Large-Scale Global and Simultaneous Inference: Estimation and Testing in Very High Dimensions

T. Tony Cai[1] and Wenguang Sun[2]

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, USA, 19104, email: tcai@wharton.upenn.edu

[2]Department of Data Sciences and Operations, University of Southern California, Los Angeles, U.S.A., 90089, email: wenguans@marshall.usc.edu

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–30

This article's doi:
10.1146/((please add article doi))

Keywords

Compound decision problem, dependence, detection boundary, false discovery rate, global inference, multiple testing, null distribution, signal detection, simultaneous inference, sparsity.

Abstract

Due to rapid technological advances, researchers are now able to collect and analyze ever large data sets. Statistical inference for big data often requires solving thousands and even millions of parallel inference problems simultaneously. This poses significant challenges and calls for new principles, theories and methodologies. The present paper gives a selective survey of some recently developed methods and results for large-scale statistical inference, including detection, estimation, and multiple testing. We begin with the global testing problem where the goal is to detect the existence of sparse signals in a data set, and then move to the problem of estimating the proportion of non-null effects. Finally, we focus on multiple testing with false discovery rate (FDR) control. The FDR provides a powerful and practical approach to large-scale multiple testing and has been successfully used in a wide range of applications. We discuss several effective data-driven procedures and also present efficient strategies to handle various grouping, hierarchical, and dependency structures in the data.

Contents

1. LARGE-SCALE INFERENCE

In current business and economic research, massive and complex data sets are collected routinely by governments, organizations, small businesses and large enterprises, with thousands and even millions of variables. The expansive data collection calls for new techniques for making *large-scale statistical inference*, which involves performing inferences on many study units simultaneously. One phenomenon that arises particularly frequently is sparsity: out of a large number of observations most of them are pure noise and only a small fraction contain signal, or information of interest. The identification of these sparse signals is challenging, similar to finding needles in a haystack. These new challenges have motivated the development of a plethora of novel concepts and powerful approaches to the important and rapidly growing field of large-scale inference. This article reviews significant progresses that have been made recently in this field, with a focus on multiple testing with false discovery rate control.

1.1. Examples

Large-scale inference techniques have been successfully applied in a wide range of fields, including financial economics, marketing analytics, social science, signal processing, and biological sciences such as genomics and neuroimaging. We start with several examples in business and social science research where large data sets are routinely collected from empirical studies.

- **Detection of anomalous events**. Anomaly is a pattern in the data that does not conform to the normal state or behavior. Important applications include the detection of credit card frauds, cyber intrusion, financial market anomalies, and covert communication. For example, techniques for reliably detecting and precisely locating credit card frauds are important for credit card companies to improve their service and reduce possible financial losses. To predict/detect frauds, it is necessary to monitor an enormous amount of transactions from many customers at the same time. This large-scale inference problem involves either producing massive amount of real-time estimates or testing thousands and even millions of hypotheses with high frequencies.

- **Selection of skilled fund managers**. In financial markets, monthly returns from a large number of mutual funds are routinely collected. As a guide to evaluate past and future performances, investors are interested in knowing the proportion of fund managers who possess true stock-picking skills (Barras et al. 2010). Furthermore, it is desirable to accurately identify skilled fund managers so that investors can build a portfolio that achieves outstanding performance. However, it is possible that some outperforming funds are due to luck and not special skills, whereas some skilled fund managers may underperform from time to time. The issue is further aggravated when thousands of mutual funds exist in the financial markets. The selection of skilled fund managers requires some formal principles to control false discoveries.
- **Evaluation of trading rules**. An important goal in financial economics is to test a large number of factors to explain the cross-sectional patterns and use these to develop/evaluate new trading strategies. However, the simultaneous investigation of a large number of factors gives rise to the issue of data snooping bias (Lo and MacKinlay 1990; Harvey and Liu 2015). That is, one may find seemingly significant but in fact spurious correlations in the data. Moreover, small or moderate effects, promoted by expansive data mining, may be overestimated and hence appear outstanding. To reduce data-snooping bias, investors are required to carry out an appropriate "haircut" for the reported effect size. However, most existing rules are ad hoc. For example, a common practice in evaluating trading rules is to discount the reported Sharpe ratio by 50%. It is desirable to develop more rigorous backtesting rules to account for the data mining effects with theoretical guarantees.
- **Comparison of academic performances**. The adequate yearly progress (AYP) study of California high schools (Rogosa 2003) aimed to compare academic performances of socio-economically advantaged (SEA) versus socio-economically disadvantaged (SED) students. In the AYP study, standard tests in mathematics were administered to 7867 schools and a z-score for comparing SEA and SED students was obtained for each school. The identification of "interesting" schools is an important step for making proper allocations of available funds. The policy-makers need to come up with an effective and fair ranking and selection procedure to analyze the yearly survey data. This involves carrying out thousands of significance tests simultaneously, and making decisions by taking into account other important factors such as school sizes and previous allocations of funds.

In the above examples, researchers or policy makers need to either estimate thousands of parameters or test thousands of hypotheses at the same time. This requires new theories and methodologies to overcome the limitations of classical methods that were developed for small studies. As a first step, we need a realistic and effective model to describe the data structure in large-scale inference problems; this is discussed in the next section.

1.2. A Two-Group Model

Suppose we are interested in making inference on n units, each represented by a summary statistic X. The cases are either *null* or *non-null*, with non-null cases referring to units exhibiting interesting patterns or abnormal behaviors, such as fraudulent credit card transactions, financial market anomalies, or fund managers with superior performance. In practice, we do not know the true states of nature but only observe a mixture of null and non-null cases. There are many ways to model sparse data but one of the most natural is to posit a mixture model

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} (1-\epsilon_n)F_0 + \epsilon_n F_1, \tag{1}$$

where the mixing proportion ϵ_n is small, F_0 is the null distribution and F_1 is the non-null or "alternative" distribution. Equivalently, for each $1 \le i \le n$, one assumes that X_i has probability $1-\epsilon_n$ of being a null case and probability ϵ_n of being a non-null case. Let f_0 and f_1 denote the densities corresponding to null and non-null cases respectively. The marginal density is given by $f(x) = (1-\epsilon_n)f_0(x) + \epsilon_n f_1(x)$. The mixture model (1) provides a powerful and convenient framework for large-scale inference and has been widely used in the literature (Efron et al. 2001; Storey 2002; Newton et al. 2004; Sun and Cai 2007).

1.3. Global and Simultaneous Inference

The tasks in large-scale inference are often complex: it is desirable to investigate a massive data set from different perspectives and possibly through multiple stages. One often starts with a few general questions regarding the global features of a large data set. A natural question is: are there any signals in the data set? For example, a credit card company wants to know if any fraudulent transactions have occurred in the previous period, and an internet security agency needs to decide whether there is cyber intrusion at a given time. These applications give rise to the anomaly or *signal detection* problem, which can be stated as a *global testing* problem

$$H_0^n : \epsilon_n = 0 \quad \text{vs.} \quad H_1^n : \epsilon_n \neq 0. \tag{2}$$

The proportion ϵ_n of non-null effects is an important quantity. For instance, the magnitude of ϵ_n can help make informative decisions in large-scale studies. For example, investors are interested in knowing how many fund managers possess true stock-picking skills, and policy makers need to decide how many schools should receive assistance/funds to reduce the large gaps between test scores. An interesting and technically challenging global inference problem is to obtain a good estimate of the non-null proportion ϵ_n.

However, global inference is often inadequate in many decision-making scenarios. For instance, investors might be interested in further identifying which fund managers are truly skilled, and credit card companies need to locate fraudulent transactions precisely to take further actions. In these situations, one needs to look at every individual case and decide whether it is null or non-null. This gives rise to a *multiple testing* problem, which involves making *simultaneous inference* on n hypotheses:

$$H_{i0}\text{: case } i \text{ is null vs. } H_{i1}\text{: case } i \text{ is non-null}, \; i = 1, \cdots, n. \tag{3}$$

Unlike global inference problems, the goal in simultaneous inference is to make precise decisions at individual levels, which is more challenging due to the increased precision required and new complications such as data snooping bias and multiple comparisons; these issues will be discussed next.

1.4. New Challenges

While searching for interesting features in the vast amount of data, researchers routinely investigate a large number of parallel problems at the same time, and many analyses may be conducted using the same data set. Common practices include multiple testing of thousands of hypotheses, simultaneous estimation of a large number of parameters, or frequent predictions on numerous outcomes. Making multiple inferences simultaneously without properly accounting for multiplicity can lead to misleading conclusions. For example, one may find seemingly significant but in fact spurious patterns in the data, or overestimate the strength of the selected associations.

The multiplicity effect in large-scale inference can be illustrated by the following spam email example (White 2000). Suppose a person wishes to demonstrate that he is a stock-picking genius. In Day 1, he sends emails to 102,400 individuals and makes predictions on the stock market in the next day: half are told that the market will go up and the other half down. In Day 2, those who received the wrong predictions will be discarded from the email list, and the remaining will get emails with new predictions: again, half up and half down. After ten trading days, the one hundred people who are still on the email list would have received ten correct predictions in a row. Without knowing the scheme or accounting for the multiplicity, these one hundred people must have been very impressed.

In addition to multiple predictions, the multiplicity effect is also a serious issue in large-scale estimation and testing problems, where repeated application of classical methods tends to yield severely biased estimates and inflation of false discoveries. For example, the identification of skilled fund managers requires looking through the past performances of a large number of funds and choosing a significance threshold to characterize the benchmark performance. However, not all fund managers who outperform the benchmark are skilled: some are truly skilled but some are just "lucky." Moreover, even if the selected managers do have some skills, their true performances may be overestimated substantially.

This paper gives a selective survey of some significant recent developments in large-scale inference, including detection, estimation, and multiple testing. Section 2 considers global inference; important topics include sparse signal detection and estimation of the proportion of the non-null effects. Section 3 focuses on multiple testing with false discovery rate (FDR) control. Several effective simultaneous testing procedures under various settings are presented. Open problems and other issues are discussed in Section 4.

2. GLOBAL INFERENCE PROBLEMS

We study a class of global inference problems that involve either testing or estimation of the global parameters under the mixture model (1): (i) testing the global hypothesis (2), (ii) estimating the non-null proportion ϵ_n and (iii) estimating the null distribution F_0.

2.1. Detection of Sparse Signals

The signal detection concerns testing against the global null hypothesis that there is no signal of interest in a data set. The problem arises in many applications, where a large number of variables are measured and only a small proportion of them possibly carry signal information. For example, in financial markets it is crucial to detect anomalies in early stage when only a small fraction of firms or markets are adversely affected. Other examples include the detection of disease outbreaks, credit card frauds and covert communication. In this section, we begin with the theory and methodology of a simple model and then move to more complicated settings.

2.1.1. Detection boundary in homoscedastic Gaussian mixtures. Suppose one observes $X_1, \cdots, X_n$ and wishes to test global hypotheses

$$\begin{aligned} &H_0^n : X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(0,1), \\ \text{v.s.} \quad &H_1^n : X_1, \ldots, X_n \overset{i.i.d.}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_n, 1). \end{aligned} \tag{4}$$

Interesting cases correspond to choices of (ϵ_n, μ_n) that are calibrated with a pair of parameters (β, r):

$$\epsilon_n = n^{-\beta}, \qquad \mu_n = \sqrt{2r \log n}, \qquad 1/2 < \beta < 1, \qquad 0 < r < 1.$$

There are two main goals in the analysis.

