ANNUAL REVIEWS

Annual Review of Economics

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, Pennsylvania 19104; email: tcai@wharton.upenn.edu

[2]Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, California 90089; email: wenguans@marshall.usc.edu

Annu. Rev. Econ. 2017. 9:411–39

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-063016-104355

JEL codes: C12, C13, C44, C55, G11

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 larger data sets. Statistical inference for big data often requires solving thousands or even millions of parallel inference problems simultaneously. This poses significant challenges and calls for new principles, theories, and methodologies. This review provides 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 nonnull 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.

1. LARGE-SCALE INFERENCE

In current business and economic research, massive and complex data sets, with thousands or even millions of variables, are collected routinely by governments, organizations, small businesses, and large enterprises. This 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 progress that has been made recently in this field, with a focus on multiple testing with false discovery rate (FDR) 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.

The first example is detection of anomalous events in financial markets. An 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 fraud, cyber intrusion, financial market anomalies, and covert communication. For example, techniques for reliably detecting and precisely locating credit card fraud are important for credit card companies to improve their service and reduce possible financial losses. To predict and detect fraud, one must monitor an enormous amount of transactions from many customers simultaneously. The detection and modeling of extreme events in time series are rapidly growing areas in financial economics. The sequential change-point analysis has been widely used in detecting anomalies and instabilities (Lumsdaine & Papell 1997, Andreou & Ghysels 2006, Fryzlewicz 2014). One outstanding challenge is identifying anomalies in the financial markets as quickly as possible while controlling the false alarm rate. These large-scale inference problems involve either producing massive amounts of real-time estimates or testing thousands or even millions of hypotheses with high frequencies.

The second example is selection of skilled fund managers. In financial markets, monthly returns from a large number of mutual funds are routinely collected. As a guide to evaluating past and future performance, 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 by the fact that thousands of mutual funds exist in the financial markets. The selection of skilled fund managers requires some formal principles to control false discoveries.

The third example is evaluation of trading rules. An important goal in financial economics is to test a large number of factors to explain cross-sectional patterns and use these to develop or evaluate new trading strategies. However, the simultaneous investigation of a large number of factors gives rise to the issue of selection bias or data-snooping bias (Lo & MacKinlay 1990, Harvey & 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 thus 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.

The fourth example is model selection in macroeconomic forecasting. Forecasting with high-dimensional time series data is an important problem in macroeconomics, where one often needs to, for example, evaluate the effects of monetary policies or predict GDP growth and Consumer Price Index inflation based on a large set of macroeconomic variables (Stock & Watson 2012). Standard techniques such as the vector autoregressive model and factor analysis must be modified with penalized variable selection techniques to avoid overfitting. In a recent study, Chudik et al. (2016) propose a new variable selection principle that involves the evaluation of the net contribution of each covariate in explaining the response while taking into account the multiplicity in simultaneous tests. By adopting a multiple testing framework as the stopping rule, both the model interpretability and forecasting performance can be much improved.

The fifth example is comparison of academic performances. The adequate yearly progress (AYP) study of California high schools (Rogosa 2003) aims to compare academic performances of socioeconomically advantaged (SEA) versus socioeconomically disadvantaged (SED) students. In the AYP study, standard tests in mathematics were administered to 7,867 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 simultaneously. 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 (e.g., a p-value or a z-value). The cases are either null or nonnull, with nonnull 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 nonnull cases. There are many ways to model sparse data, but one of the most natural is to posit a mixture model

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

where the mixing proportion ϵ_n is small, F_0 is the null distribution, and F_1 is the nonnull 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 nonnull case. If the summary statistics (e.g., p-values) are comparable across units, then Equation 1 would be suitable under heterogeneity. It is commonly assumed that $F_0 = \text{Unif}(0,1)$ if X_i is a p-value and $F_0 = N(0,1)$ if X_i is a z-value. Let f_0 and f_1 denote the densities corresponding to null and nonnull cases, respectively. The marginal density is given by $f(x) = (1-\epsilon_n)f_0(x) + \epsilon_n f_1(x)$. The mixture model (Equation 1) provides a

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 & 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 whether there are 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,

$$H_0^n : \epsilon_n = 0 \ \text{ versus } \ H_1^n : \epsilon_n \neq 0. \qquad 2.$$

The proportion ϵ_n of nonnull effects is an important quantity. For instance, the magnitude of ϵ_n can help one 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 or funds to reduce the large gaps between test scores. An interesting and technically challenging global inference problem is obtaining a good estimate of the nonnull 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 might 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 nonnull. This gives rise to a multiple testing problem, which involves making simultaneous inferences on n hypotheses:

$$H_{i0}\text{: case } i \text{ is null versus } H_{i1}\text{: case } i \text{ is nonnull, } i = 1, \ldots, n. \qquad 3.$$

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 are 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. On the first day, he sends emails to 102,400 individuals and makes predictions on the stock market activity in the next day: Half are told that the market will go up and the other half that it will go down. On the second day, those who received the wrong predictions will be discarded from the email list, and the remainder will get emails with new predictions: again, half up and half

down. After 10 trading days, the 100 people who are still on the email list will have received 10 correct predictions in a row. Without knowing the scheme or accounting for the multiplicity, these 100 people must be 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 substantially overestimated.

This review provides 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 nonnull effects. Section 3 focuses on multiple testing with 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

Global inference problems include the testing and estimation of unknown parameters that capture the overall structure of all units. Under the mixture model (Equation 1), we study a class of interrelated global inference problems: (*a*) testing the global hypothesis (Equation 2), (*b*) estimating the nonnull proportion ϵ_n, and (*c*) 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 the early stage, when only a small fraction of firms or markets are adversely affected. Other examples include the detection of disease outbreaks, credit card fraud, 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 $X_1, \ldots, X_n$ and wishes to test global hypotheses

$$H_0^n : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1) \qquad 4.$$
$$\text{versus } H_1^n : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_n, 1).$$

We consider the following choices of (ϵ_n, μ_n) that are calibrated with a pair of parameters (β, r):

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

This β–r range means that the fraction of signals is small and the magnitude of the signals is only moderately large. This calibration leads to an interesting and subtle global testing problem (Donoho & Jin 2004, Meinshausen & Rice 2006, Cai et al. 2007).

