. We chose the dependent setting \( (\rho = 0.3) \) to show the harder case. Both the two-group and regression simulations were repeated 20 times only, due to the large computation cost of multiple sample sizes.

We used different criteria or plots to verify the assumptions. For Assumption 1, we calculated the ratio of the number of null target variables to the number of null decoys, on the simulation data of different sample sizes. For Assumption 2, we showed the proportion of decoy variables in nonnull ones, and its variation as the sample size increases. For Assumption 3, we drew split violin plots to compare the score distributions of decoy variables and null target ones. Figure 1 shows the results, with Figure 1A, B, and C corresponding to Assumptions 1, 2, and 3, respectively. Each subfigure contains three plots, which correspond to the normal \( a = 2.5 \) scenario, the gamma \( \alpha = 7 \) scenario, and the regression \( A = 3.5 \) scenario, respectively. In Figure 1A, B, the light blue points represent the results of individual simulations, and the dark blue points represent the means of 20 repetitions. For Figure 1C, only one repetition is shown for each sample size.

As depicted in Figure 1A, the ratios of target to decoy null variables are dispersed around one in all three scenarios, providing overall support for Assumption 1. However, the regression results exhibit larger variances, and the means for sample sizes of 5000 and 6000 deviate further from one. This deviation could potentially be attributed to the randomness and imperfect ability of the Lasso score in distinguishing the weak-signal variables, particularly when the sample size is not sufficiently large.

In Figure 1B, we see that the proportion of nonnull decoys in the two-group scenario is below 1% for all tested sample sizes and decreases to zero when the sample size is larger than 7 and 11, respectively, for the normal and gamma data. For the regression scenario, the proportion is higher, but it decreases monotonically as the sample size increases. Due to the heavy burden of knockoff computation, we did not carry out simulations on larger sample sizes, in which case we believe that \(|D_1|\) can become small enough if not zero. In summary, we empirically verified Assumption 2 that \(|D_1|\) is small and decreases as the sample size increases. In practice, the actual ratio of \(|D_1|\) to \(m_1\) and its decreasing rate are related to the data type/distribution and the scoring statistic.

Finally, Assumption 3 is validated by Figure 1C, where the split violin plots show similar distributions between decoy and null target variables, for normal, gamma, and regression scenarios. In addition, the two-sample Kolmogorov-Smirnov test was also carried out to validate Assumption 3 in the three scenarios. For each scenario, we calculated \( P \)-values for 20 repetitions in different sample sizes. We counted the numbers and proportions of rejections with the significant level 0.05 for the three scenarios. Specifically, among the 220 hypotheses (11 sample sizes * 20 repetitions), six for the normal scenario and 16 for the gamma were rejected, with the rejection proportions 2.7% and 7.3%, respectively. For the regression scenario, 10 out of 120 (6 sample sizes * 20 repetitions) were rejected, taking the proportion of 8.3%. The rejection proportions are small for all these scenarios, which indicates that for most cases there is no significant difference between the score distributions of the decoy and null target variables. Overall, the decoy density is a decent approximation of the null target density.

### 4.3 Simulation results on the two-group study

The simulation results are arranged in three parts, that is, \( \pi_0 \) estimation, fdr estimation, and FDR-related results. For the latter two, we only show the results when \( \pi_0 = 0.8 \) in the main text, and more results are given in Section S2.2 of the Appendix S1.
Figure 1 Validation of the assumptions on simulation data. (A) Ratios of null target to null decoy variables; (B) Proportions of decoy variables in nonnulls; (C) Split violin plots of decoy and null target final scores.
4.3.1 Results of $\pi_0$ estimation

Due to the importance of $\pi_0$ in fdr estimation, we first compared the accuracy of $\pi_0$ estimation by the TDfdr and locfdr methods.

Note that $\pi_0$ could not be compared because locfdr works on all variables instead of target ones. We chose a series of real $\pi_0$ values for simulation, and obtained $\pi_0$ estimates using the TDfdr and locfdr methods. Two $\pi_0$ estimates were obtained by the locfdr method, locfdr− and locfdr+, respectively. For the real values of $\pi_0$, we chose $\pi_0 = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99,$ and $1$. The scatter plots of $\pi_0$ estimates for normal and gamma data are shown in Figure 2A,B.

It is shown that TDfdr estimated $\pi_0$ more accurately than locfdr, especially in the cases where the real $\pi_0$ was less than 0.8. In most cases, TDfdr achieved the highest accuracy. Moreover, TDfdr kept a decent level of variances of estimations.