1. Determine the *detection boundary*, which gives the smallest possible signal strength r as a function of the sparsity parameter β such that reliable detection is possible.
2. Construct *adaptive* optimal tests, which simultaneously achieve vanishing probability of error for all values of (r, β) inside the detectable region.

Under model (4), Ingster (1998) and Donoho and Jin (2004) showed that there exists a detection boundary

$$r^*(\beta) = \begin{cases} \beta - \frac{1}{2}, & 1/2 < \beta \le 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1, \end{cases} \tag{5}$$

which separates the testing problem into two regions: the *detectable region* and the *undetectable region* (Figure 1). When (β, r) belongs to the interior of the undetectable region, the sum of Type I and Type II errors for testing the global null must tend to 1 and no test can asymptotically distinguish the two hypotheses (4). However when (β, r) belongs to the interior of the detectable region, there are tests for which both Type I and Type II errors tend to zero. In applications such as the identification of skilled fund managers, it is desirable to precisely select the fund managers who have true stock-picking skills. The goal is more ambitious and can only be achieved in a subset of the detection region when $r > \beta$ (*classifiable region*, Cai and Sun 2016). Inside the classifiable region, observations can be separately into null cases and non-null cases with negligible classification errors.

2.1.2. Methodologies for sparse detection. In the very sparse situation, most tests based on empirical moments have no power in detection. To construct adaptive optimal procedures, Ingster (1999) considered generalized likelihood ratio (GLR) tests over a growing discretized set of (β, r)-pairs and established its asymptotic adaptive optimality. A more elegant solution is provided by Donoho and Jin (2004), who proposed a testing procedure based on Tukey's Higher Criticism statistic and showed that it attains the optimal detection boundary (5).

The Higher Criticism test consists of three simple steps. First, for each $1 \le i \le n$, obtain a p-value by $p_i = \bar{\Phi}(Y_i) \equiv P\{N(0,1) \ge Y_i\}$, where $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0,1)$. Second, sort the p-values in the ascending order $p_{(1)} < p_{(2)} < \ldots < p_{(n)}$. Last,

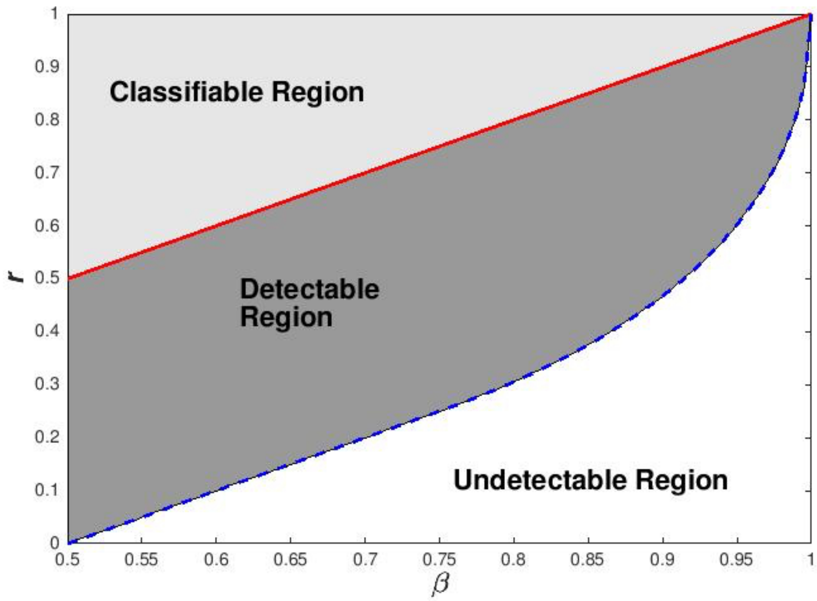

Figure 1

The detection boundary (dashed line) divides the β-r plane into the undetectable and detectable regions. It provides an optimality benchmark for the global testing problem (4). The higher criticism procedure attains the boundary and is hence *fully efficient*. Cai et al. (2007) showed that ϵ_n can be estimated consistently in the entire detectable region. The *classification boundary* (solid line; Cai et al. 2007; Cai and Sun 2016) gives the precise condition under which the observations can be separated into signals and noises with negligible misclassification rate.

define the Higher Criticism statistic as

$$HC_n^* = \max_{\{1 \leq i \leq n\}} HC_{n,i}, \qquad \text{where} \qquad HC_{n,i} = \sqrt{n}\left[\frac{i/n - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}\right], \tag{6}$$

and reject the null hypothesis H_0 when HC_n^* is large. The key ideas can be illustrated as follows. When $Y \sim N(0, I_n)$, $p_i \overset{iid}{\sim} U(0,1)$ and so $HC_{n,i} \approx N(0,1)$. Therefore, by the well-known results from empirical processes (e.g. Shorack and Wellner 2009), $HC_n^* \approx \sqrt{2 \log \log n}$, which grows to ∞ very slowly. In contrast, if $Y \sim N(\mu, I_n)$ where some of the coordinates of μ is nonzero, then $HC_{n,i}$ has an elevated mean for some i, and HC_n^* could grow to ∞ algebraically fast. Consequently, Higher Criticism is able to separate two hypotheses even in the very sparse case. Unlike the GLR test, the HC test is *optimally adaptive* in the sense that it attains the detection boundary without requiring the knowledge of the unknown parameters (β, r).

The above results have been generalized along various directions. Jager and Wellner (2007) proposed a family of goodness-of-fit tests based on the Rényi divergences, including the higher criticism test as a special case. The detection boundary with correlated noise and known variance was established in Hall and Jin (2010), where a modified version of the higher criticism was shown to achieve the corresponding optimal boundary.

2.1.3. Signal Detection under General Mixture Models.

The homoscedastic Gaussian mixture (4) is highly restrictive and idealized. In many applications, the signal strength varies among the non-null cases, violating the assumption of constant μ_n under the alternative. A natural question is the following: What is the detection boundary if μ_n varies with a distribution $\mathbb{P}_n$? Cai et al. (2011) considered a heteroscedastic Gaussian mixture model, which can be viewed as taking the signal strength under the alternative to be $\mathbb{P}_n = N(A_n, \tau^2)$. Writing σ^2 for $1+\tau^2$, under such a model, the detection problem aims to test

$$\begin{aligned} & H_0^n : Y_i \overset{i.i.d.}{\sim} N(0,1) \\ \text{v.s.} \quad & H_1^n : Y_i \overset{i.i.d.}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(A_n, \sigma^2). \end{aligned} \tag{7}$$

Cai et al. (2011) discovered that the detection problem behaves very differently in two regimes: the *sparse regime* where $1/2 < \beta < 1$ and the *dense regime* where $0 < \beta \leq 1/2$. Furthermore, a double-sided version of the higher criticism test was shown to be optimally adaptive in the whole detectable region in both the sparse and dense regimes, in spite of the very different detection boundaries and heteroscedasticity effects in the two cases. Classical methods have treated the detections of sparse and dense signals separately. In real practice, however, the information of the signal sparsity is usually unknown, the adaptivity of the modified higher criticism test is thus a practically useful property.

Cai and Wu (2014) considered the problem of sparse mixture detection in a more general model (1) where the distributions are not necessarily Gaussian and the non-null effects are not necessarily a binary vector. They obtained an explicit formula for the fundamental limit of the general testing problem under mild conditions on the mixture, which are in particular satisfied by the Gaussian and generalized Gaussian null distributions. These general results recover and extend all previously mentioned detection boundary results in a unified manner. The optimal adaptivity of the higher criticism procedure is also generalized far beyond the setup in Ingster (1999), Donoho and Jin (2004) and Cai et al. (2011). In the most general case, it turns out that detection boundary is determined by the asymptotic behavior of the log-likelihood ratio $\log \frac{dF_0}{dF_1}$ evaluated at an appropriate quantile of the null distribution.

2.2. Estimation of the Proportion of Non-null Effects

The proportion of non-null effects is an important quantity that is of significant interest in its own right. For example, in financial markets investors are interested in knowing the proportion of fund managers who possess true stock-picking skills. It is also one of the key quantities in the implementation of many large-scale multiple testing procedures. See, for example, Efron et al. (2001); Sun and Cai (2007); Storey (2007). The development of useful estimates of ϵ_n along with the corresponding statistical analysis is a challenging task. Recent work includes that of Langaas et al. (2005); Meinshausen and Rice (2006); Cai et al. (2007); Jin and Cai (2007) and Cai and Jin (2010).

2.2.1. Tail-based approach.

Schweder and Spjøtvoll (1982) proposed an intuitive method for estimating the proportion of null hypotheses using p-value plots. The methodology is developed for the general mixture model (1). To illustrate how it works, we simulated $n = 1000$ observations from a simple two-point normal mixture $F(x) = (1-\epsilon_n)N(0,1) + \epsilon_n N(2,1)$. The proportion of non-null hypotheses is $\epsilon_n = 0.2$. The histogram of the p-values is shown in panel (a) of Figure 2. Under the sparsity assumption, the majority of large p-values should come from the null distribution. Let λ be a sufficiently large threshold, say

$\lambda = 0.5$. Denote $W(\lambda) = \#\{i : p_i > \lambda\}$. Since the p-values to the right of the threshold roughly follow a uniform distribution, the expected counts covered by light grey bars can be approximated as $\mathbb{E}\{W(\lambda)\} \approx n(1 - \epsilon_n)(1 - \lambda)$. Setting the expected and actual counts equal, we obtain an estimate

$$\hat{\epsilon}_n(\lambda) = 1 - \frac{W(\lambda)}{n(1 - \lambda)}. \tag{8}$$

The p-value plotting method proposed in Schweder and Spjøtvoll (1982) is described in Panel (b) of Figure 2. The grey curve plots $1 - p_i$ against their rank. Then a straight line is fitted through the left portion of the grey curve and extended all the way to the right. The interception point gives the estimated proportion of null cases. In Benjamini and Hochberg (2000), this graphical method was formalized as an asymptotically equivalent step-wise least-slope estimator. See also Benjamini et al. (2006).

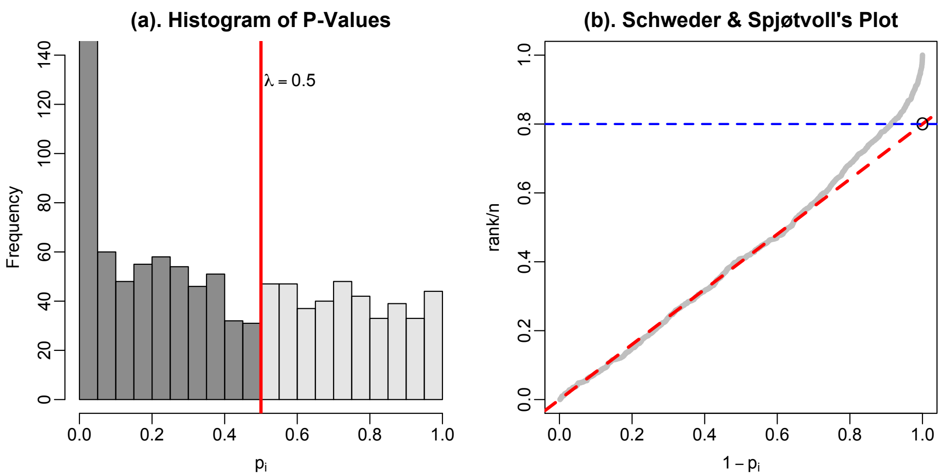

Figure 2

Tail-based methods for estimating ϵ_n. Data are simulated from a two-point normal mixture model $0.8 \cdot N(0, 1) + 0.2 \cdot N(2, 1)$. Panel (a) illustrates equation (8) with $\lambda = 0.5$. The p-values from right part of the histogram, represented by light grey bars, follow a uniform distribution approximately. Panel (b) illustrates the graphical solution in Schweder and Spjøtvoll (1982). The straight line was fitted through the p-values in the left via an "eyeball" method. The intersection point (○) shows that the estimated proportion of null cases is 0.8.