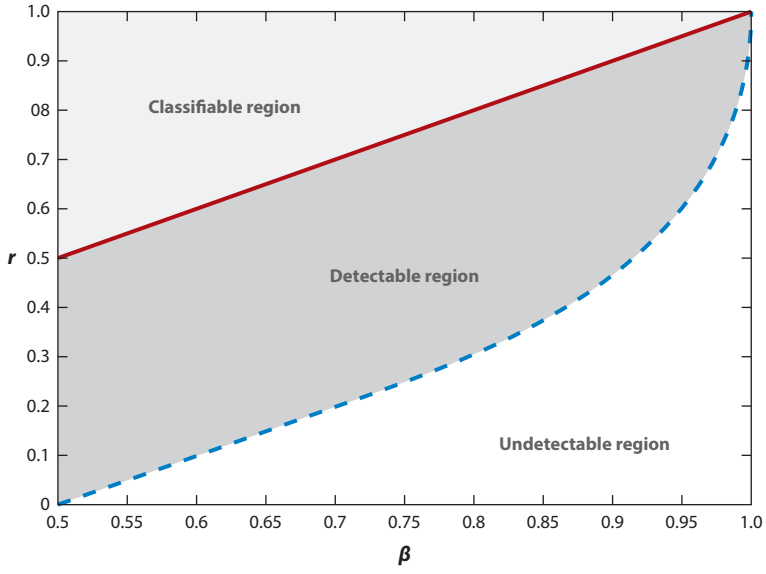

Figure 1

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 (Equation 4). The higher criticism procedure attains the boundary and is, thus, fully efficient. Cai et al. (2007) show that ϵ_n can be estimated consistently in the entire detectable region. The classification boundary (*solid line*) (Cai et al. 2007, Cai & Sun 2016) gives the precise condition under which the observations can be separated into signals and noises with a negligible misclassification rate.

There are two main goals of this analysis. The first is to 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. The second is to construct adaptive optimal tests, which simultaneously achieve vanishing probability of error for all values of (r, β) inside the detectable region.

Under Equation 4, Ingster (1998) and Donoho & Jin (2004) show that there exists a detection boundary

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

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 contained in Equation 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 where $r > \beta$ (the classifiable region; Cai & Sun 2016). Inside the classifiable region, observations can be separated into null cases and nonnull cases with negligible classification errors.

2.1.2. Methodologies for sparse detection. In the very sparse situation $3/4 < \beta < 1$, most tests based on empirical moments have no power in detection. To construct adaptive optimal procedures, Ingster (1998) considers generalized likelihood ratio (GLR) tests over a growing discretized set of (β, r) pairs and establishes their asymptotic adaptive optimality. A more elegant solution is provided by Donoho & Jin (2004), who propose a testing procedure based on Tukey's higher criticism (HC) statistic and show that it attains the optimal detection boundary (Equation 5).

The HC 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 $p_{(1)} < p_{(2)} < \cdots < p_{(n)}$. Last, define the HC statistic as

$$\mathrm{HC}_n^* = \max_{\{1 \le i \le n\}} \mathrm{HC}_{n,i}, \qquad 6.$$

where $\mathrm{HC}_{n,i} = \sqrt{n}\left[\frac{i/n - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}\right]$, and reject the null hypothesis H_0 when HC_n^* is large. The key ideas can be illustrated as follows. When $Y \sim N(0, I_n)$ holds true, we have $p_i \overset{\text{i.i.d.}}{\sim} U(0,1)$, and so $\mathrm{HC}_{n,i} \approx N(0,1)$. Therefore, by the well-known results from empirical processes (e.g., Shorack & Wellner 2009), it follows that $\mathrm{HC}_n^* \approx \sqrt{2 \log\log n}$, which grows to ∞ very slowly. In contrast, if $Y \sim N(\mu, I_n)$, where some of the coordinates of μ are nonzero, then $HC_{n,i}$ has an elevated mean for some i, and HC_n^* could grow to ∞ algebraically fast. Consequently, the HC test 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 knowledge of the unknown parameters (β, r).

The above results have been generalized along various directions. Jager & Wellner (2007) propose a family of goodness-of-fit tests based on the Rényi divergences, including the HC test as a special case. The detection boundary with correlated noise and known variance is established by Hall & Jin (2010), who show that a modified version of the HC test achieves the corresponding optimal boundary.

2.1.3. Signal detection under general mixture models. The homoscedastic Gaussian mixture (Equation 4) is highly restrictive and idealized. In many applications, the signal strength varies among the nonnull 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) consider a heteroscedastic Gaussian mixture model, which can be viewed as taking the signal strength under the alternative to be $\mathbb{P}_n = N(A_n, \tau^2)$. If σ^2 is written for $1 + \tau^2$, then, under such a model, the detection problem aims to test

$$H_0^n : Y_i \overset{\text{i.i.d.}}{\sim} N(0,1) \qquad 7.$$

$$\text{versus} \quad H_1^n : Y_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)N(0,1) + \epsilon_n N(A_n, \sigma^2).$$

Cai et al. (2011) discover that the detection problem behaves very differently in two regimes: the sparse regime, where $1/2 < \beta < 1$, and the dense regime, where $0 < \beta \le 1/2$. Furthermore, they show that a double-sided version of the HC test is 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 regimes. Classical methods have treated the detections of sparse and dense signals separately. In practice, however, the information of the signal sparsity is usually unknown, and the adaptivity of the modified HC test is, thus, a practically useful property.

Cai & Wu (2014) consider the problem of sparse mixture detection in a more general model (Equation 1) where the distributions are not necessarily Gaussian and the nonnull effects are

not necessarily a binary vector. They obtain an explicit formula for the fundamental limit of the general testing problem under mild conditions on the mixture, which are specifically 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 HC procedure is also generalized far beyond the setup in the works of Ingster (1998), Donoho & Jin (2004), and Cai et al. (2011). In the most general case, the 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 Nonnull Effects

The proportion of nonnull 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. The proportion of nonnull effects is also one of the key quantities in the implementation of many large-scale multiple testing procedures (see, for example, Efron et al. 2001, Storey 2007, Sun & Cai 2007). The development of useful estimates of ϵ_n, along with the corresponding statistical analysis, is a challenging task (for recent work, see Langaas et al. 2005, Meinshausen & Rice 2006, Cai et al. 2007, Jin & Cai 2007, Cai & Jin 2010).