4.3.2 Results of fdr estimation

The boxplots of fdr estimation RMSEs of the normal and gamma data ($\pi_0 = 0.8$) are shown in Figure 3A,B, respectively. For the results corresponding to other values of $\pi_0$, see Figures S2 and S3.

**FIGURE 2** $\pi_0$ estimation results of two-group data. (A) Estimated vs real values of $\pi_0$ of normal data; (B) Estimated vs real values of $\pi_0$ of gamma data.
As Figure 3A shows, locfdr+ has smaller RMSEs than locfdr−, indicating that the transformation of input scores increased the accuracy of locfdr. In the comparison between TDfdr and other methods, the RMSEs of TDfdr are significantly smaller than those of locfdr−, locfdr+, and ash. Though decreasing as the group difference (a) increases, the RMSEs of ash remain the largest. Moreover, the variances of TDfdr’s RMSEs are also relatively small. Overall, the fdr estimation by TDfdr is more accurate and stable than locfdr and ash for normal data.

For gamma data, the boxplots of fdr RMSEs in Figure 3B show a similar trend to normal data. The RMSEs of TDfdr are much lower than those of locfdr− and ash, and are lower than or comparable to those of locfdr+.

As shown in Figures S2 and S3, locfdr+ and ash give larger RMSEs than TDfdr for both normal and gamma data. In contrast, the medians of RMSEs of TDfdr remain less than 0.05 in all cases, and the variances of TDfdr’s RMSEs are less than or comparable to those of locfdr and ash, demonstrating TDfdr’s robustness and stability to various conditions.

### 4.3.3 Results of FDR comparison

With the FDR calculated from fdr, the performance of TDfdr was also evaluated in terms of FDR control and power. For FDR control, we calculated the realized FDR of rejected variables as the mean of observed false discovery proportions.

![Figure 3](image-url) **Figure 3** Local false discovery rate (fdr) estimation results of two-group data. (A) RMSEs of fdr estimation of normal data (π₀ = 0.8); (B) RMSEs of fdr estimation of gamma data (π₀ = 0.8).
(FDPs) in all 50 repetitions, and then drew plots of realized FDR vs FDR control threshold, to see whether different methods are able to control FDR under varying thresholds. Points lying under the dashed line \( x = y \) represent good control of FDR.

For normal data, Figure 4A and Figure S4 show that locfdr− seriously failed to control FDR, and ash output unsatisfying FDR control results with only two to three successes in the \( a = 3, \pi_0 = 0.8 \) configuration. The remaining three methods (TDfdr, locfdr+, and TDFDR) yielded acceptable control of FDR. More specifically, locfdr+ performed conservatively in some cases, and in contrast, both TDfdr and TDFDR obtained realized FDRs closer to the nominal ones.

For gamma data, similar results were observed in Figure 4B and Figure S5. Both TDfdr and locfdr+ achieved decent FDR control, while TDfdr realized the FDR closer to the thresholds than locfdr+. TDFDR behaved even more liberally than TDfdr, controlling the FDR to the exact level of nominal ones. While ash represents more serious inflation of FDR than that in normal.

In general, among all the methods, locfdr+ estimated the FDR most conservatively, which may sacrifice power when selecting important variables. On the contrary, the realized FDR by locfdr− and ash deviated far from the thresholds, meaning serious failure in FDR control. TDfdr and TDFDR are the two which controlled the FDR most closely to the given thresholds, though in some cases the FDR slightly got out of control.