Langaas et al. (2005) showed that the estimate given by (8) always has a downward bias, i.e. $\mathbb{E}\{\hat{\epsilon}_n(\lambda)\} \leq \epsilon_n(\lambda)$ for all λ. There is a tradeoff in the choice of λ: a larger λ would reduce the bias but increase the variance. To choose a proper λ, Storey (2002) and Storey and Tibshirani (2003) proposed a bootstrapping method and a spline-smoothing method, respectively. In Langaas et al. (2005), the choice of λ is investigated systematically, and a class of estimators based on nonparametric MLEs were developed.

However, tail-based methods are in general biased; they are only consistent in a limited class of models satisfying the so-called "purity" condition (i.e. the non-null density has thinner tails than that of a standard normal). Moreover, the data tail is not scale invariant and consequently the accuracy of tail based methods depends on the degree of heteroscedasticity of the data.

2.2.2. Frequency-domain approach. Jin and Cai (2007) demonstrated that information on the null distribution and non-null proportion is well-preserved in the frequency domain instead of the spatial domain. They further proposed a frequency-domain approach to estimating the proportion. The estimator is robust against heteroscedasticity and is shown to be consistent for a wide class of parameter spaces. Numerical results demonstrate that it outperforms competing tail-based methods.

Consider the Gaussian mixture model

$$X_i \overset{iid}{\sim} (1-\epsilon_n)N(\mu_0, \sigma_0^2) + \epsilon_n Q_n, 1 \le i \le n, \tag{9}$$

where $N(\mu_0, \sigma_0^2)$ is the null distribution with possibly unknown parameters μ_0 and σ_0^2, and Q_n is a general Gaussian location-scale mixture with the density $q(x) = \int \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})dH_n(\mu,\sigma)$ for some mixing distribution H_n. We only discuss the case with known null parameters. See Jin and Cai (2007) for a modified procedure for the case with unknown null parameters. Then we can re-normalize X_j and assume, without loss of generality, $\mu_0 = 0$ and $\sigma_0 = 1$. The marginal density f of X_j becomes

$$f(x) = (1-\epsilon)\phi(x) + \epsilon \int \phi(\frac{x-\mu}{\sigma})dH_n(\mu,\sigma). \tag{10}$$

Jin and Cai's method can be described as follows. Introduce the empirical characteristic function $\varphi_n(t) = \frac{1}{n}\sum_{j=1}^n e^{itX_j}$, and its expectation, the characteristic function $\varphi(t) = \frac{1}{n}\sum_{j=1}^n e^{it\mu_j - \frac{\sigma_j^2 t^2}{2}}$, where $i = \sqrt{-1}$. Let $\omega(\xi)$ be a bounded, continuous and symmetric density function supported in $(-1,1)$. Define the phase function $\psi_n(t;\omega) = \int \omega(\xi)e^{\frac{t^2\xi^2}{2}}\varphi_n(t\xi)d\xi$. Fix $\gamma \in (0, 1/2)$ and let $t_n(\gamma) = \inf\{t : t > 0,\ |\varphi(t)| \le n^{-\gamma}\}$, the estimator is defined as

$$\hat{\epsilon}_n(\gamma;\omega) = 1 - \mathrm{Re}\left\{\psi_n(t_n(\gamma);\omega)\right\}, \tag{11}$$

where $\mathrm{Re}(z)$ stands for the real part of z. In Jin and Cai (2007) and Jin (2008), three different choices of $\omega(\xi)$ are recommended, namely the uniform density, the triangle density, and the smooth density that is proportional to $\exp(-\frac{1}{1-|\xi|^2}) \cdot 1_{\{|\xi|<1\}}$.

2.2.3. Optimality theory. The detection theory developed in Ingster (1999) and Donoho and Jin (2004) provides a benchmark for a theory of consistent estimation. However, the theoretical analysis for estimation of the proportion contains further challenges that are not present in the detection problem. For example, the procedure in Meinshausen and Rice (2006) is only capable of estimating ϵ_n consistently on a subset of the detectable region, failing to achieve the optimality benchmark of the detection boundary. Cai et al. (2007) developed an effective data-driven method for a two-point homoscedastic Gaussian mixture model $X_i \overset{iid}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_n, 1), 1 \le i \le n$ and showed that the estimator is rate-optimal within a logarithmic factor. In contrast to the results in Meinshausen and Rice (2006), the results in Cai et al. (2007) imply that it is possible to estimate ϵ_n consistently over the entire detectable region.

The optimality theory for estimating π_n was further developed in Cai and Jin (2010) for the general Gaussian mixture model (9). Cai and Jin (2010) introduced a modified estimator,

$$\hat{\epsilon}_n(\gamma) = \left(1 - \frac{1}{n}\sum_{j=1}^n e^{\frac{t^2}{2}}\cos(tX_j)\right)\Bigg|_{t=\sqrt{2\gamma\log n}} = 1 - n^{-(1-\gamma)}\sum_{j=1}^n \cos(\sqrt{2\gamma\log n}X_j). \tag{12}$$

The estimator $\hat{\epsilon}_n(\gamma)$ given in (12) can be viewed as a special case of $\hat{\epsilon}_n(\gamma;\omega)$, where instead of being a density function as in (11), ω is a point mass concentrated at 1. Cai and Jin (2010) obtained the convergence rate of the proposed estimator $\hat{\epsilon}_n(\gamma)$ and established a matching lower bound for the minimax rate. The results show that the estimator $\hat{\epsilon}_n(\gamma)$ given in (12) adaptively attains the optimal rate of convergence for a large collection of parameter spaces.

2.3. Estimation of the Null Distribution

Conventionally F_0 is assumed to be known and referred to as the *theoretical null*. It was argued by Efron (2004) that in large-scale inference problems, the use of theoretical null is incorrect and the choice of the null distribution has a huge impact on subsequent analysis. Efron further proposed the concept of *empirical null* and argued that the empirical evidence in the data determines the normal state and the null distribution should be estimated from the data. For the AYP example in Section 1.1, the empirical null is estimated to be $N(1.89, 1.81^2)$, which is substantially different from the theoretical null $N(0,1)$. This deviation can be attributed to unobserved covariates, unknown correlations or a large proportion of uninterestingly small effects.

Efron (2004) proposed a simple method to estimate the null parameters utilizing the central peak of the histogram. Jin and Cai (2007) proposed a class of more powerful estimators based on the empirical characteristic function and Fourier analysis. They further show that the proposed estimators are uniformly consistent over a wide class of parameters. Optimality theory was developed in Cai and Jin (2010). The empirical null approach in Efron (2004) and the estimation methods in Jin and Cai (2007) assume that all null cases follow a common distribution $N(\mu_0, \sigma_0^2)$. However, in applications such as the AYP study, a common null distribution does not exist. This issue was considered in Sun and McLain (2012), where Jin and Cai's method is extended to estimate the *composite null* distribution with an external covariate.

3. MULTIPLE TESTING PROBLEMS

Multiple testing is a useful approach to extract valuable insights from massive data. Its recent developments, epitomized by false discovery rate methodologies, have greatly influenced a wide range of scientific and business disciplines. This section reviews some important concepts and recent progresses of this field.

3.1. Multiplicity, Error Rate and Power Concepts

When performing a hypothesis test, two types of errors may be committed: rejecting a hypothesis when it is null (type I error), or failing to reject a hypothesis when it is non-null (type II error). A Type I error means finding a pattern that does not exist in the data (false discovery), whereas a Type II error indicates missing out an interesting pattern that actually exists (missed discovery). In practice, one cannot entirely eliminate the chance of committing decision errors. However, the consequences of the two types of errors are usually different, with a type I error being regarded as a more serious mistake. Define Type I and II error rates as the probability of making the respective type of error. The classical formulation in single hypothesis testing aims to control the type I error rate at a prespecfied level α while minimizing the Type II error rate.

When n hypotheses are tested simultaneously, the outcomes of all tests can be summarized in Table 1. In the multiple testing setting, it is desirable to assess the overall performance of a testing procedure by combining all decisions together. The multiplicity, which leads to inflation of Type I errors, becomes a serious issue. Next we discuss some widely used concepts for measuring the *overall error rate* in multiple testing.

Table 1 Classification of tested hypotheses

	Claimed non-significant	Claimed significant	Total
Null	N_{00}	N_{10}	n_0
Non-null	N_{01}	N_{11}	n_1
Total	S	R	n

3.1.1. Family-Wise Error Rate (FWER). The FWER is defined as the probability of making at least one Type I error in the family, e.g. $\text{FWER} = \mathbb{P}(N_{10} \geq 1)$, where N_{10} is the number of false positive findings. It has been widely used as an overall error measure when multiple hypotheses are tested at the same time. A per-comparison error rate (PCER) procedure, which repeatedly tests each hypothesis at level α, fails to control the FWER. The most well-known FWER procedure is the Bonferroni correction, which conducts individual tests at level α/m instead of α. Bonferroni method can be further improved by step-wise methods such as Holm's procedure and Hommel's procedure (Holm 1979; Hommel 1988; Hochberg 1988), or resampling based methods (Westfall and Young 1993). We refer interested readers to Shaffer (1995) and Hochberg and Tamhane (2009) for an extensive review of FWER methodologies. A useful extension of the FWER is the k-FWER, which is defined as the probability of making k or more Type I errors in the family. The k-FWER controlling procedures are more powerful than FWER methods; recent works include Lehmann and Romano (2005a), Romano and Shaikh (2006) and Sarkar (2007).

3.1.2. False Discovery Rate (FDR) . The FWER is a very strict criterion. When thousands and even millions of hypotheses are tested simultaneously, the FWER procedures often become excessively conservative and fail to identify most useful signals. This often results in the waste of expensive studies and possible financial losses. In large-scale settings, a more powerful and practical error rate concept is the false discovery rate (FDR, Benjamini and Hochberg 1995). Under the FDR paradigm, one is willing to tolerate some Type I errors, provided that the number is small relative to the total number of rejections. Define the false discovery proportion

$$\text{FDP} = \begin{cases} N_{10}/R, & \text{if } R > 0 \\ 0, & \text{if } R = 0 \end{cases}. \tag{13}$$

Then the FDR is the expectation of the FDP

$$\text{FDR} = \mathbb{E}(\text{FDP}) = \mathbb{E}\left(\frac{N_{10}}{R}\,\middle|\, R > 0\right)\mathbb{P}(R > 0). \tag{14}$$

The FDR concept reflects the tradeoff between false discoveries and true discoveries in practice, and is connected to minimax estimation theory (Abramovich et al. 2006) and compound decision theory (Sun and Cai 2007). Other closely related measures include the positive false discovery rate (pFDR, Storey 2003) and the marginal false discovery rate (mFDR, Genovese and Wasserman 2002). The difference among various FDR measures seem to be non-essential in large-scale testing problems. For example, the pFDR and mFDR are equivalent when test statistics come from a random mixture model (Storey 2003). Genovese and Wasserman (2002) showed that, under mild conditions, mFDR = FDR + $O(m^{-1/2})$.