2.2.1. Tail-based approach. Schweder & Spjøtvoll (1982) propose an intuitive method for estimating the proportion of null hypotheses using p-value plots. The methodology is developed for the general mixture model (Equation 1). To illustrate how it works, we simulate $n = 1{,}000$ observations from a simple two-point normal mixture $F(x) = (1-\epsilon_n)N(0,1) + \epsilon_n N(2,1)$. The proportion of nonnull hypotheses is $\epsilon_n = 0.2$. The histogram of the p-values is shown in **Figure 2*a***. 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 $W(\lambda) = \#\{i : p_i > \lambda\}$. Because the p-values to the right of the threshold roughly follow a uniform distribution, the expected counts 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

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

The p-value plotting method proposed by Schweder & Spjøtvoll (1982) is described in **Figure 2*b***. Benjamini & Hochberg (2000) formalize this graphical method as an asymptotically equivalent stepwise least-slope estimator (see also Benjamini et al. 2006).

Langaas et al. (2005) show that the estimate given by Equation 8 always has a downward bias, i.e., $\mathbb{E}\{\hat{\epsilon}_n(\lambda)\} \le \epsilon_n(\lambda)$ 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 & Tibshirani (2003) propose a bootstrapping method and a spline-smoothing method, respectively. Langaas et al. (2005) investigate the choice of λ systematically and develop a class of estimators based on nonparametric maximum likelihood estimates (MLEs).

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 nonnull 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 & Cai (2007) demonstrate that information on the null distribution and nonnull proportion is well-preserved in the frequency domain but not in the

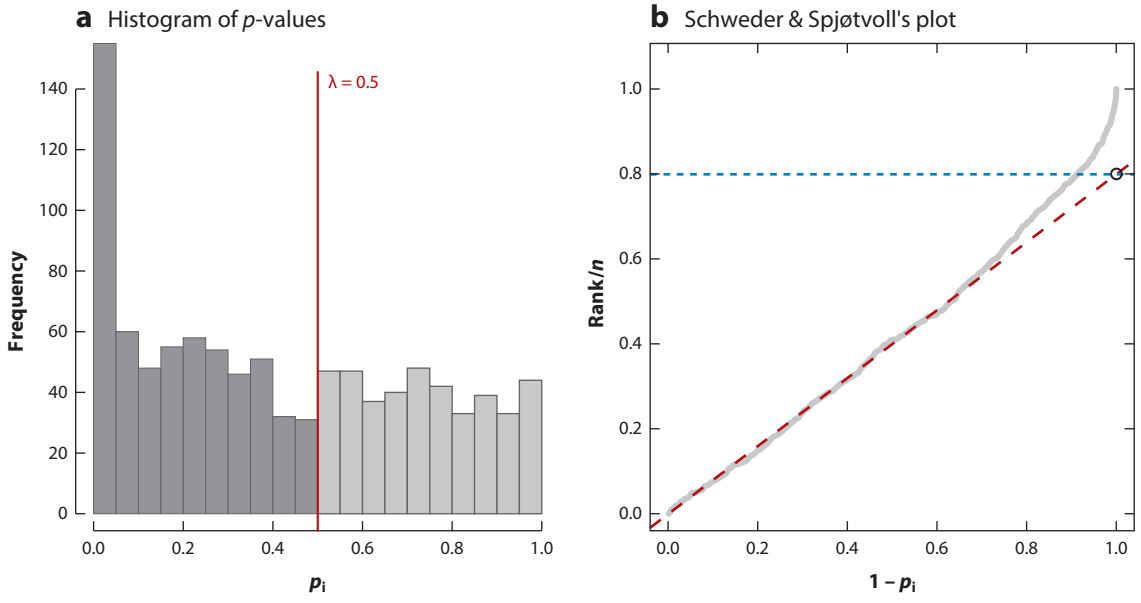

Figure 2

Tail-based methods for estimating ϵ_n. Data are simulated from a two-point normal mixture model $0.8 \cdot N(0,1) + 0.2 \cdot N(2,1)$. (*a*) Histogram of *p*-values, illustrating Equation 8 with $\lambda = 0.5$. The *p*-values to the right of the red line (*light gray bars*) follow a uniform distribution approximately. (*b*) The graphical solution of Schweder & Spjøtvoll (1982). The gray curve plots $1 - p_i$ against their rank. A horizontal line is fitted through the left portion of the gray curve and extended all the way to the right via an eyeball method. The circle represents the intersection point, which gives the estimated proportion of null cases (*blue dashed line*). The intersection point (*unfilled circle*) shows that the estimated proportion of null cases is 0.8.

spatial domain. They further propose a frequency-domain approach to estimating the proportion. The estimator is robust against heteroscedasticity and has been 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

$$X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)N(\mu_0, \sigma_0^2) + \epsilon_n Q_n, i = 1, \ldots, n, \qquad 9.$$

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 $q(x) = \int \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})\mathrm{d}H_n(\mu,\sigma)$ for some mixing distribution H_n. We discuss only the case with known null parameters (see Jin & Cai 2007 for a modified procedure for the case with unknown null parameters). Then, we can renormalize X_j and assume, without loss of generality, that $\mu_0 = 0$ and $\sigma_0 = 1$. The marginal density f of X_j becomes

$$f(x) = (1-\epsilon_n)\phi(x) + \epsilon_n \int \phi\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}H_n(\mu,\sigma). \qquad 10.$$

Jin & Cai's (2007) 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 as $\psi_n(t;\omega) = \int \omega(\xi)e^{\frac{t^2\xi^2}{2}}\varphi_n(t\xi)\mathrm{d}\xi$. 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

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

where Re(z) stands for the real part of z. In this case, the phase function $\psi_n(t)$ is used to obtain an estimate of ϵ_n, and the choice of γ reflects the bias–variance trade-off in the estimate. Typically, we set $\gamma = 0.1$ (see Cai & Jin 2010 for more discussions). In studies by Jin & 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 $\exp\left(-\frac{1}{1-|\xi|^2}\right)\cdot 1_{\{|\xi|<1\}}$. In our numerical experiments, the triangle density tends to work slightly better. Cai & Jin (2010) show that choosing ω to be the point mass at 1 leads to an asymptotically rate-optimal estimator of ϵ_n.