For TDfdr, the inflation of FDR occurred more frequently when the effect size was small, \( \pi_0 \) was large and the FDR threshold was small, in which case the number of variables with high scores (which can surpass the selection threshold) is small. Among these high-scoring variables, there may be an extremely small number of null ones, which might be labelled as decoy or target in the process of competition. Due to the randomness of the sample data and competition procedure, the number of high-scoring decoy variables can be more, or less with equal probability, than the number of high-scoring null target variables. On the one hand, if the number of high-scoring decoy variables is more than that of the null target, then the score distribution of null targets, \( f_{0t} \), will be overestimated in the tail domain, leading to overestimated FDR, and consequently selection results with lower realized FDR than the nominal threshold. In some extreme cases where the FDR thresholds are very small, the overestimated FDR can result in selection results including no null variable, which produces the FDP equaling zero. On the other hand, if the number of high-scoring decoy variables is less than that of the null target, then \( f_{0t} \) will be underestimated in the tail, and consequently caused underestimated FDR. While after selection through the FDR threshold, the liberal FDR estimation will lead to selected variables containing more nulls than expectation, resulting in a much higher FDP than the nominal threshold. For example, if 10 variables are selected with the nominal FDR threshold = 1%, among which there is only 1 null variable, then the FDP for this time is as high as 10%, which is much higher than 1%. Among the 50 repetitions of simulation, we found that these two cases happened with nearly equal frequencies, producing one half of zero FDPs and another half of high FDPs, which resulted in the realized FDR (mean of FDPs) far beyond 1%. We also calculate the medians of FDPs for several cases, it showed that the medians are much close to the nominal FDR thresholds. This kind of inflation of FDR also happens more frequently when the sample size is small, which aggravated the unbalance of the numbers of high-scoring decoy and null target variables. However, the simulation results showed that increasing the sample size from \( g = 5 \) to \( g = 20 \) perfectly avoided the inflation of FDR for TDfdr (see Section S2.2.3 in Appendix S1).

The average power results for normal data are plotted in Figure 4C and Figure S6. Due to the uncontrollable FDR by locfdr−, we here only display the power results of the remaining three methods. TDfdr performed decently in most of the situations, with robustness to different configurations. Compared to locfdr+, TDfdr obtained comparable or better results in most cases, but became slightly worse in some easier cases where the group difference was larger (shown in the right-hand side of Figure S6). Ash gave distinctive power results which surpass all other methods on all the configurations, yet the poor FDR controlling makes these results less meaningful considering the more important quality of the selections.

For gamma data, Figure 4D and Figure S7 show that TDfdr produced power results all surpassing locfdr+, while locfdr+ output almost all-zero power in the hardest case (when \( a = 6 \) and \( \pi_0 = 0.8 \)). Similarly, ash obtained the highest power results on most of the configurations, which would be preferable if the FDR controlling could be done better.

To sum up, TDfdr showed higher power, especially for gamma data. Locfdr+ achieved slightly superior performance to TDfdr in a few cases, but from the overall perspective, it was less stable to produce a valid selection of variables. Ash output the best power results but was not able to control FDR well in most cases, which might be due to the violation of the “unimodal assumption,” or in other words, it is not suitable to deal with the data in our simulations.
FIGURE 4  False discovery rate (FDR) control and power results of two-group data ($\pi_0 = 0.8$). (A) FDR control results of normal data; (B) FDR control results of gamma data; (C) Power results of normal data; (D) Power results of gamma data.
4.4 Simulation results on knockoff-based variable selection

As a competition-based procedure, the knockoff filter calculates statistics with different signs and absolute values, which play the similar roles of labels and final scores, respectively, in the framework of TDFDR. We used the statistics of the knockoff filter as input for TDfdr and locfdr to estimate fdr, and demonstrated the universality of our method on competition-based procedures. In addition, we evaluated the FDR results in comparison with the knockoff filter, illustrating the ability of our method in FDR control and power. Again, we show here only the results of $\pi_0 = 0.8$ for fdr estimation, FDR control and power calculated in 50 repetitive simulations, and give more results in Section S2.3 of Appendix S1.

4.4.1 Results of $\pi_0$ estimation

Figure 5A,B corresponding to independent and dependent cases respectively, shows the $\pi_0$ estimation results of TDfdr and locfdr+. It can be seen that locfdr+ mistakenly estimated $\pi_0$ to be one, while the estimation by TDfdr was much more accurate.

**Figure 5** $\pi_0$ estimation results of regression data. (A) Estimated vs real values of $\pi_0$ for independent cases; (B) Estimated vs real values of $\pi_0$ for dependent cases.
4.4.2 Results of fdr comparison

Figure 6A and Figure S12 compare the fdr estimation RMSEs of TDfdr, locfdr+ and ash for independent variables. For all three signal amplitudes (A) when \( \pi_0 = 0.8 \), TDfdr yielded more accurate estimations than locfdr+ and ash. The full results in Figure S12 show that for all the cases but one (A = 2.5 & \( \pi_0 = 0.9 \)), TDfdr estimated fdr with less median error, demonstrating superior performance to the other two methods.