The FDR is fundamentally different from the FWER by providing a powerful and cost-effective framework to handle large-scale testing problems. Although the subject of FDR is still relatively new, it has already exhibited enormous impacts on many scientific and business fields. This article reviews its important recent developments.

3.1.3. Power and Optimality. In single hypothesis testing, the power is defined as the probability of correctly rejecting a non-null hypothesis. The fundamental Neyman-Pearson lemma shows that the likelihood ratio test is the *most powerful test* in the sense that it maximizes the power at a pre-specified test level α.

The power concept can be generalized in different ways as we move to multiple testing. We shall use the expected number of true positives

$$\text{ETP} = \mathbb{E}(N_{11}) \tag{15}$$

in this article. Other related measures include the average power (Spjøtvoll 1972; Storey 2007; Efron 2007b), the false negative/non-discovery rate (FNR, Genovese and Wasserman 2002; Sarkar 2004):

$$\text{FNR} = \mathbb{E}\left(\frac{N_{01}}{S} \,\middle|\, S > 0 \right) \mathbb{P}(S > 0),$$

the missed discovery rate (MDR, Taylor et al. 2005) and the non-discovery rate (NDR, Haupt et al. 2011). Under mild conditions (Cao et al. 2013), maximizing the ETP is asymptotically equivalent to minimizing the FNR or MDR. An FDR procedure is said to be *valid* if it controls the FDR at the nominal level α, and *optimal* if it has the largest ETP among all valid FDR procedures at level α.

3.2. P-Value Based Methodologies for FDR Control

In single hypothesis testing, p-value is a fundamental statistic: we decide whether a hypothesis should be rejected by comparing the p-value with the test level α. A widely used strategy in multiple testing is to first rank the hypotheses according to individual p-values and then choose a cutoff along the ranking. This section reviews p-value based FDR methodologies; their limitations and optimal FDR control will be discussed in Section 3.3.

3.2.1. Benjamini-Hochberg's (BH) procedure. Let $\{p_i : 1 \le i \le n\}$ be the p-values from individual tests. Denote $p_{(1)} \le p_{(2)} \le \cdots \le p_{(n)}$ the ordered p-values and $H_{(1)}, \cdots, H_{(n)}$ the corresponding hypotheses. The BH procedure first uses a step-up comparison to decide a p-value threshold:

$$\text{Let } k = \max\{i : p_{(i)} \le i\alpha/n\}, \tag{16}$$

then rejects all hypotheses $H_{(j)}, j = 1, \cdots, k$. This method can be intuitively explained as follows. Suppose the cutoff is $p_{(i)}$ and i hypotheses are rejected. Because the null p-values follow a uniform distribution, one expects to have $n_0 p_{(i)}$ significant p-values from the null and the FDP can be estimated by $\hat{Q}_j = n_0 p_{(i)}/i$. In practice, n_0 is not known but can be approximated by n. The corresponding estimated FDP is then $\tilde{Q}_j = np_{(i)}/i$. To maximize the power, we choose the largest i such that $\tilde{Q}_i \le \alpha$, which leads directly to the BH procedure (16).

The BH procedure is easy to implement and has a simple graphical representation. To illustrate, we simulate $n = 60$ observations from a random mixture model $(1 - \epsilon_n)N(0, 1) + \epsilon_n N(2.5, 1)$ with $\epsilon_n = 0.25$. In Figure 3, the discrete points are ranked p-values plotted against their indices. The straight lines correspond to the right hand side of equation (16), where the slope is the prespecified FDR level α. The p-value threshold is given by the last crossing point between the p-value curve and the straight line.

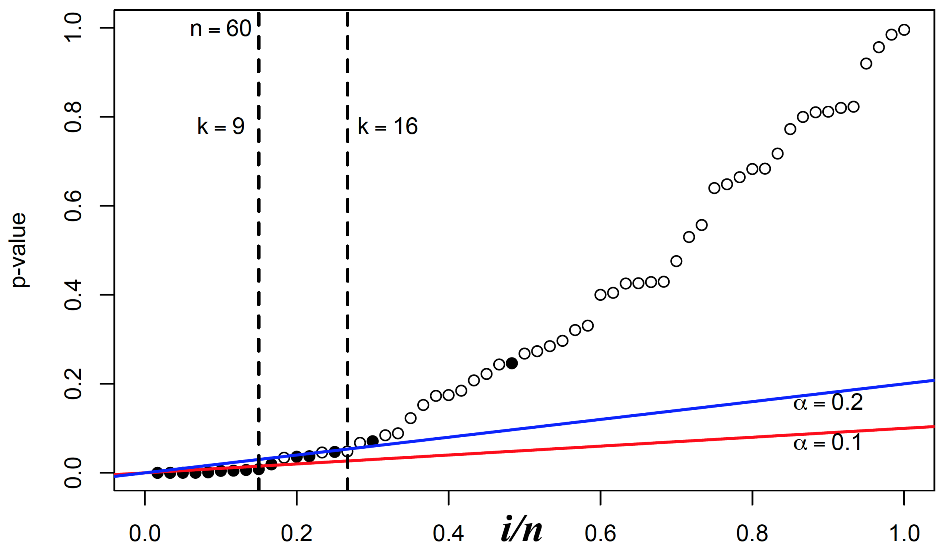

Figure 3

An graphical illustration of the BH procedure: ● and ○ stand for non-null and null cases, respectively. The FDR thresholds are computed as the largest intersection point of the p-value curve and straight line, whose slope corresponds to the test level. At $\alpha = 0.1$, 9 hypotheses are rejected with no false positive. At $\alpha = 0.2$, 16 hypotheses are rejected with 3 false positives.

Benjamini and Hochberg (1995) showed that Procedure (16) controls the FDR at the nominal level when the p-values are independent. The BH procedure remains valid for FDR control under positive regression dependency and weak dependency (Benjamini and Yekutieli 2001; Storey et al. 2004). The BH threshold is usually larger than the FWER threshold, leading to a more powerful procedure with more rejections. The power gain over FWER methods becomes more pronounced as the number of tests increases. This makes the method more suitable for large-scale simultaneous inference.

3.2.2. Adaptive p-value procedure. The BH procedure is conservative because it controls the FDR at level $(1 - \epsilon_n)\alpha$ instead of α, where ϵ_n is the proportion of non-null cases. Benjamini and Hochberg (2000), Genovese and Wasserman (2002), and Storey (2002) proposed to estimate ϵ_n from data and further utilize it to construct more powerful procedures.

Let $\hat{\epsilon}_n$ be an estimate of ϵ_n. Then the *adaptive p-value procedure* (Benjamini and

Hochberg 2000) operates as follows.

$$\text{Let } k = \max\{i : P_{(i)} \le i\alpha/[(1-\hat{\epsilon}_n)n]\}, \text{ then rejects all } H_{(i)}, i \le k. \tag{17}$$

We can see that in (17), the BH procedure is carried out at an adjusted FDR level $\alpha/(1-\hat{\epsilon}_n)$. Therefore by incorporating the estimated proportion, the procedure is *adaptive* to the sparsity information in the data. Numerical results show that the power of BH method can be improved, and the efficiency gain increases with ϵ_n.

3.2.3. Oracle and plug-in p-value procedures. Let $G_1(t)$ be the cumulative distribution function (CDF) of the p-value of a non-null case and $G(t)$ is the mixture CDF. Consider a random mixture model for p-values:

$$G(t) = (1-\epsilon_n)t + \epsilon_n G_1(t). \tag{18}$$

The marginal FDR (mFDR) for a given cutoff t (e.g. we reject H_i if $p_i < t$) is defined as

$$Q(t) = \frac{\mathbb{E}(N_{10})}{\mathbb{E}(R)} = \frac{(1-\epsilon_n)t}{G(t)}.$$

If G_1 is concave, then the solution to $Q(t) = \alpha$, denoted by u^*, is unique. The *oracle p-value procedure* reject H_i if $p_i < u^*$. It is *optimal* in the sense that it has the smallest FNR among all p-value based procedures at mFDR level α (Genovese and Wasserman 2002). However, this optimality result only holds within the class of p-value based methods.

When G and ϵ_n are unknown, we use their estimates $\hat{G}$ and $\hat{\epsilon}_n$ to obtain the estimated FDR level $\hat{Q}(t) = (1-\hat{\epsilon}_n)t/\hat{G}(t)$. The estimation of $\hat{\epsilon}_n$ has been discussed in Section 2.2. G is commonly estimated by the empirical CDF $\hat{G}(t) = m^{-1}\sum_{i=1}^m \mathbb{I}\{p_i < t\}$, where $\mathbb{I}(\cdot)$ is an indicator function. Hence a class of *plug-in* FDR procedures can be constructed (Genovese and Wasserman 2002, 2004) as follows.

$$\text{Let } t(\hat{p}, \hat{G}) = \sup\{t : \hat{Q}(t) \le \alpha\}. \quad \text{Reject } H_i \text{ if } p_i < t(\hat{p}, \hat{G}). \tag{19}$$

Equation (19) reveals the connection between a multiple testing problem and an *FDR estimation* problem. The BH procedure and adaptive p-value procedure can be identified as special cases in the class. For example, if we choose $\hat{\epsilon}_n = 0$ and $\hat{G}(t)$ as the empirical CDF, then (19) reduces to the well-known BH procedure. Genovese and Wasserman (2004) developed a stochastic process framework for multiple testing and showed that, when consistent estimates of G and p are chosen, the class of plug-in procedures (19) are *asymptotically valid* and exhaustive. That is, the FDR is controlled at level $\alpha + o(1)$.

3.2.4. The q-value procedure. The p-value has a nice interpretation and provides a convenient framework for testing a single hypothesis, e.g. we reject the null if the p-value is less than α. The q-value (Storey 2003) can be viewed as an analogue of the p-value in the FDR paradigm in the sense that if we want to carry out an FDR analysis at level α, then we can obtain the q-value for each test and reject H_i if its q-value is less than α. The q-value has gained great popularity in large-scale "omics" research such as genomics and proteomics (Tusher et al. 2001) due to its convenience and nice interpretation.

Roughly speaking, the q-value of a test measures the fraction of false discoveries when that test is *just* rejected. Consider the random mixture model (18), the positive FDR (pFDR) is defined as $\mathrm{pFDR}(t) = \mathbb{E}\left(\frac{N_{10}}{R}\,\middle|\, R > 0\right) = (1-\epsilon_n)t/G(t)$, where t is the p-value cutoff. The q-value of H_i is the smallest FDR level such that H_i can be rejected:

$$q(p_i) = \inf_{t \geq p_i} \{\mathrm{pFDR}(t)\} = \inf_{t \geq p_i} \left\{ \frac{(1-\epsilon_n)t}{G(t)} \right\}. \tag{20}$$

In practice, we estimate ϵ_n and G as $\hat{\epsilon}_n$ and $\hat{G}$. Suppose all hypotheses are arranged in ascending order of p-values $p_{(1)}, \cdots, p_{(m)}$. Then the q-value procedure works as follows.

$$\text{Let } \hat{q}\left(p_{(i)}\right) = \frac{(1-\hat{\epsilon}_n)p_{(i)}}{\hat{G}\left(p_{(i)}\right)}. \text{ Reject } H_{(i)} \text{ if } \hat{q}\left(p_{(i)}\right) \leq \alpha. \tag{21}$$

The q-value is computed for an individual case but has a global interpretation: it reflects the relative significance of a single test by taking into account of the p-values from all other tests. By comparing (21) with (19), we can see that the q-value procedure belongs to the class of plug-in methods.