2.2.3. Optimality theory. The detection theory developed by Ingster (1998) and Donoho & 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 of Meinshausen & 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) develop an effective data-driven method for a two-point homoscedastic Gaussian mixture model $X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)N(0,1)+\epsilon_n N(\mu_n,1), 1 \le i \le n$, and show that the estimator is rate optimal within a logarithmic factor. In contrast to the results of Meinshausen & Rice (2006), the results of Cai et al. (2007) imply that it is possible to estimate ϵ_n consistently over the entire detectable region.

The optimality theory for estimating ϵ_n is further developed by Cai & Jin (2010) for the general Gaussian mixture model (Equation 9). Consider γ as defined in Equation 11. Cai & Jin (2010) introduce a modified estimator

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

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

2.3. Estimation of the Null Distribution

Conventionally, F_0 is assumed to be known and is referred to as the theoretical null. Efron (2004) argues that, in large-scale inference problems, the use of the theoretical null is incorrect and the choice of the null distribution has a huge impact on subsequent analysis. Efron further proposes the concept of the empirical null and argues that the empirical evidence in the data determines the normal state and that 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) proposes a simple method to estimate the null parameters utilizing the central peak of the histogram. Jin & Cai (2007) propose 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 is developed

by Cai & Jin (2010). The empirical null approach of Efron (2004) and the estimation methods of Jin & 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 by Sun & McLain (2012), who extend the method of Jin & Cai (2007) to estimate the composite null distribution with an external covariate.

3. MULTIPLE TESTING PROBLEMS

Multiple testing is a useful approach to extracting valuable insights from massive data. Recent developments in multiple testing, epitomized by FDR methodologies, have greatly influenced a wide range of scientific and business disciplines. This section reviews some important concepts and recent progress in this field.

3.1. Multiplicity, Error Rate, and Power Concepts

Two types of errors may be committed when performing a hypothesis test: rejecting a hypothesis when it is null (type I error) or failing to reject a hypothesis when it is nonnull (type II error). A type I error occurs when one finds a pattern that does not exist in the data (false discovery), whereas a type II error occurs when one misses 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 type II error rates as the probability of making type I and type II errors, respectively. The classical formulation in single-hypothesis testing aims to control the type I error rate at a prespecified level α while minimizing the type II error rate.

When n hypotheses are tested simultaneously, the outcomes of all tests can be summarized as 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.

3.1.1. Family-wise error rate. The family-wise error rate (FWER) is defined as the probability of making at least one type I error in the family, e.g., FWER $= \mathbb{P}(N_{10} \geq 1)$, where N_{10} is the number of false positive findings. The FWER has been widely used as an overall error measure when multiple hypotheses are tested at the same time. A per-comparison error rate 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 level α. The Bonferroni method can be further improved by stepwise methods, such as Holm's procedure and Hommel's procedure (Holm 1979, Hochberg 1988, Hommel 1988), or resampling-based methods (Westfall & Young 1993). We refer interested readers to Shaffer

Table 1 Classification of tested hypotheses

Hypotheses	Claimed nonsignificant	Claimed significant	Total
Null	N_{00}	N_{10}	n_0
Nonnull	N_{01}	N_{11}	n_1
Total	S	R	n

(1995) and Hochberg & 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 (for recent work, see Lehmann & Romano 2005a, Romano & Shaikh 2006, Sarkar 2007).

3.1.2. False discovery rate. The FWER is a very strict criterion. When thousands or 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 FDR (Benjamini & 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 (FDP) as

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

Thus, the FDR is the expectation of the FDP:

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

The FDR concept reflects the trade-off between false discoveries and true discoveries in practice and is connected to minimax estimation theory (Abramovich et al. 2006) and compound decision theory (Sun & Cai 2007). Other closely related measures include the positive FDR (pFDR; Storey 2003) and the marginal FDR (mFDR; Genovese & Wasserman 2002). The differences among various FDR measures seem to be nonessential 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 & Wasserman (2002) show that, under mild conditions, $\mathrm{mFDR} = \mathrm{FDR} + O(n^{-1/2})$. The FDR is fundamentally different from the FWER because it provides 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.

3.1.3. Power and optimality. In single-hypothesis testing, the power is defined as the probability of correctly rejecting a nonnull 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 prespecified test level α.

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

$$\mathrm{ETP} = \mathbb{E}(N_{11}), \tag{15.}$$

in this article. Other related measures include the average power (Spjøtvoll 1972, Efron 2007b, Storey 2007), the false negative/nondiscovery rate (FNR; Genovese & Wasserman 2002, Sarkar 2004), $\mathrm{FNR} = \mathbb{E}\left(\frac{N_{01}}{S} \,\middle|\, S > 0\right) \mathbb{P}(S > 0)$, and the missed discovery rate (MDR; Taylor et al. 2005). 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.

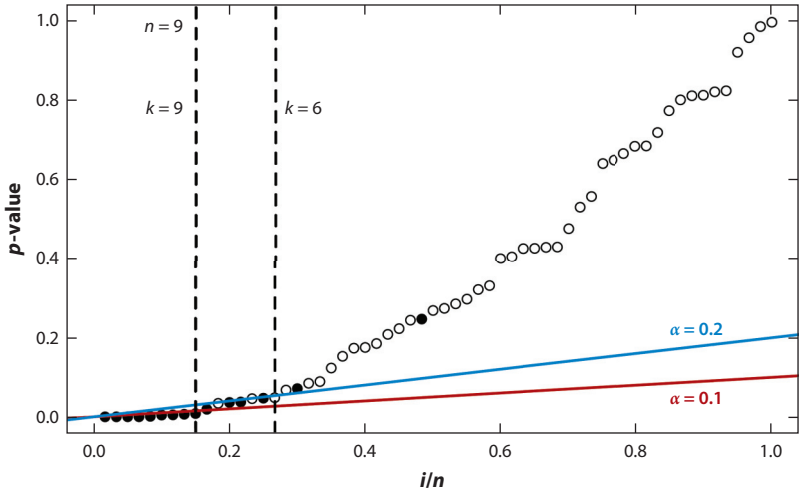

Figure 3

A graphical illustration of the Benjamini-Hochberg procedure; filled and unfilled circles stand for nonnull and null cases, respectively. The false discovery rate thresholds are computed as the largest intersection point of the p-value curve and the straight line, the slope of which 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 three false positives.