Figure 6B and Figure S13 show the comparison of fdr for variables with dependency. In the existence of dependency, the number of wins of TDfdr against locfdr+ decreased compared to the independent cases. However, the advantage of TDfdr over locfdr+ is still obvious overall, despite the slight inferiority in some cases to locfdr+. While ash output less competitive results in most cases.

In brief, these results demonstrate that for most configurations of null proportions and signal amplitudes, TDfdr yielded more accurate fdr estimation than locfdr and ash.

4.4.3 Results of FDR comparison

For FDR, we compared the deduced FDRs from fdr estimated by TDfdr, locfdr and ash, with the FDR given by the knockoff filter.

As shown in Figure 7A, B, for different FDR thresholds, TDfdr, locfdr+, and the knockoff filter all succeeded in controlling FDR when \( \pi_0 = 0.8 \), even with the dependency between variables, while ash failed the FDR control in all the cases. Moreover, among the three methods which are able to control FDR, TDfdr controlled FDR most liberally, with the realized FDR closest to the line \( x = y \). In contrast, the knockoff filter and locfdr+, especially the latter, were too

![Figure 6](image_url)  
**Figure 6** Local false discovery rate (fdr) estimation results of regression data (\( \pi_0 = 0.8 \)). (A) RMSEs of fdr estimation for independent cases; (B) RMSEs of fdr estimation for dependent cases.
False discovery rate (FDR) control and power results of regression data ($\pi_0 = 0.8$). (A) FDR control results for independent cases; (B) FDR control results for dependent cases; (C) Power results for independent cases; (D) Power results for dependent cases.
conservative in FDR control. This trend directly led to higher power of TDfdr. However, for some cases shown in the figures in Section S2.3.2 in Appendix S1, TDfdr did not control FDR well especially when the thresholds were small. And ash succeeded in FDR control in the $A = 3.5, \pi_0 = 0.9$ and $A = 3.5, \pi_0 = 0.95$ cases.

Regarding power, for all the thresholds in both independent (Figure 7C) and dependent cases (Figure 7D), TDfdr achieved the highest power except ash, while the highest power results of ash only make sense in the two to three cases when the FDR is under control. Locfdr+ was barely able to select any significant variables, as shown by the power line around zero. This also explained why its results of control FDR were so close to zero. Note that the FDR is by definition zero when the selection set is empty.

When $\pi_0 = 0.9$ and 0.95 (Figures S14-S17), we observed similar results. TDfdr still had higher power than locfdr+ and the knockoff filter, and again, locfdr+ selected few significant variables. TDfdr controlled FDR well in all cases except when $\pi_0 = 0.95$ and the FDR threshold was 1% to 2% (independent cases).

5 | REAL DATA ANALYSIS

We applied the TDfdr method to two real datasets, including a two-sample COVID-19 dataset and a regression dataset of HIV drug resistance, and compared it with other methods.

5.1 | COVID-19 data

We utilized a dataset from samples of COVID-19 sera to evaluate the performance of TDfdr. The results from the original paper were employed as a reference. Moreover, the methods of locfdr (including locfdr- and locfdr+) and ash were tested for comparison.

The dataset was from a study of proteomic characterization of COVID-19 patient sera, and contains the measurements of 894 proteins in the serum samples of 93 subjects divided into four groups, as shown in Table 2. Missing data were processed by first deleting the all-missing proteins, and then filling the remaining missing values with zeros. As a result, the number of variables (proteins) was trimmed from 894 to 791.

To identify the significantly differentially expressed proteins related to the COVID-19 disease, the original paper first employed three case-control comparisons with the “Healthy” group serving as the control, that is, (1) Severe vs Healthy, (2) Nonsevere vs Healthy, and (3) Non-COVID-19 vs Healthy, and then reported the final proteins as the union of the first two comparisons excluding the third. For the original results, 105 differentially expressed proteins for COVID-19 patients were discovered using the combined criteria of FDR and fold-change.

Referring to the original paper, we used the 5% FDR threshold and substituted the FDR estimation results with the outcomes obtained from TDfdr, locfdr, and ash. Due to the randomness inside TDfdr, we repeated its procedure for 49 times, and calculated the median number of selected proteins. Locfdr−, locfdr+ and ash were run once, as they provide deterministic results. The number of proteins selected through the four aforementioned comparisons, the number of reported proteins, and the number of their intersections with the original 105 proteins were given in Table 3 for comparison.