3.2.5. Other error rate concepts and methodologies. In situations where the FDP is highly variable, the false discovery exceedance (FDX, Genovese and Wasserman 2004) provides a useful alternative to the FDR. Let $0 \leq \tau \leq 1$ be a pre-specified *tolerance level*, the FDX at level τ is $\mathrm{FDX}_\tau = \mathbb{P}(\mathrm{FDP} > \tau)$, the tail probability that the FDP exceeds a given bound. The goal is to construct a testing procedure satisfying $\mathrm{FDX} \leq \alpha$. The FDX control takes into account the variability of the FDP, and is desirable with correlated tests where variability of FDP is very high. See Lehmann and Romano (2005b), Genovese and Wasserman (2006), and Roquain and Villers (2011) for recent development in FDX theories and methodologies.

Other important p-value based FDR procedures include the augmentation procedure (van der Laan et al. 2004), two-stage linear procedure (Benjamini et al. 2006), and resampling procedures (Tusher et al. 2001), among others. The resampling methods are attractive in many applications because the p-values and adjusted p-values can be estimated without making any parametric assumptions on the joint distribution of the test statistics. Moreover, the correlation structure and distributional characteristics of the data can be preserved. Algorithms for computing adjusted p-values are introduced, for example, in Westfall and Young (1993) and Dudoit et al. (2003).

There are a range of other error measures in the multiple testing literature, including the FWER, k-FWER, FDR, generalized FDR, marginal FDR, positive FDR, FDX, false cluster rate, weighted FDR, overall FDR, outer-node FDR, and focus-level FDR. These concepts are useful but may cause confusion. Benjamini (2010) provided a good summary of error measures and discussed how to match proper error rates with inference needs.

3.3. Optimal FDR Control: A Decision-Theoretic Approach

In multiple testing, we aim to separate the non-null cases from null cases. A testing procedure can be represented by a binary rule $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n) \in \{0,1\}^n$, where $\delta_i = 0/1$ indicates that we claim that case i is null/non-null. Multiple testing is a *compound decision problem* (Robbins 1951) since all tests are combined and evaluated together.

The development of a multiple testing procedure involves two steps: (i) deriving a test statistic T_i that ranks hypotheses from the most significant to the least significant, and (ii) setting a cutoff t for T_i to control the FDR at α. This leads to a thresholding rule:

$$\delta_i = \mathbb{I}(T_i < t), i = 1, \cdots, n. \tag{22}$$

We can see that T_i, which determines the ranking of hypotheses, plays a central role in multiple testing. In conventional FDR procedures, the default choice for T_i has been the p-value. Sun and Cai (2007) developed a compound decision theoretic framework and showed that the p-value is not a fundamental building block in large-scale testing problems. The next sections survey results on *optimal* and *asymptotically optimal* FDR procedures and show that all p-value methods can be uniformly improved.

3.3.1. Oracle FDR procedure. Consider an ideal setup where an oracle knows p, f_0 and f_1. To develop the oracle rule, we consider two problems in turn: (i) what is *oracle statistic* that gives the optimal ranking of all tests? (ii) What is the *oracle cutoff* that controls the FDR and maximizes the ETP?

Consider model (1). Suppose we obtain a z-value from each test. Sun and Cai (2007) showed that the optimal test statistic in the oracle setting is the *local false discovery rate*

$$\mathrm{Lfdr}(z_i) = \frac{(1-\epsilon_n)f_0(z_i)}{f(z_i)}. \tag{23}$$

Now consider a class of FDR procedures of the form $\delta_i(t) = \mathbb{I}\{\mathrm{Lfdr}(z_i) < t\}, \text{ for } 1 \le i \le n$, where $0 \le t \le 1$ is a cutoff. The next step is to find the oracle cutoff that controls the FDR at level α with the largest ETP (15). To this end, denote $Q_{\mathrm{OR}}(t)$ the FDR level when the cutoff for Lfdr is t. Define the oracle cutoff as the largest cutoff allowed under the FDR constraint $t_{\mathrm{OR}} = \sup\{t : Q_{\mathrm{OR}}(t) \le \alpha\}$. Finally, we introduce the *oracle FDR procedure* as a thresholding rule based on Lfdr and t_{OR}: $\boldsymbol{\delta}_{\mathrm{OR}} = (\delta^i_{\mathrm{OR}} : 1 \le i \le n)$, where

$$\delta^i_{\mathrm{OR}} = \mathbb{I}\{\mathrm{Lfdr}(z_i) < t_{\mathrm{OR}}\}. \tag{24}$$

Sun and Cai (2007) showed that the oracle rule (24) is optimal for FDR control in the sense that it has the largest ETP among all FDR procedures at level α.

The Lfdr statistic has a Bayesian interpretation: $\mathrm{Lfdr}(z_i) = \mathbb{P}(\text{case } i \text{ is null} \mid z_i)$ (Efron et al. 2001). It captures all important distributional information in the mixture model (1). The expression (23) implies that we actually rank the hypotheses according to the ratio f_0/f, and the ranking is more efficient than that based on p-values. An interesting consequence of using the Lfdr statistic is that we may accept a more "extreme" observation while rejecting a less extreme observation, which implies that the rejection region is asymmetric. This point will be illustrated in Section 3.3.3 using the mutual funds data.

3.3.2. A data-driven procedure. The oracle procedure cannot be implemented in practice since both the Lfdr and t_{OR} are unknown. We discuss how to estimate the unknown quantities. Let $\hat{\epsilon}_n$, $\hat{f}_0$ and $\hat{f}$ be estimates of ϵ_n, f_0 and f, respectively. The estimation of ϵ_n is discussed in Section 2. The null density f_0 is either taken as a known theoretical null, i.e. the standard normal density, or is estimated as an *empirical null* using methods in Efron (2004) and Jin and Cai (2007). The mixture density f can be obtained as a standard kernel density estimator with bandwidth chosen by cross validation (Silverman 1986). Then the Lfdr statistic can be estimated as

$$\widehat{\mathrm{Lfdr}}_i = \frac{(1-\hat{\epsilon}_n)\hat{f}_0(z_i)}{\hat{f}(z_i)}.$$

Next, we derive a *data-driven* procedure that mimics the oracle procedure. We use the "ranking followed by thresholding" idea to motivate a step-wise method. Denote $\widehat{\mathrm{Lfdr}}_{(1)} \le$

$\ldots \le \widehat{\text{Lfdr}}_{(n)}$ the ordered Lfdr statistics. Suppose j hypotheses are rejected along the ranking, then the actual FDR level can be estimated as $\hat{Q}_{OR}(j) = \frac{1}{j}\sum_{i=1}^{j} \widehat{\text{Lfdr}}_{(i)}$, the moving average of the top j ordered statistics [cf. Sun and Cai (2007)]. To fulfill the FDR constraint and maximize the power, we propose the following step-wise procedure:

$$\text{Let } k = \max\left\{ j : \frac{1}{j}\sum_{i=1}^{j} \widehat{\text{Lfdr}}_{(i)} \le \alpha \right\}, \text{ then reject all } H^{(i)},\ i = 1, \cdots, k. \tag{25}$$

The goals of global FDR control and individual case interpretation are naturally unified in the data-driven procedure (25). Moreover, with the consistent estimators proposed in Jin and Cai (2007), Sun and Cai (2007) showed that the data-driven procedure is *asymptotically valid and optimal* in the sense that the data-driven procedure controls the FDR at level $\alpha + o(1)$, and has an FNR level of $\text{FNR}_{\text{OR}} + o(1)$, where FNR_{OR} is the FNR level of the oracle procedure.

3.3.3. Analysis of mutual funds data: a comparison of p-value and Lfdr. Consider a normal mixture model with three components:

$$(1 - \epsilon_n^- - \epsilon_n^+)N(0,1) + \epsilon_n^- N(\mu^-, 1) + \epsilon_n^- N(\mu^+, 1),$$

where ϵ_n^- and ϵ_n^+ are the proportions of negative and positive non-null cases, respectively. The model was considered in Barras et al. (2010) for analysis of mutual funds data, where $N(0,1)$, $N(\mu^-,1)$, $N(\mu^+,1)$ are used to describe the distributions of zero alpha funds, unskilled funds and skilled funds, respectively. We choose a setting so that the main findings in Barras et al. (2010) can be roughly matched. Specifically, $n = 5000$ z-values are simulated from the mixture model with $\mu^- = -2.5$, $\mu^+ = 3$, $\epsilon_n^- = 0.15$ and $\epsilon_n^+ = 0.05$. Hence many funds have underperformance but few have outperformance. The histograms of zero, positive and negative components are plotted in different colors in Figure 4, with a mixture density curve fitted to the observed bars.

In practice we do not know the true states of nature but only observe a mixture of the three types of funds. It is desirable to identify both skilled and unskilled funds. We apply the BH procedure (Benjamini and Hochberg 1995), adaptive p-value (AP, Benjamini and Hochberg 2000) procedure and the data-driven Lfdr procedure (Sun and Cai 2007) to the data set at $\alpha = 0.1$. The results are summarized in Table 2.

Table 2 Analysis summary for simulated mutual funds data.

Methods	# Rejections	# True Rejections	FDP	Lower cutoff	Upper cutoff
BH	572	532	0.07	−2.53	2.53
AP	633	579	0.085	−2.41	2.41
Lfdr	694	626	0.098	−2.18	2.73

We can see that the Lfdr procedure controls the false discovery proportion (FDP) more precisely compared to the p-value based methods. Moreover, it correctly identifies more non-zero alpha funds compared to the p-value based methods. The efficiency gain is due to the *adaptivity* of the Lfdr procedure. Concretely, the mixture is an asymmetric distribution with ϵ_n^- being higher than ϵ_n^+, hence we are more likely to find signals in the negative component. Therefore it makes sense to adopt an *asymmetric rejection region* when selecting nonzero

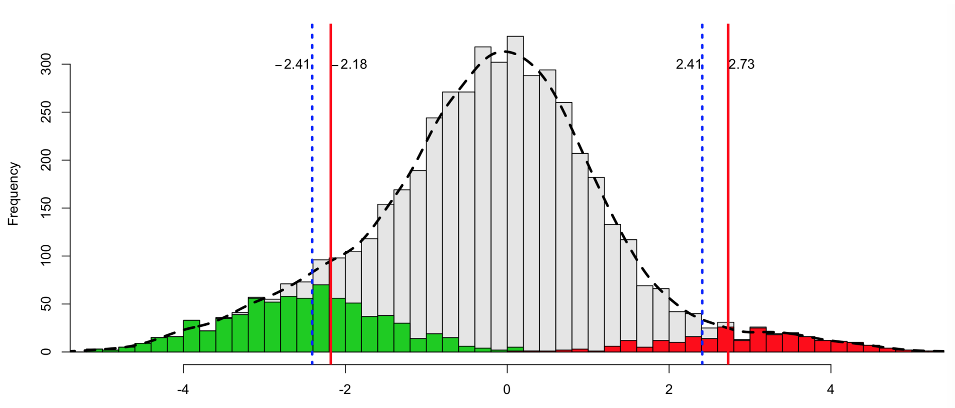

Figure 4

Mutual funds example: symmetric vs. asymmetric rejection regions. The normal mixture model is $0.8 \cdot N(0,1) + 0.15 \cdot N(-2.5,1) + 0.05 \cdot N(3,1)$ with a higher proportion of negative alpha funds. It makes sense to adopt an *asymmetric rejection region* as we are more likely to find signals in the negative part. The Lfdr procedure allows to accept an observation located further away from 0 while rejecting an observation closer to 0. In contrast, p-value based methods are not adaptive to the asymmetry of the distribution. The rejection region of the Lfdr method is given by $z < -2.18$ or $z > 2.73$. In contrast, the rejection region of the AP method is $|z| > 2.41$.

alpha funds. The Lfdr procedure is adaptive in the sense that it produces asymmetric regions automatically (without having to estimate ϵ_n^- and ϵ_n^+!). We can see from Figure 4 that the rejection region of the AP method is $|z_i| > 2.41$, whereas the rejection region of the Lfdr procedure is $z_i < -2.18$ and $z_i > 2.73$. It is interesting to note that the Lfdr procedure rejects observation $z = -2.2$ but does not reject observation $z = 2.6$. This will never be encountered by a p-value method which always has symmetric rejection regions.