3.2. p-Value-Based Methodologies for False Discovery Rate Control

In single-hypothesis testing, the 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 are discussed in Section 3.3.

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

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

and then rejects all hypotheses $H_{(j)}, j = 1, \ldots, k$. 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 (Equation 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 red and blue 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.

Benjamini & Hochberg (1995) show that the procedure described by Equation 16 controls the FDR at the nominal level when the p-values are independent. The dependence issue is important and is discussed in Section 3.5. 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 level α, where ϵ_n is the proportion of nonnull cases. Benjamini & Hochberg (2000), Genovese & Wasserman (2002), and Storey (2002) propose 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 & Hochberg 2000) operates as follows:

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

We can see that, in Equation 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 the 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 nonnull case and $G(t)$ be the mixture CDF. Consider a random mixture model for p-values:

$$G(t) = (1-\epsilon_n)t + \epsilon_n G_1(t). \qquad 18.$$

The 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 rejects 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 & 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 $\hat{\epsilon}_n$ is 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. Thus, a class of plug-in FDR procedures can be constructed (Genovese & Wasserman 2002, 2004) as follows:

$$\text{Let } t(\hat{p}, \hat{G}) = \sup\{t : \hat{Q}(t) \le \alpha\}. \qquad \text{Reject } H_{i0} \text{ if } p_i < t(\hat{p}, \hat{G}). \qquad 19.$$

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 Equation 19 reduces to the well-known BH procedure. Genovese & Wasserman (2004) develop a stochastic process framework for multiple testing and show that, when consistent estimates of G and ϵ_n are chosen, the class of plug-in procedures (Equation 19) is 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 analog 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_{i0} if its q-value is less than α. The q-value has gained great popularity in large-scale "omics" research, such as research in 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 (Equation 18): The pFDR is defined as $\text{pFDR}(t) = \mathbb{E}\left(\frac{N_{10}}{R} \middle| R > 0\right) = (1-\epsilon_n)t/G(t)$, 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:

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

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 $p_{(1)}, \ldots, p_{(n)}$. Then, the q-value procedure works as follows:

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

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 all of the p-values from all other tests. By comparing Equation 21 with Equation 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 & Wasserman 2004) provides a useful alternative to the FDR. Let $0 \leq \tau \leq 1$ be a prespecified tolerance level; the FDX at level τ is $\text{FDX}_\tau = \mathbb{P}(\text{FDP} > \tau)$, the tail probability that the FDP exceeds a given bound. The goal is to construct a testing procedure satisfying $\text{FDX} \leq \alpha$. 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 & Romano 2005b, Genovese & Wasserman 2006, and Roquain & 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), the two-stage linear procedure (Benjamini et al. 2006), and resampling procedures (Tusher et al. 2001). 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, by Westfall & 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, mFDR, pFDR, 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) provides a good summary of error measures and discusses how to match proper error rates with inference needs.

3.3. Optimal False Discovery Rate Control: A Decision-Theoretic Approach

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

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

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

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 & Cai (2007) develop a compound decision theoretic framework and show 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 false discovery rate procedure. Consider an ideal setup where an oracle knows ϵ_n, f_0, and f_1. To develop the oracle rule, we consider two questions in turn: (*a*) What is the oracle statistic that gives the optimal ranking of all tests? (*b*) What is the oracle cutoff that controls the FDR and maximizes the ETP?

Consider Equation 1. Suppose we obtain a z-value from each test. Sun & Cai (2007) show that the optimal test statistic in the oracle setting is the local FDR (lFDR):

$$\text{lFDR}(z_i) = \frac{(1-\epsilon_n) f_0(z_i)}{f(z_i)}. \qquad 23.$$

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

$$\delta^i_{\text{OR}} = \mathbb{I}\{\text{lFDR}(z_i) < t_{\text{OR}}\}. \qquad 24.$$

Sun & Cai (2007) show that the oracle rule (Equation 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: $\text{lFDR}(z_i) = \mathbb{P}(\text{case } i \text{ is null} \mid z_i)$ (Efron et al. 2001). It captures all important distributional information in the mixture model (Equation 1). The expression in Equation 23 implies that we actually rank the hypotheses according to the ratio f_0/f and that 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 is illustrated in Section 3.3.3 using data on mutual funds.

3.3.2. A data-driven procedure. The oracle procedure cannot be implemented in practice because both the lFDR and t_{OR} are unknown. In this section, 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 estimated as an empirical null using methods proposed by Efron (2004) and Jin & 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

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

Next, we derive a data-driven procedure that mimics the oracle procedure. We use the idea of ranking followed by thresholding to motivate a stepwise method. Use $\widehat{\text{lFDR}}_{(1)} \le \cdots \le \widehat{\text{lFDR}}_{(n)}$ to denote the ordered lFDR statistics. Suppose j hypotheses are rejected along the ranking; then, the actual FDR level can be estimated as $\hat{Q}_{\text{OR}}(j) = \frac{1}{j}\sum_{i=1}^{j}\widehat{\text{lFDR}}_{(i)}$, the moving average of the top j ordered statistics (Sun & Cai 2007). To fulfill the FDR constraint and maximize the power, we propose the following stepwise procedure:

$$\text{Let } k = \max\left\{j : \frac{1}{j}\sum_{i=1}^{j}\widehat{\text{lFDR}}_{(i)} \le \alpha\right\}, \text{ then reject all } H_{(i)0},\ i = 1,\ldots,k. \qquad 25.$$

The goals of global FDR control and individual case interpretation are naturally unified in the data-driven procedure (Equation 25). Moreover, with the consistent estimators proposed by Jin & Cai (2007), Sun & Cai (2007) show 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*-values and local false discovery rate. 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 nonnull cases, respectively. The model is considered by Barras et al. (2010) for analysis of mutual funds data, where $N(0,1)$, $N(\mu^-,1)$, and $N(\mu^+,1)$ are used to describe the distributions of zero alpha funds, unskilled funds, and skilled funds, respectively. We choose a setting in which the main findings of Barras et al. (2010) can be roughly matched. Specifically, $n = 5{,}000$ z-values are simulated from the mixture model with $\mu^- = -2.5$, $\mu^+ = 3$, $\epsilon_n^- = 0.15$, and $\epsilon_n^+ = 0.05$. Thus, many funds have underperformance but few have outperformance. The histograms of zero, positive, and negative components are illustrated 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 instead 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 & Hochberg 1995), the adaptive p-value (AP; Benjamini & Hochberg 2000) procedure, and the data-driven lFDR procedure (Sun & Cai 2007) to the data set at $\alpha = 0.1$. The results are summarized in **Table 2**.