In the individual case-control comparisons, the TDfdr, locfdr, and ash methods generally identified a greater number of significant proteins compared to the original results, with the exception of locfdr+ which only reported three proteins in the “Non-COVID-19 vs Healthy” comparison. In terms of the final reported proteins, both locfdr− and locfdr+ selected 144 proteins, ash selected 143 proteins, and TDfdr selected 124 proteins. However, when intersected with the original

| Table 2 | Sample grouping details. |
|---------|--------------------------|
| Group name | Group size | Description |
| Severe | 28 | Serum samples from severe COVID-19 patients |
| Nonsevere | 37 | Serum samples from nonsevere COVID-19 patients |
| Non-COVID-19 | 25 | Serum samples from non-COVID-19 patients |
| Healthy | 28 | Serum samples from healthy subjects |

\* Non-COVID-19 patients represent those who are negative for the SARS-CoV-2 nucleic acid test but have clinical characteristics similar to COVID-19 patients.
TABLE 3  Numbers of selected proteins by different methods and their intersections with the original result.

| Methods       | Severe vs Healthy | Nonsevere vs Healthy | Non-COVID-19 vs Healthy | Final report\* | Intersection of the final report with original proteins |
|---------------|-------------------|----------------------|-------------------------|----------------|-------------------------------------------------------|
| Original      | 120               | 43                   | 28                      | 105            | -                                                     |
| TDfdr         | 143               | 52                   | 31                      | 124            | 104                                                   |
| locfdr−       | 153               | 83                   | 26                      | 144            | 104                                                   |
| locfdr+       | 130               | 71                   | 3                       | 144            | 103                                                   |
| ash           | 186               | 61                   | 73                      | 143            | 87                                                   |

*Final reported proteins = (Severe vs Healthy) ∪ (Nonsevere vs Healthy) \ (Non-COVID-19 vs Healthy).

FIGURE 8  Comparison Venn plot of TDfdr and original results. “TDfdr(S-)” represents the results of set operation (Severe vs Healthy) \ (Non-COVID-19 vs Healthy) from TDfdr, and “TDfdr(N-)” represents the results of set operation (Nonsevere vs Healthy) \ (Non-COVID-19 vs Healthy) from TDfdr. Similarly, “Orig(S-)” and “Orig(N-)” represent the results of the corresponding set operations from the original study.

results, TDfdr and locfdr− both obtained 104 overlapping proteins, very close to the total of 105. Locfdr+ identified 103 overlapping proteins, slightly fewer than the former two methods. Ash yielded only 87 overlapping proteins, suggesting a potential lack of control over the realized FDR. Among these methods, TDfdr demonstrated the highest number of overlapping proteins with the fewest selections, indicating a high power with a low realized FDR.

Further, we analyzed the intersection and the difference of the protein sets detected through TDfdr and the original method. TDfdr was run three times, and the intersecting proteins were believed to be of high confidence and were subjected to analysis. We compared four sets of proteins through the Venn plot which is shown in Figure 8. For the proteins specific to the severe COVID-19 patients, the original paper reported 97 significant ones, while TDfdr reported 123 with high confidence, as shown in the areas of “Orig(S-)” and “TDfdr(S-),” respectively, in the Venn plot. Of the 97 proteins, 96 were included in the results of TDfdr. Moreover, TDfdr found another 20 severe proteins which were not reported in the original study. For the differentially expressed proteins in the nonsevere COVID-19 group (marked as “TDfdr(N-)” and “Orig(N-)”), TDfdr’s result covered all the 33 proteins that were found in the original paper, with six newly reported ones.

Regarding the 123 significant severe-COVID-19 proteins selected by TDfdr, we carried out pathway analysis to demonstrate their biological functions in the pathways. The R package “clusterProfiler” was utilized to search for important pathways and make visualizations. Figure 9 shows the results, where the x-axis of the plot represents the number of proteins in the corresponding pathways. The 20 most significant pathways are displayed, and the small values of the adjusted P-values demonstrate that these pathways were enriched from the genes with high confidence.

Among these significant pathways, two were found concordant with those reported in the original paper. Specifically, the pathway “GOBP platelet degranulation” in Figure 9 corresponds exactly to the “platelet degranulation” in the original paper, and the “GOBP complement activation” is similar to the “complement system” in the original paper. The third
FIGURE 9 Pathways associated to the 123 severe proteins.

pathway enriched in the original paper is called “Macrophage function,” which is closely related to immune response, and it also has a corresponding pathway in our results called “GOB Phumoral immune response”. In summary, the three pathway clusters found in the original paper can all correspond to the ones in our analysis results.