3.4. Multiple Testing with External and Structural Information

Conventional multiple testing procedures implicitly assume that data are collected from repeated or identical experimental conditions, and hence hypotheses are exchangeable. However, in many applications, data are known to be collected from heterogeneous sources and form into groups. Moreover, relevant domain knowledge, such as external covariates, scientific insights, prior data and hierarchical structure, is often available alongside the primary data set in many studies. Exploiting such information in an efficient manner promises to enhance both the interpretability of research results and precision of statistical inference.

3.4.1. Heterogeneity and grouping. The problem of multiple testing with groups and related problems are studied in Efron (2008); Ferkingstad et al. (2008); Cai and Sun (2009); Hu et al. (2012), among others. For example, in the AYP study discussed in Section 1.1, the estimated null densities of the z-values for large schools is much wider than those in medium and small schools. In the brain imaging study considered by Schwartzman et al. (2008), the null cases for the front and back halves of the brain centered on different means, and the density of the back half is narrower. The differences in the null distributions have significant impacts on the outcomes of multiple testing procedures.

Efron (2008) introduced the *multi-group mixture model* to handle the heterogeneity in the data. Suppose $X_1, \cdots, X_n$ can be divided into K groups:

$$X_{ki} \sim f_k = (1 - \pi_{1k}) f_{k0} + \pi_{1k} f_{k1},\ i = 1, \cdots, n_k,\ k = 1, \cdots, K. \tag{26}$$

The group memberships are assumed to be known. Three strategies for testing grouped hypotheses have been considered in the literature. First, the *pooled analysis* simply ignores the information of group labels and conducts a global analysis on the combined sample at a given FDR level α. It is argued by Efron (2008) that a pooled FDR analysis is problematic because highly significant cases from one group may be hidden among the nulls from another group, while insignificant cases may be possibly enhanced. Efron (2008) suggested the second approach, namely the *separate analysis*, which first conducts an FDR analysis at level α within each group, and then combines the testing results from all groups. It was shown by Efron (2008) that the separate analysis controls the FDR. However, the choice of identical FDR levels across all groups can be suboptimal. Cai and Sun (2009) showed that both the separate and pooled analyses can be uniformly improved by a third approach, the conditional Lfdr (CLfdr) method, which enjoys features from both pooled and separate analyses. Let $\hat{p}_k$, $\hat{f}_{k0}$ and $\hat{f}_k$ be estimates of the unknown quantities in (26). Then the CLfdr procedure operates as follows:

1. Calculate the plug-in CLfdr statistic $\widehat{\text{CLfdr}}_{ki} = (1 - \hat{p}_k)\hat{f}_{k0}(x_{ki})/\hat{f}_k(x_{ki})$.
2. Combine and rank the plug-in CLfdr values from all groups. Denote by $\widehat{\text{CLfdr}}_{(1)}, \cdots,$ $\widehat{\text{CLfdr}}_{(n)}$ the ranked values and $H_{(1)}, \cdots, H_{(n)}$ the corresponding hypotheses.
3. Reject all $H_{(i)}$, $i = 1, \cdots, l$, where $l = \max\left\{i : (1/i)\sum_{j=1}^{i} \widehat{\text{CLfdr}}_{(j)} \le \alpha\right\}$.

It is important to note that in step 1, the external information of group labels is utilized to calculate the CLfdr; this is the feature from a separate analysis. However, in steps 2 and 3, the group labels are dropped and the rankings of all hypotheses are determined globally; this is the feature from a pooled analysis. Cai and Sun (2009) showed that the CLfdr procedure is *asymptotically valid and optimal.* Unlike for the separate analysis, the group-wise FDR levels of the CLfdr procedure, which are in general different from α, are adaptively weighted among groups.

3.4.2. External weights. In multiple testing, the hypotheses being investigated often become "unequal" in light of external information, which may be reflected by differential attitudes towards the relative importance of testing units or the severity of decision errors. The use of weights provides an effective strategy to incorporate informative domain knowledge in large-scale testing problems. In the literature, various weighting methods have been advocated for a range of multiple comparison problems (Genovese et al. 2006; Roeder and Wasserman 2009; Roquain and Van De Wiel 2009). A popular scheme, referred to as the *decision weights* approach, involves modifying the error criteria or power functions (Benjamini and Hochberg 1997). The idea is to employ two sets of positive constants $\boldsymbol{a} = \{a_i : i = 1, \cdots, n\}$ and $\boldsymbol{b} = \{b_i : i = 1, \cdots, n\}$ to take into account the costs and gains of multiple decisions. Let δ_i be the decision for H_i. The weighted false discovery rate (wFDR) is defined as

$$\text{wFDR} = \mathbb{E}\left\{\sum_{i=1}^{n} a_i(1 - \theta_i)\delta_i\right\} / \mathbb{E}\left(\sum_{i=1}^{n} a_i\delta_i\right),$$

where a_i is the weight indicating the severity of a false positive decision. For example, a_i is taken as the cluster size in the spatial cluster analyses conducted in Benjamini and Heller

(2007) and Sun et al. (2015). As a result, rejecting a larger cluster erroneously corresponds to a more severe decision error. To compare the effectiveness of different weighted multiple testing procedures, we define the expected number of true positives ETP $= \mathbb{E}\left(\sum_{i=1}^{n} b_i\theta_i\delta_i\right)$, where b_i is the weight indicating the power gain when H_i is rejected correctly. The use of b_i provides a useful scheme to incorporate informative domain knowledge. In spatial data analysis, correctly identifying a larger cluster that contains signal may correspond to a larger b_i, indicating a greater decision gain. By combining the concerns on both the error criterion and power function, the goal in weighted multiple testing is to

$$\text{maximize the ETP subject to the constraint wFDR} \leq \alpha. \tag{27}$$

Basu et al. (2015) developed an asymptotically optimal solution to (27). The key step involves a conceptualization of the constrained optimization problem (27) as an expanding knapsack problem, followed by an application of the classical ideas in Neyman-Pearson Lemma. This leads to a fast greedy algorithm that substantially speeds up conventional knapsack algorithms with optimality guarantees. Moreover, the optimality theory reveals that the optimal ranking depends on the pre-specified wFDR level, an interesting phenomenon unknown in previous works.

3.4.3. Hierarchical structure and logical correlation. In many applications, the data are aggregated to different resolution levels and it is desirable to test hypotheses in a hierarchical fashion. Hierarchical analysis is also useful in large-scale pattern recognition problems. When the signals are sparse, it is desirable to first separate signals from massive and noisy data (testing) and then determine the patterns of the selected signals (classification). The task can be described as finding needles of various shapes in a haystack. Important applications include hierarchical testing in oncological genetics, fault detection and classification in control engineering, and satellite surveillance for coarse to fine interpretation of visual images. The pattern discovery process can be described by a decision tree with multiple levels, where decisions are made at finer and finer resolution levels going from the top to bottom of the tree. At each node of a given level, we have three possible actions: (i) testing: deciding whether a unit contains one of the patterns of interest; (ii) classification: assigning the selected subjects to a specific pattern categories (classification); and (iii) indecision: selecting a subject as a signal but does not specify its pattern.

In hierarchical testing, important error measures for summarizing the whole decision process include full-tree and outer-node FDR's (Yekutieli 2008), the focus level FDR (Goeman and Mansmann 2008), the mixed directional FDR (Guo et al. 2010), and the overall false discovery rate (Sun and Wei 2015). Moreover, a hierarchical decision rule needs to fulfill a genuine logical relationship, that is, a case is rejected only if its parent node is rejected. Various methods have been developed for the adjustment of statistical significance according to the hierarchical structure, as well as the logical and error rate constraints; see Blanchard and Geman (2005), Goeman and Mansmann (2008), Yekutieli (2008), Meinshausen (2008), Goeman and Solari (2010) and Sun and Wei (2015). Recent works on multiple comparison issue in multi-stage and sequential testing problems include Benjamini et al. (2006), Lin (2006), Dmitrienko et al. (2007), Benjamini and Heller (2007), Posch et al. (2009), Liang and Nettleton (2010), Sarkar et al. (2013), Benjamini and Bogomolov (2014), and Cai and Sun (2016). Hierarchical testing is also related to the control of *directional errors* in multiple testing; see Guo et al. (2010), and Goeman et al. (2010) for related theories and methodologies.

3.5. Multiple Testing Under Dependency

Observations arising from large scale testing problems are often dependent. However, conventional FDR procedures rely heavily on the independence assumption, and the correlation among hypotheses is typically ignored. There are two important questions regarding the dependence issue: (i) what is the impact of dependence on the conventional FDR analysis? (ii) How to construct new FDR procedures for dependent tests?

3.5.1. Impact of dependence in multiple testing. The impact of dependence has been extensively studied in the multiple testing literature. The results can be roughly divided into two types. First, it has been shown that the classical BH procedure is valid for controlling the FDR under different dependency assumptions, indicating that it is safe to apply conventional methods as if the tests were independent (see Benjamini and Yekutieli 2001; Sarkar 2002; Storey et al. 2004; Wu 2008; Clarke and Hall 2009, among others). On the other hand, Efron (2007a) and Schwartzman and Lin (2011) showed that correlation usually degrades statistical accuracy, affecting both estimation and testing. High correlation also results in high variability of testing results and hence the irreproducibility of scientific findings; see Owen (2005); Finner et al. (2007) for related discussions. These results suggest that dependency has negative impact and must be adjusted for multiple testing, especially when the correlations are very high. Leek and Storey (2008) and Friguet et al. (2009) studied multiple testing under the factor models and showed that by subtracting the common factors out, the dependence structure can be greatly weakened. Efron (2007a) and Fan et al. (2012) discussed how to take into account the dependence structure and obtain more accurate FDR estimates for a given p-value threshold. However, these p-value based methods still suffer from efficiency loss when the dependence structure is highly informative.