We can see that the lFDR procedure controls the FDP more precisely compared to the p-value-based methods. Moreover, it correctly identifies more nonzero 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^+; thus, we are more likely to find signals in the negative component. Therefore, it makes sense to adopt an asymmetric rejection region when selecting nonzero 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 $z_i > 2.73$. Interestingly, 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.

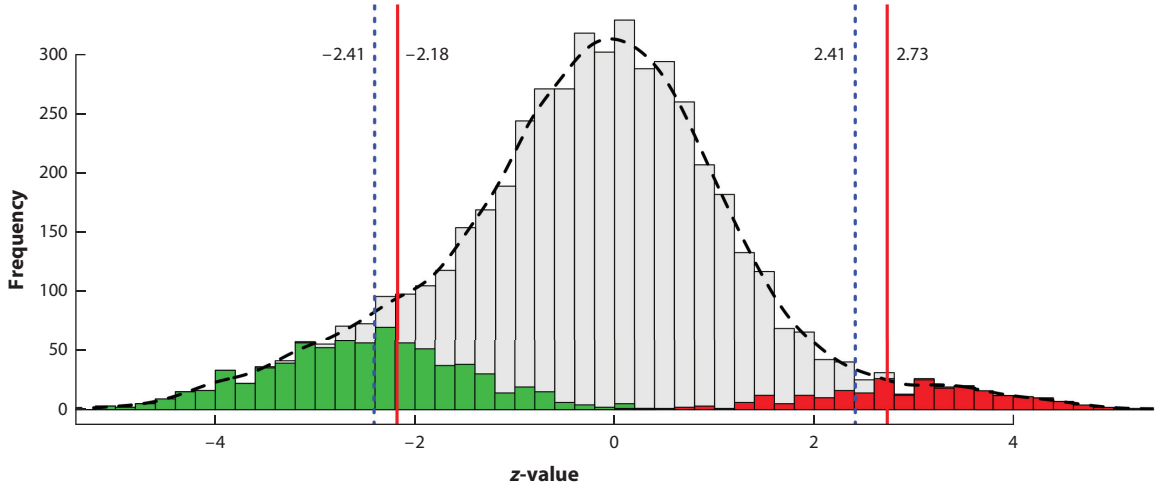

Figure 4

Mutual funds example, showing symmetric versus 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 one is more likely to find signals in the negative part. The local false discovery rate (lFDR) procedure allows us 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 adaptive p-value method is $|z| > 2.41$. The grey bars represent observations from the mixture distribution. The green and red bars correspond to observations from $N(-2.5, 1)$ and $N(3, 1)$, respectively.

We then adjust the p-value threshold of BH as 0.0185, which leads to a symmetric rejection region $z < -2.35$ and $z > 2.35$ with an FDP of 0.098 (the numbers of rejections and false positives are 668 and 66, respectively; see the last row of **Table 2**). Thus the comparison is on an equal footing. At the same FDP level, the lFDR method rejects more true positives than the p-value method (626 versus 602). This demonstrates the superiority of the lFDR ranking.

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, thus, that 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

Table 2 Analysis summary for simulated mutual funds data

Methods	Number of rejections	Number of 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
Adjusted BH	668	602	0.098	−2.35	2.35

Abbreviations: AP, adaptive p-value; BH, Benjamini-Hochberg; FDP, false discovery proportion; lFDR, local false discovery rate.

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 the precision of statistical inference.

3.4.1. Heterogeneity and grouping. The problem of multiple testing with groups and related problems are studied by Efron (2008), Ferkingstad et al. (2008), Cai & Sun (2009), and 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 are much wider than those for 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 center 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) introduces the multigroup mixture model to handle the heterogeneity in the data. Suppose $X_1, \ldots, X_n$ can be divided into K groups:

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

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. Efron (2008) argues that a pooled FDR analysis is problematic because highly significant cases from one group may be hidden among the nulls from another group, whereas insignificant cases may possibly be enhanced. Efron (2008) suggests a second approach, the separate analysis, which first conducts an FDR analysis at level α within each group and then combines the testing results from all groups. Efron (2008) shows that the separate analysis controls the FDR. However, the choice of identical FDR levels across all groups can be suboptimal. Cai & Sun (2009) show that both the separate and pooled analyses can be uniformly improved by a third approach, the conditional lFDR (clFDR) method, which enjoys features of both pooled and separate analyses. Let $\hat{p}_k$, $\hat{f}_{k0}$, and $\hat{f}_k$ be estimates of the unknown quantities in Equation 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 $\widehat{\text{clFDR}}_{(1)}, \ldots, \widehat{\text{clFDR}}_{(n)}$ the ranked values and by $H_{(1)}, \ldots, H_{(n)}$ the corresponding hypotheses.
3. Reject all $H_{(i)}$, $i = 1, \ldots, l$, 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 the first step, the external information of group labels is utilized to calculate the clFDR; this is the feature from the separate analysis. However, in the second and third steps, the group labels are dropped and the rankings of all hypotheses are determined globally; this is the feature from the pooled analysis. Cai & Sun (2009) show that the clFDR procedure is asymptotically valid and optimal. Unlike in the separate analysis, the groupwise 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 toward 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 & Wasserman 2009, Roquain & Van De Wiel 2009). A popular scheme, referred to as the decision weights approach, involves modifying the error criteria or power functions (Benjamini & Hochberg 1997). The idea is to employ two sets of positive constants, $\mathbf{a} = \{a_i : i = 1, \ldots, n\}$ and $\mathbf{b} = \{b_i : i = 1, \ldots, n\}$, 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 by Benjamini & 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 $\text{ETP} = \mathbb{E}\left(\sum_{i=1}^{n} b_i\theta_i\delta_i\right)$, 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. With the combination of the concerns on both the error criterion and power function, the goal in weighted multiple testing is to maximize the ETP subject to the constraint $\text{wFDR} \leq \alpha$.