5.2 HIV data

To demonstrate the performance of TDfdr on the regression model, we employed it on an HIV dataset, which was also utilized to evaluate the knockoff method. 13

As described in the original paper, 13 the dataset consists of drug resistance measurements and genotype information from samples of Human Immunodeficiency Virus Type 1 (HIV-1). The task is to detect mutations in the genotypes of HIV-1 that are associated with drug resistance. Specifically, the response variable $y_i$ is given by the log-fold increase of lab-tested drug resistance in the $i$th sample, and the design matrix $X$ has entries $X_{ij} \in \{0, 1\}$, indicating presence or absence of mutation $j$ in the $i$th sample.

In the study conducted on the HIV data, we evaluated the performance of TDfdr, locfdr, ash, and the knockoff filter. We reported the significant variables at different FDR thresholds of 0.01, 0.05, 0.1, and 0.2. Specifically, we used locfdr+ as the implementation of locfdr due to the failure of locfdr− on the Lasso scores. Within locfdr+, we chose the “maximum likelihood” method to estimate the null distribution (nulltype = 1). Both locfdr+ and the knockoff filter utilized the Lasso $\lambda$ score as input. However, obtaining the SE of the Lasso $\lambda$ score was challenging. To address this, we fitted a linear model for each drug with the response and calculated the $\hat{\beta}/\hat{\Sigma}$ score to serve as the input for ash. To comprehensively assess the performance of TDfdr, we provided both the $\lambda$ score and $\hat{\beta}/\hat{\Sigma}$ score as inputs, resulting in two versions: TDfdr-lbd and TDfdr-bs, respectively. In our analysis, we focused on the resistance of seven drugs, and the names of these drugs were displayed as subtitles in Figure 10.

From the comparison between TDfdr-lbd and TDfdr-bs, the latter generally yielded more selected mutations that were verified, indicating that the $\hat{\beta}/\hat{\Sigma}$ score is more effective in measuring the significance of variables in the HIV dataset compared to the Lasso $\lambda$ score. Among the methods utilizing the $\hat{\beta}/\hat{\Sigma}$ score, ash selected more verified mutations than TDfdr-bs in most cases, demonstrating the effectiveness of ash when its assumptions were satisfied. On the other hand, the inferior performance of TDfdr-bs compared to ash may be attributed to the relatively low dimensionality (in the range of hundreds) of the HIV dataset, which generally affects the performance of nonparametric methods in comparison to parametric ones. However, this is merely a conjecture since the verified set of mutations can only serve as a reference,
FIGURE 10  Numbers of selected mutations by the knockoff filter, locfdr+, TDfdr-lbd, TDfdr-bs and ash. The bars in blue color represent the numbers of mutations that were verified in medical assays, and the orange bars represent the mutations reported by the method yet not verified. (A) FDR threshold = 1%; (B) FDR threshold = 5%; (C) FDR threshold = 10%; (D) FDR threshold = 20%.
with other mutations remaining unverified. Regarding the methods using the Lasso $\lambda$ score, TDfdr-lbd performed comparably to the other two methods (the knockoff filter and locfdr+), or better, especially in cases where the FDR threshold is small.

Overall, in most cases, the TDfdr method selected a greater number of significant mutations in total, including some unverified ones, compared to other methods that utilize the same score. The knockoff filter met the serious problem of vanished power when the FDR threshold is low (1%). This reveals a drawback of competition-based procedures, which have a “+1” correction in their FDR estimation formula to achieve FDR control. Such “+1” correction has a side effect of dramatically decreasing the power when the number of significant variables are small and the FDR threshold is low.

### 6 | CONCLUSION

In this paper, we proposed an fdr estimation method, TDfdr, with and for the competition-based procedures. Taking advantage of competitive decoy variables, TDfdr provides an innovative way of estimating $\pi_0$ and $f_0$. Then TDfdr leverages the iteration framework of kerfdr to estimate the fdr nonparametrically on target variables. Compared to many existing methods, TDfdr extends the scope of input and improves the accuracy of fdr estimation. In general, the framework of TDfdr can be applied to any competition-based procedures, such as TDFDR and the knockoff filter.