3.5.2. Exploiting dependence for multiple testing. Some empirical studies Some empirical studies have demonstrated that dependence can be utilized to improve the precision of inference. The idea is to aggregate weak signals from individuals and pool information from nearby observations by exploiting high correlations. Genovese et al. (2006) and Benjamini and Heller (2007), Sun and Cai (2009), and Sun and Wei (2011) showed that incorporating functional, spatial and temporal correlations into a multiple testing procedure may greatly improve the power and accuracy of conventional methods.

To see why the dependence structure can be helpful, consider the following example. Suppose one observes a mixture of null and non-null hypotheses and expects that *the non-null cases appear in clusters*. Suppose the observed sequence is

$$\cdots, -2.8, -3.4, x_1, -3.2, -2.9, \cdots, 0.2, -0.3, x_2, 0.01, 1, \cdots,$$

where $x_1 = x_2 = 2$. Heuristically we can argue that x_1 is likely to come from the non-null distribution because there is evidence in the sample that it is in a cluster with negative effects. In contrast, x_2 is likely to be a random large observation that comes from a cluster of null effects. Therefore it is natural to assign different significance levels to x_1 and x_2 even if the observed values are the same. However, x_1 and x_2 have the same p-values if inspected alone. Next we discuss how to systematically incorporate the structural information among the hypotheses in multiple testing. We first consider a simple and widely used model and then move to more complicated settings.

3.5.3. Hidden Markov models. Hidden Markov model (HMM) is a widely used and effective tool for modeling the dependency structure (Rabiner 1989). Suppose we observe a mixture of null and non-null hypotheses and expect that the non-nulls appear in clusters. In an HMM, the sequence of the unknown (hidden) null and non-null states is assumed to form a Markov chain $(\theta_i)_1^n = (\theta_1, \cdots, \theta_n) \in \{0, 1\}^n$. The observed data values $\boldsymbol{x} = (x_1, \cdots, x_n)$ are independent conditional on the hidden states $(\theta_i)_1^n$. Let ϑ denote the collection of all HMM parameters.

Sun and Cai (2009) showed that under the HMM dependency, the optimal test statistic is the *local index of significance* $\mathrm{LIS}_i = \mathbb{P}_\vartheta(\theta_i = 0|\boldsymbol{x})$, which can be computed using a fast forward-backward algorithm. The LIS is superior than the p-value as it utilizes the HMM dependence to pool information from nearly observations. The information from the whole sequence is integrated to calculate the LIS statistic. By using LIS, the signal to noise ratio is increased and the procedure is more robust against local disturbance.

In practice, we estimate the HMM parameters by $\hat{\vartheta}$ and use a plug-in statistic $\widehat{\mathrm{LIS}}_i = \mathbb{P}_{\hat{\vartheta}}(\theta_i = 0|\boldsymbol{x})$. The maximum likelihood estimate is commonly used and is strongly consistent and asymptotically normal (Leroux 1992; Bickel et al. 1998). The MLE can be computed using the EM algorithm or other standard optimization schemes. Denote by $\widehat{\mathrm{LIS}}_{(1)}, \cdots, \widehat{\mathrm{LIS}}_{(n)}$ the ranked plug-in test statistics and $H_{(1)}, \cdots, H_{(n)}$ the corresponding hypotheses. The following data-driven procedure can be used for FDR control:

$$\text{Let } k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} \widehat{\mathrm{LIS}}_{(j)} \le \alpha \right\}, \text{ then reject all } H_{(i)},\ i = 1, \cdots, k. \tag{28}$$

Sun and Cai (2009) showed that the data-driven procedure controls the FDR at level $\alpha + o(1)$, and is asymptotically optimal. Numerical results from both simulated and real data show that conventional p-value based methods can be greatly improved. At the same FDR level, the number of false positives is greatly reduced and the statistical power to reject a non-null is substantially increased. This indicates that dependence can make the testing problem easier and can be a *blessing* if incorporated properly.

3.5.4. Random field model: Point-wise inference. The multiple comparison issue has been raised in a wide range of spatial analyses such as brain imaging (Genovese et al. 2002; Heller et al. 2006; Schwartzman et al. 2008), disease mapping and surveillance (Green and Richardson 2002), and network analysis (Wei and Li 2007). When the intensities of signals have a spatial pattern, it is expected that incorporating the underlying dependence structure can significantly improve the power and accuracy of conventional methods. We discuss how to extend the methodology in an HMM to spatial settings.

Let S be a spatial domain. Consider the random field model (RFM) $\boldsymbol{X} = \{X(s) : s \in S\}$ in Pacifico et al. (2004) for spatial multiple testing: $X(s) = \mu(s) + \epsilon(s)$, where $\mu(s)$ is the unobserved random process and $\epsilon(s)$ is the noise process. Assume that there is an underlying state $\theta(s)$ associated with each location s with one state being dominant ("background"). In applications, an important goal is to identify locations that exhibit significant deviations from background. This can be formulated as a multiple testing problem. Let $\theta(s) \in \{0, 1\}$ be an indicator such that $\theta(s) = 1$ if location s contains signal and $\theta(s) = 0$ otherwise. For each location we make a decision $\delta(s) = 1$ if the null is rejected and $\delta(s) = 0$ otherwise. The *decision process* for the whole spatial domain S is denoted by $\boldsymbol{\delta} = \{\delta(s) : s \in S\}$. Let $\nu(\cdot)$ denote the Lebesgue/counting measure for a continuous/discrete domain. The spatial

FDR can be defined as

$$\text{FDR} = \mathbb{E}\left(\frac{\nu(S_{FP})}{\nu(R)}\middle|\, \nu(R) > 0\right)\mathbb{P}(\nu(R) > 0)$$

where $R = \{s \in S : \delta(s) = 1\}$ is the rejection area, and $S_{\text{FP}} = \{s \in S : \theta(s) = 0, \delta(s) = 1\}$ is the false positive area.

Let $\boldsymbol{x}^N = (x_1, \cdots, x_N)$ denote the observed values. Suppose an oracle knows all RFM parameters, denoted by Ψ. The oracle statistic for point-wise inference is $T_{OR}(s) = \mathbb{P}_\Psi\{\theta(s) = 0|\boldsymbol{x}^N\}$. However, this requires testing an uncountable number of hypotheses for all $s \in S$, which is impossible in practice. Sun et al. (2015) showed that a continuous decision process can be described, within a small margin of error, by a finite number of decisions on a grid of pixels. Concretely, the strategy is to divide a continuous S into n "pixels," pick one point in each pixel, and use the decision at that point to represent all decisions in the pixel. Let $\cup_{i=1}^n S_i$ be a partition of S. Pick a point s_i from each S_i. Let $T_{OR}^{(1)} \le T_{OR}^{(2)} \le \cdots \le T_{OR}^{(n)}$ denote the ordered oracle statistics and $S_{(i)}$ the corresponding regions. In a point-wise inference, define $R_j = \cup_{i=1}^j S_{(i)}$ and $r = \max\left\{j : \nu(R_j)^{-1}\sum_{i=1}^j T_{OR}^{(i)}\nu(S_{(i)}) \le \alpha\right\}$. The rejection area is given by $R = \cup_{i=1}^r S_{(i)}$. This procedure can be implemented efficiently under a Bayesian computational framework, which involves hierarchical modeling and MCMC computing. See Sun et al. (2015) for detailed algorithms.

3.5.5. Cluster-wise/set-wise inference. When the interest is on the behavior of a process over sub-regions, the testing units become spatial clusters instead of individual locations. Combining simultaneous tests in sets or clusters can improve statistical power and provide new research insights (Benjamini and Heller 2008; Sun and Wei 2011).

Let $\mathcal{C} = \{C_1, \cdots, C_K\}$ denote the set of (known) clusters of interest. In many applications it is desirable to incorporate the cluster size or other spatial variables in the error measure. Let ϑ_k be a binary variable which equals 0/1 if cluster k is null/non-null and 0 otherwise. The decision for cluster k is denoted a binary indicator Δ_k, where $\Delta_k = 1$ if cluster k is claimed to be significant and $\Delta_k = 0$ otherwise. We use the *false cluster rate* (FCR) to measure the overall error rate of a cluster-wise procedure:

$$\text{FCR} = \mathbb{E}\left\{\frac{\sum_k w_k(1-\vartheta_k)\Delta_k}{(\sum_k w_k\Delta_k) \vee 1}\right\}, \tag{29}$$

where w_k are cluster specific weights which are often pre-specified in practice. For example, one can take $w_k = \nu(C_k)$, the size of a cluster, to indicate that a false positive cluster with larger size would account for a larger error.

Let $C_1, \cdots, C_K$ be the clusters and $\mathcal{H}_1, \cdots, \mathcal{H}_K$ the corresponding hypotheses. The oracle statistic for cluster-wise inference is $T_{OR}(C_k) = P_\Psi(\vartheta_k = 0|\boldsymbol{x}^N)$. Let $T_{(1)}^c \le \cdots \le T_{(K)}^c$ be the ordered $T_{OR}(C_k)$ values, and $\mathcal{H}_{(1)}, \cdots, \mathcal{H}_{(K)}$ and $w_{(1)}, \cdots, w_{(K)}$ the corresponding hypotheses and weights, respectively. Let $r = \max\left\{j : \{\sum_{k=1}^j w_{(k)}\}^{-1}\sum_{k=1}^j w_{(k)}T_{(k)}^c \le \alpha\right\}$. Then reject $\mathcal{H}_{(1)}, \cdots, \mathcal{H}_{(r)}$. This procedure controls the FCR at level α, and can be implemented by MCMC algorithms. See Sun et al. (2015) for details.

3.5.6. Arbitrary dependence. Our discussions have focused on situations where dependency structures can be well estimated from data. The problem of FDR control under arbitrary and unknown dependence still requires further research. Benjamini and Yekutieli (2001) showed that performing the BH procedure at level $\alpha/(\sum_{i=1}^{n} 1/i)$ always control the FDR at level α under arbitrary dependence. However, such an adjustment is too conservative and often unnecessary in practice. It remains an open issue on how to estimate the unknown dependence and utilize the information to construct more powerful tests.

4. DISCUSSION AND OTHER TOPICS

Statistical inference for high-dimensional covariance structures is an active and important area of research. Driven by a wide range of applications, there have been significant recent developments on the methods and theory for testing of the global covariance structures and simultaneous testing of a large number of hypotheses on the local covariance structures with FDP and FDR control. High dimensionality and dependency impose significant challenges in the construction and analysis of the testing procedures. The present paper does not cover this important topic. We refer interested readers to Cai (2016) for a comprehensive review on global testing for the covariance, correlation, and precision matrices, and multiple testing for the correlations, Gaussian graphical models, and differential networks.

Another topic that is not discussed in this paper is simultaneous inference for high-dimensional regression models, which has received much recent attention. See, for example, Lockhart et al. (2014), Zhang and Zhang (2014), Javanmard and Montanari (2014), Van de Geer et al. (2014), Liu and Luo (2014), Barber et al. (2015), Xia et al. (2015), and Cai and Guo (2016).