Basu et al. (2015) develop an asymptotically optimal solution to this problem. The key step involves a conceptualization of the constrained optimization problem as an expanding knapsack problem, followed by an application of the classical ideas in the 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 prespecified 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 increasingly finer resolution levels going from the top to the bottom of the tree. At each node of a given level, we have three possible actions: (*a*) testing, i.e., deciding whether a unit contains one of the patterns of interest; (*b*) classification, i.e., assigning the selected subjects to specific pattern categories (classification); and (*c*) indecision, i.e., selecting a subject as a signal but not specifying its pattern.

In hierarchical testing, important error measures for summarizing the whole decision process include the full-tree and outer-node FDRs (Yekutieli 2008), the focus-level FDR (Goeman & Mansmann 2008), the mixed-directional FDR (Guo et al. 2010), and the overall FDR (Sun & 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 & Geman 2005, Goeman & Mansmann 2008, Meinshausen 2008, Yekutieli 2008, Goeman & Solari 2010, Sun & Wei 2015). Recent works

on multiple comparison issues in multistage and sequential testing problems include those of Benjamini et al. (2006), Lin (2006), Benjamini & Heller (2007), Dmitrienko et al. (2007), Posch et al. (2009), Liang & Nettleton (2010), Sarkar et al. (2013), Benjamini & Bogomolov (2014), and Cai & 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: (*a*) What is the impact of dependence on the conventional FDR analysis? (*b*) How can we construct new FDR procedures for dependent tests?

3.5.1. Impact of dependence in multiple testing. The performance of a multiple testing procedure is reflected in both the power and validity. Benjamini & Yekutieli (2001) show that applying the BH procedure at level $\alpha/(\sum_{i=1}^{n} 1/i)$ always controls the FDR at level α under arbitrary dependence among the p-values. However, such an adjustment is too conservative and often unnecessary in practice. In the multiple testing literature, extensive efforts have been devoted to the study of the impact of dependence on the validity of FDR control when applying the BH procedure to dependent tests without any adjustments. 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 some regularity conditions (see, e.g., Benjamini & Yekutieli 2001, Sarkar 2002, Storey et al. 2004, Wu 2008, Clarke & Hall 2009). For many applications in economics and finance, these assumptions do not hold. Practitioners should verify them carefully and proceed with caution. Second, Efron (2007a) and Schwartzman & Lin (2011) show that correlation usually degrades statistical accuracy, affecting both estimation and testing. High correlation also results in high variability of testing results and, thus, irreproducibility of scientific findings (see Owen 2005 and Finner et al. 2007 for related discussions). These results suggest that dependency has a negative impact and must be adjusted for multiple testing, especially when the correlations are very high. Leek & Storey (2008) and Friguet et al. (2009) study multiple testing under the factor models and show that, by subtracting the common factors out, the dependence structure can be greatly weakened. Efron (2007a) and Fan et al. (2012) discuss how to take into account the dependence structure and obtain more accurate FDR estimates for a given p-value threshold.

The problem of FDR control under general dependence structures still requires much research. For example, existing work has been focused mainly on validity, and the important power issue has been largely ignored. In fact, most p-value-based methods suffer from efficiency loss when the dependence structure is highly informative. More powerful testing procedures can be constructed by exploiting the correlations between the tests. It remains an open problem to develop a general framework to estimate the unknown dependence structure and then use it for efficient multiple testing. We discuss in the following sections some recent progress in this direction, focusing on settings where the dependency structures can be well estimated from data.

3.5.2. Exploiting dependence for multiple testing. 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 by exploiting high correlations. The works of Benjamini & Heller (2007), Sun & Cai (2009), and Sun & Wei (2011) show 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 nonnull hypotheses and expects that the non-null cases will appear in clusters. Suppose the observed sequence is

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

where $x_1 = x_2 = 2$. Heuristically, we can argue that x_1 is likely to come from the nonnull 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. The 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 nonnull hypotheses and expect that the nonnulls appear in clusters. In an HMM, the sequence of the unknown (hidden) null and nonnull states is assumed to form a Markov chain $(\theta_i)_1^n = (\theta_1, \ldots, \theta_n) \in \{0,1\}^n$. The observed data values $\mathbf{x} = (x_1, \ldots, x_n)$ are independent conditional on the hidden states $(\theta_i)_1^n$. Let ϑ denote the collection of all HMM parameters.

Sun & Cai (2009) show that, under the HMM dependency, the optimal test statistic is the local index of significance (LIS), $\mathrm{LIS}_i = \mathbb{P}_\vartheta(\theta_i = 0|\mathbf{x})$, which can be computed using a fast forward–backward algorithm. The LIS is superior to the p-value because it utilizes the HMM dependence to pool information from nearly observations. The information from the whole sequence is integrated to calculate the LIS statistic. With the use of 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 $\widehat{\mathrm{LIS}}_i = \mathbb{P}_{\hat{\vartheta}}(\theta_i = 0|\mathbf{x})$. The MLE is commonly used and is strongly consistent and asymptotically normal (Leroux 1992, Bickel et al. 1998). The MLE can be computed using the expectation-maximization algorithm or other standard optimization schemes. Denote by $\widehat{\mathrm{LIS}}_{(1)}, \ldots, \widehat{\mathrm{LIS}}_{(n)}$ the ranked plug-in test statistics and by $H_{(1)}, \ldots, H_{(n)}$ the corresponding hypotheses. The following data-driven procedure can be used for FDR control:

$$\text{Let } k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} \widehat{\mathrm{LIS}}_{(j)} \le \alpha \right\}, \text{ then reject all } H_{(i)},\ i = 1, \ldots, k. \qquad 27.$$

Sun & Cai (2009) show 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 nonnull 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: pointwise 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 & Richardson 2002), and network analysis (Wei & 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. In this section, 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) $\mathbf{X} = \{X(s) : s \in S\}$ of 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 measure for a continuous domain (or a counting measure for a discrete domain). The spatial FDR can be defined as