Existing fdr estimation methods such as locfdr and ash have specific assumptions about the distribution of input scores and are unable to utilize scores produced by competition procedures. In contrast, TDfdr, while having prerequisite assumptions, does not impose strict requirements on the score distribution and is applicable to general scores, including those from machine learning models. We provided theoretical analyses and validations through simulations to support our assumptions. It can be demonstrated that these assumptions generally hold as the sample size increases with high-dimensional inputs. The simulation results validated this trend and exhibited exemplary sample sizes that are easily achievable in real-world scenarios.

Simulations on two-group data and regression data demonstrated the higher accuracy and better stability of TDfdr compared to traditional methods like locfdr and the recently popular ash. TDfdr also provided more accurate estimation of $\pi_0$ than locfdr. In terms of FDR control and power, locfdr required statistic transformations to gain better FDR control, which sacrifices power. Ash exhibited the highest power but failed in FDR control for most cases. In contrast, TDfdr consistently provided stable FDR control results and, in most cases, offered the most powerful selection with controlled FDR. When compared to other FDR control methods, TDfdr also performed comparably or even superiorly to TDFDR and the knockoff filter.

The results on two real datasets demonstrated the high power of TDfdr and its robustness to FDR control levels. In the COVID-19 analysis, TDfdr identified 104 overlapping biomarkers (out of 105) compared to the original ones, with a lower FDR than the locfdr method. For the HIV data, TDfdr selected more significant mutations, including some unverified ones, especially when the FDR threshold was small.

TDfdr offers advantages over competition procedures for FDR control in two aspects. First, it provides a confidence assessment of individual variables. Second, it is less sensitive to small FDR control levels, overcoming the conservatism induced by the “+1” correction in FDR control methods, such as the knockoff filter.

In an era of high-dimensional data and the need for multiple hypothesis testing, TDfdr presents a new and accurate way to estimate fdr, capable of utilizing a wide range of scores. The idea of null proportion estimation can be explored in other frameworks related to multiple hypothesis testing. Overall, TDfdr demonstrates its effectiveness and potential for accurate fdr estimation, making it a valuable tool in multiple hypothesis testing and variable selection.

### ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (Grants 2022YFA1004801 and 2022YFA1304603) and the National Natural Science Foundation of China (Grant 32070668).

### CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

### DATA AVAILABILITY STATEMENT

The COVID-19 data that support the findings of this study are openly available in ProteomeXchange Consortium at [https://www.iprox.org/](https://www.iprox.org/), Project ID: IPX0002106000 and IPX0002171000. The HIV data that support the findings of this
study are openly available in HIVDB at http://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/. The R package for the TDfdr method can be downloaded from http://fugroup.amss.ac.cn/software/TDFDR/TDlocfdr.html.