Multiple testing is often used as a selection or screening step in the overall analysis. *Selective inference*, which involves making further inference on the selected variables, is an important area that requires much research on formal theoretical principles and practical methodologies. Making valid inference after multiple testing or model selection is a challenging task because the estimates of the post-selection variables would be biased if the selection effects are not taken into account. Post-selection inference techniques are useful in classical statistical problems such as the estimation of many normal means and simultaneous confidence intervals (Benjamini and Yekutieli 2005; Brown and Greenshtein 2009; Efron 2011), as well as rapidly growing areas such as high-dimensional regression and sparse principal components analysis; see Yekutieli (2012), Hwang and Zhao (2013), Berk et al. (2013), Benjamini and Bogomolov (2014), Taylor and Tibshirani (2015) and Lee et al. (2016) for recent developments in this direction.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The research of Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334, and the research of Wenguang Sun was supported in part by NSF grant DMS-CAREER 1255406.

LITERATURE CITED

Abramovich, F. and Benjamini, Y. and Donoho, D.L. and Johnstone, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics* 34: 584–653.

Barber, R. F., E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* 43: 2055–2085.

Barras, L., O. Scaillet, and R. Wermers (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance* 65:179–216.

Basu, P., T. T. Cai, K. Das, and W. Sun (2015). Weighted false discovery rate control in large-scale multiple testing. *Technical Report. arXiv:1508.01605.*

Benjamini, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal* 52:708–721.

Benjamini, Y. and M. Bogomolov (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76: 297–318.

Benjamini, Y. and R. Heller (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* 102: 1272–1281.

Benjamini, Y. and R. Heller (2008). Screening for partial conjunction hypotheses. *Biometrics* 64: 1215–1222.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* 57: 289–300.

Benjamini, Y. and Y. Hochberg (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* 24: 407–418.

Benjamini, Y. and Y. Hochberg (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 25: 60–83.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93: 491–507.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* 29: 1165–1188.

Benjamini, Y. and D. Yekutieli (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* 100: 71–81.

Berk, R., L. Brown, A. Buja, K. Zhang, L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics* 41: 802–837.

Bickel, P. J., Y. Ritov, T. Ryden (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics* 26: 1614–1635.

Blanchard, G. and D. Geman (2005). Hierarchical testing designs for pattern recognition. *Ann. Statist.* 33: 1155–1202.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 37: 1685–1704.

Cai, T. T. (2016). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annual Review of Statistics and Its Application* (to appear).

Cai, T. T. and Z. Guo (2016). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics to appear*, arXiv:1506.05539.

Cai, T. T., X. J. Jeng, and J. Jin (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73: 629–662.

Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38:100–145.

Cai, T. T., J. Jin, and M. G. Low (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* 35: 2421–2449.

Cai, T. T. and W. Sun (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* 104: 1467–1481.

Cai, T. T. and W. Sun (2016). Optimal screening and discovery of sparse signals with applications to multistage high-throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology). To Appear*.

Cai, T. T. and Y. Wu (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory* 60: 2217–2232.

Cao, H., W. Sun, and M. R. Kosorok (2013). The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika* 100: 495–502.

Clarke, S. and P. Hall (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* 37: 332–358.

Dmitrienko, A., B. L. Wiens, A. C. Tamhane, and X. Wang (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat. Med.* 26: 2465–2478.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* 32: 962–994.

Dudoit, S., J. P. Shaffer, and J. C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* 99: 96–104.

Efron, B. (2007a). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* 102: 93–103.

Efron, B. (2007b). Size, power and false discovery rates. *Ann. Statist.* 35: 1351–1377.

Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat.* 2: 197–223.

Efron, B. (2011). Tweedies formula and selection bias. *Journal of the American Statistical Association* 106: 1602–1614.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc. 96*, 1151–1160.

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

Ferkingstad, E., A. Frigessi, H. Rue, G. Thorleifsson, and A. Kong (2008). Unsupervised empirical bayesian multiple testing with external covariates. The Annals of Applied Statistics, 2:714–735.

Finner, H., T. Dickhaus, and M. Roters (2007). Dependency and false discovery rate: asymptotics. *Ann. Statist.* 35: 1432–1455.

Friguet, C., M. Kloareg, and D. Causeur (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* 104: 1406–1415.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B* 64: 499–517.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *Ann. Statist.* 32: 1035–1061.

Genovese, C. R., N. A. Lazar, and T. Nichols (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15: 870–878.

Genovese, C. R., K. Roeder, and L. Wasserman (2006). False discovery control with p-value weighting. *Biometrika* 93: 509–524.

Genovese, C. R. and L. Wasserman (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association* 101: 1408–1417.

Goeman, J. J. and U. Mansmann (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 24: 537–544.

Goeman, J. J. and A. Solari (2010). The sequential rejection principle of familywise error control. *Ann. Statist.* 38: 3782–3810.

Goeman, J. J., A. Solari, and T. Stijnen (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in medicine* 29: 2117–2125.

Green, P. J. and S. Richardson (2002). Hidden markov models and disease mapping. *Journal of the American statistical association* 97: 1055–1070.

Guo, W., S. K. Sarkar, and S. D. Peddada (2010). Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics* 66: 485–492.

Hall, P. and J. Jin (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* 38: 1686–1732.

Harvey, C. R. and Y. Liu (2015). Backtesting. *Journal of Portfolio Management* 42: 13–28.

Haupt, J., R. M. Castro, and R. Nowak (2011). Distilled Sensing: Adaptive Sampling for Sparse Detection and Estimation. *IEEE T. Inform. Theory* 57: 6222–6235.

Heller, R., D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini (2006). Cluster-based analysis of fmri data. *Neuroimage* 33: 599–608.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800–802.

Hochberg, Y. and A. C. Tamhane (2009). Multiple comparison procedures.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* 6: 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika* 75: 383–386.

Hu, J. X., H. Zhao, and H. H. Zhou (2012). False discovery rate control with groups. *Journal of the American Statistical Association* 105: 1215–1227.

Hwang, J. G. and Z. Zhao (2013). Empirical bayes confidence intervals for selected parameters in high-dimensional data. *Journal of the American Statistical Association* 108: 607–618.

Ingster, Y. I. (1998). Minimax detection of a signal for l^n-balls. *Math. Methods Statist.* 7: 401–428.

Jager, L. and J. A. Wellner (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics* 35: 2018–2053.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* 15: 2869–2909.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. Roy. Statist. Soc. B* 70: 461–493.

Jin, J. and T. T. Cai (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* 102: 495–506.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. Roy. Statist. Soc. B* 67: 555–572.

Lee, J. D., D. L. Sun, Y. Sun, J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44: 907–927.

Leek, J. T. and J. D. Storey (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* 105: 18718–18723.

Lehmann, E. and J. P. Romano (2005a). Generalizations of the familywise error rate. *The Annals of Statistics* 33: 1138–1154.

Lehmann, E. L. and J. P. Romano (2005b). *Testing statistical hypotheses* (Third ed.). Springer Texts in Statistics. New York: Springer.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* 40: 127–143.

Liang, K. and D. Nettleton (2010). A hidden markov model approach to testing multiple hypotheses on a tree-transformed gene ontology graph. *Journal of the American Statistical Association* 105: 1444–1454.

Lin, D. Y. (2006). Evaluating statistical significance in two-stage genomewide association studies. *Am. J. Hum. Genet.* 78: 505–509.

Liu, W. and S. Luo (2014). Hypothesis testing for high-dimensional regression models. Technical report.

Lo, A. W. and A. C. MacKinlay (1990). Data-snooping biases in tests of financial asset pricing models. *Review of financial studies* 3: 431–467.

Lockhart, R., J. Taylor, R. J. Tibshirani, and R. Tibshirani (2014). A significance test for the lasso. *Annals of statistics* 42: 413–468.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika* 95: 265–278.

Meinshausen, N. and J. Rice (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* 34: 373–393.

Newton, M. A., A. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5: 155–176.

Owen, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc., Ser. B* 67: 411–426.

Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *Journal of the American Statistical Association* 99: 1002–1014.

Posch, M., S. Zehetmayer, and P. Bauer (2009). Hunting for significance with the false discovery rate. *J. Amer. Statist. Assoc.* 104: 832–840.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77: 257–286.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, Berkeley and Los Angeles, pp. 131–148. University of California Press.

Roeder, K. and L. Wasserman (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics* 24: 398–413.

Rogosa, D. (2003). Accuracy of api index and school base report elements: 2003 academic performance index, California department of education.

Romano, J. P. and A. M. Shaikh (2006). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics* 34: 1850–1873.

Roquain, E. and M. A. Van De Wiel (2009). Optimal weighting for false discovery rate control. *Electronic journal of statistics* 3: 678–711.

Roquain, E. and F. Villers (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *The Annals of Statistics* 39: 584–612.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* 30: 239–257.

Sarkar, S. K. (2004). Fdr-controlling stepwise procedures and their false negatives rates. *J. Stat. Plan. Infer.* 125: 119–137.

Sarkar, S. K. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *Ann. Statist.* 35: 2405–2420.

Sarkar, S. K., J. Chen, and W. Guo (2013). Controlling the false discovery rate in two-stage combination tests for multiple endpoints. *J. Am. Statist. Assoc.* 108: 1385–1401.

Schwartzman, A., R. F. Dougherty, and J. E. Taylor (2008). False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Stat.* 2: 153–175.

Schwartzman, A. and X. Lin (2011). The effect of correlation in false discovery rate estimation. *Biometrika* 98: 199–214.

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* 69: 493–502.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology* 46: 561–584.

Shorack, G. R. and J. A. Wellner (2009). *Empirical processes with applications to statistics*, Volume 59. Siam.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press.

Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math.*

Statist. 43: 398–411.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. B* 64: 479–498.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* 31: 2013–2035.

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

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. B* 66: 187–205.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100: 9440–9445.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* 102: 901–912.

Sun, W. and T. T. Cai (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. B* 71: 393–424.

Sun, W. and A. McLain (2012). Multiple testing of composite null hypotheses in heteroscedastic models *J. Amer. Statist. Assoc.* 107: 673–687.

Sun, W., B. J. Reich, T. T. Cai, M. Guindani, and A. Schwartzman (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77: 59–83.

Sun, W. and Z. Wei (2011). Large-scale multiple testing for pattern identification, with applications to time-course microarray experiments. *J. Amer. Statist. Assoc.* 106: 73–88.

Sun, W. and Z. Wei (2015). Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika* 32: 1823–1831.

Taylor, J., R. Tibshirani, and B. Efron (2005). The miss ratefor the analysis of gene expression data. *Biostatistics* 6: 111–117.

Taylor, J. and R. J. Tibshirani (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences* 112: 7629–7634.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 98: 5116–5121.

Van de Geer, S., P. Bühlmann, Y. Ritov, R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42: 1166–1202.

van der Laan, M. J., S. Dudoit, and K. S. Pollard (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical applications in genetics and molecular biology* 3.

Wei, Z. and H. Li (2007). A markov random field model for network-based analysis of genomic data. *Bioinformatics* 23: 1537–1544.

Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279. John Wiley & Sons.

White, H. (2000). A reality check for data snooping. *Econometrica* 68: 1097–1126.

Wu, W. B. (2008). On false discovery control under dependence. *Ann. Statist.* 36: 364–380.

Xia, Y., T. Cai, and T. T. Cai (2015). Two-sample tests for high-dimensional linear regression with an application to detecting interactions. Technical report, Department of Statistics, University Pennsylvania.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *J. Amer. Statist. Assoc.* 103: 309–316.

Yekutieli, D. (2012). Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74: 515–541.

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