ORCID
Xiaoya Sun https://orcid.org/0000-0001-9776-2236

REFERENCES
1. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100(16):9440-9445.
2. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plan Inf. 2004;124(2):379-398.
3. Keich U, Kertesz-Farkas A, Noble WS. Improved false discovery rate estimation procedure for shotgun proteomics. J Proteome Res. 2015;14(8):3148-3161.
4. The M, MacCoss MJ, Noble WS, Käll L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. J Am Soc Mass Spectrom. 2007;38(1):43-56.
5. Schwartzman A, Dougherty RF, Lee J, Ghahremani D, Taylor JE. Empirical null and false discovery rate analysis in neuroimaging. Neuroimage. 2009;44(1):71-82.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Methodol. 1995;57(1):289-300.
7. Tukey JW. The Problem of Multiple Comparisons. Princeton: Princeton University; 1953.
8. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc. 2001;96(456):1151-1160.
9. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002;23(1):70-86.
10. He K, Li M, Yan F, Gong F, Sun X. A direct approach to false discovery rates by decay permutations. arXiv:1804.08222 2018.
11. He K, Li M, Yan F, Gong F, Sun X. Null-free false discovery rate control using decay permutations. Acta Math Appl Sin. 2022;38(2):235-253.
12. Emery K, Hasam S, Noble WS, Keich U. Multiple competition-based FDR control and its application to peptide detection. In: Schwartz R, ed. Research in Computational Molecular Biology. Cham, Switzerland: Springer International Publishing; 2020:54-71.
13. He K, Keich U. Controlling the FDR in variable selection via multiple knockoffs. arXiv:1911.09442 2019.
14. Yang L, Fan Y, Lv J, Noble WS. DeepPINK: reproducible feature selection in deep neural networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. Advances in Neural Information Processing Systems. Vol 31. Montréal, Canada: Curran Associates, Inc.; 2018.
15. He K. Multiple Hypothesis Testing Methods for Large-Scale Peptide Identification in Computational Proteomics [Master’s thesis]. Beijing, China: University of Chinese Academy of Sciences. 2013.
16. He K, Fu Y, Zeng W-F, et al. A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. arXiv:1501.00537 2015.
17. Candès E, Fan Y, Janson L,Lv J. Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. J R Stat Soc B Stat Methodol. 2018;80(3):551-577.
18. Barber RF, Candès EJ, Samworth RJ. Robust inference with knockoffs. Ann Stat. 2020;48(3):1409-1431.
19. Barber RF, Candès E. On the construction of knockoffs in case–control studies. Stat. 2019;8(1):e225.
20. Barber RF, Ramdas A. The p-filter: multilayer false discovery rate control for grouped hypotheses. J R Stat Soc B Stat Methodol. 2017;79(4):1247-1268.
21. Rahchert A, Keich U. Controlling the false discovery rate via competition: is the +1 needed? Stat Prob Lett. 2023;197:109819.
22. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc. 2001;96(456):1151-1160.
23. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002;23(1):70-86.
24. Pan W, Lin J, Chap TL. A mixture model approach to detecting differentially expressed genes with microarray data. Funct Integr Genomics. 2003;3(3):117-124.
25. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics. 2003;19(10):1236-1242.
26. Allison DB, Gadbury GL, Heo M, et al. A mixture model approach for the analysis of microarray gene expression data. Comput Stat Data Anal. 2002;39(1):1-20.
27. Broberg P. A new estimate of the proportion unchanged genes in a microarray experiment. Genome Biol. 2004;5(5):1-10.
28. Aubert J, Bar-Hen A, Daudin J-J, Robin S. Determination of the differentially expressed genes in microarray experiments using local FDR. BMC Bioinform. 2004;5(1):1-9.
29. Liao JG, Lin Y, Selvanayagam ZE, Shih WJ. A mixture model for estimating the local false discovery rate in DNA microarray analysis. Bioinformatics. 2004;20(16):2694-2701.
34. Do K-A, Müller P, Tang F. A Bayesian mixture model for differential gene expression. *J R Stat Soc Ser C Appl Stat.* 2005;54(3):627-644.
35. Robin S, Bar-Hen A, Daudin J-J, Pierre L. A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput Stat Data Anal.* 2007;51(12):5483-5493.
36. Guedj M, Robin S, Celisse A, Nuel G. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinform.* 2009;10(1):84.
37. Stephens M. False discovery rates: a new deal. *Biostatistics.* 2016;18(2):275-294.
38. Jeong S-O, Choi D, Jang W. A semiparametric mixture method for local false discovery rate estimation from multiple studies. *Ann Appl Stat.* 2020;14(3):1242-1257.
39. Bickel DR, Rahal A. Correcting false discovery rates for their bias toward false positives. *Commun Stat Simul Comput.* 2021;50(11):3699-3713.
40. Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J Proteome Res.* 2008;7(1):286-292.
41. Ma K, Vitek O, Nesvizhskii AI. A statistical model-building perspective to identification of ms/ms spectra with peptideprophet. *BMC Bioinform.* 2012;13(16):S1.
42. Langaas M, Lindqvist BH, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J R Stat Soc Ser B Stat Methodol.* 2005;67(4):555-572.
43. Efron B. Size, power and false discovery rates. *Ann Stat.* 2007;35(4):1351-1377.
44. Neumann A, Bodnar T, Dickhaus T. Estimating the proportion of true null hypotheses under dependency: a marginal bootstrap approach. *J Stat Plan Inf.* 2021;210:76-86.
45. Biswas A, Chakraborty S, Baruah VJ. Estimation of the proportion of true null hypotheses under sparse dependence: adaptive FDR controlling in microarray data. *Stat Methods Med Res.* 2022;31(5):917-927.
46. Emery K. Controlling the FDR through Multiple Competition [PhD thesis]. University of Sydney. 2020.
47. Scott DW. *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: John Wiley & Sons; 2015.
48. Shen B, Yi X, Sun Y, et al. Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell.* 2020;182(1):59-72.

**SUPPORTING INFORMATION**
Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Sun X, Fu Y. Local false discovery rate estimation with competition-based procedures for variable selection. *Statistics in Medicine.* 2024;43(1):61-88. doi: 10.1002/sim.9942