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

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

Let $\mathbf{x}^N = (x_1, \ldots, x_N)$ denote the observed values. Suppose an oracle knows all RFM parameters, denoted by Ψ. The oracle statistic for pointwise inference is $T_{\mathrm{OR}}(s) = \mathbb{P}_\Psi\{\theta(s) = 0|\mathbf{x}^N\}$. However, this requires testing an uncountable number of hypotheses for all $s \in S$, which is impossible in practice. Sun et al. (2015) show 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 $T_{\mathrm{OR}}^{(1)} \le T_{\mathrm{OR}}^{(2)} \le \cdots \le T_{\mathrm{OR}}^{(n)}$ denote the ordered oracle statistics and $S_{(i)}$ the corresponding regions. In a pointwise inference, define $R_j = \cup_{i=1}^j S_{(i)}$ and $r = \max\left\{j : \nu(R_j)^{-1}\sum_{i=1}^j T_{\mathrm{OR}}^{(i)}\nu(S_{(i)}) \le \alpha\right\}$. 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 Markov chain Monte Carlo (MCMC) computing (see Sun et al. 2015 for detailed algorithms).

3.5.5. Clusterwise and setwise inference. When the focus is on the behavior of a process over subregions, 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 & Heller 2008, Sun & Wei 2011).

Let $\mathcal{C} = \{C_1, \ldots, C_K\}$ 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/nonnull and 0 otherwise. The decision for cluster k is denoted by 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 clusterwise procedure:

$$\mathrm{FCR} = \mathbb{E}\left\{\frac{\sum_k w_k(1 - \vartheta_k)\Delta_k}{(\sum_k w_k\Delta_k) \vee 1}\right\}, \qquad 28.$$

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

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 in 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 review does not cover this important topic. We refer interested readers to Cai (2017) 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.

The recent innovation of FDR control also provides a powerful regularization method for estimation of sparse vectors (Abramovich et al. 2006), large covariance matrices (Bailey et al. 2014), and Gaussian graphical models (Liu 2013). Sharp asymptotic optimality results have been established by exploiting the data-adaptive nature of the BH thresholding scheme (Abramovich et al. 2006, Wu & Zhou 2013). These works reveal the interesting connection between testing and estimation problems in high-dimensional inference. Another topic that is not discussed in this review is simultaneous inference for high-dimensional regression models, which has received much recent attention (see, for example, Javanmard & Montanari 2014, Liu & Luo 2014, Lockhart et al. 2014, Van de Geer et al. 2014, Zhang & Zhang 2014, Barber & Candès 2015, Xia et al. 2017, Cai & Guo 2017). Recent works also reveal that multiple testing, and in particular the FDR control, provides a promising regularization principle for variable selection in high-dimensional regression models (Liu & Luo 2014, Chudik et al. 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 postselection variables are biased if the selection effects are not taken into account. Postselection inference techniques are useful in classical statistical problems such as the estimation of many normal means and simultaneous confidence intervals (Benjamini & Yekutieli 2005, Brown & Greenshtein 2009, Efron 2011), as well as rapidly growing areas such as high-dimensional regression and sparse principal components analysis (see Leeb & Pötscher 2005; Stoye 2009; Belloni et al. 2012, 2014a,b; Yekutieli 2012; Hwang & Zhao 2013; Berk et al. 2013; Benjamini & Bogomolov 2014; Taylor & 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 T.T.C. was supported in part by National Science Foundation (NSF) grants DMS-1208982 and DMS-1403708 and National Institutes of Health grant R01 CA127334; the research of W.S. was supported in part by NSF grant DMS-CAREER 1255406. The authors would like to thank the Editor and reviewers for helpful comments and references.

LITERATURE CITED

Abramovich F, Benjamini Y, Donoho DL, Johnstone IM. 2006. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Stat.* 34:584–653

Andreou E, Ghysels E. 2006. Monitoring disruptions in financial markets. *J. Econom.* 135(1):77–124

Bailey N, Pesaran M, Smith LV. 2014. *A multiple testing approach to the regularisation of large sample correlation matrices*. CESifo Work. Pap. 4834, CESifo Group, Munich

Barber RF, Candès EJ. 2015. Controlling the false discovery rate via knockoffs. *Ann. Stat.* 43:2055–85

Barras L, Scaillet O, Wermers R. 2010. False discoveries in mutual fund performance: measuring luck in estimated alphas. *J. Finance* 65:179–216

Basu P, Cai TT, Das K, Sun W. 2015. Weighted false discovery rate control in large-scale multiple testing. arXiv:1508.01605 [stat.ME]

Belloni A, Chen D, Chernozhukov V, Hansen C. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6):2369–429

Belloni A, Chernozhukov V, Hansen C. 2014a. High-dimensional methods and inference on structural and treatment effects. *J. Econ. Perspect.* 28(2):29–50

Belloni A, Chernozhukov V, Kato K. 2014b. Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* 102(1):77–94

Benjamini Y. 2010. Simultaneous and selective inference: current successes and future challenges. *Biom. J.* 52:708–21

Benjamini Y, Bogomolov M. 2014. Selective inference on multiple families of hypotheses. *J. R. Stat. Soc. B* 76:297–318

Benjamini Y, Heller R. 2007. False discovery rates for spatial signals. *J. Am. Stat. Assoc.* 102:1272–81

Benjamini Y, Heller R. 2008. Screening for partial conjunction hypotheses. *Biometrics* 64:1215–22

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

Benjamini Y, Hochberg Y. 1997. Multiple hypotheses testing with weights. *Scand. J. Stat.* 24:407–18

Benjamini Y, Hochberg Y. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* 25:60–83

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

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29:1165–88

Benjamini Y, Yekutieli D. 2005. False discovery rate—adjusted multiple confidence intervals for selected parameters. *J. Am. Stat. Assoc.* 100:71–81

Berk R, Brown L, Buja A, Zhang K, Zhao L. 2013. Valid post-selection inference. *Ann. Stat.* 41:802–37

Bickel PJ, Ritov Y, Ryden T. 1998. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Stat.* 26:1614–35

Blanchard G, Geman D. 2005. Hierarchical testing designs for pattern recognition. *Ann. Stat.* 33:1155–202

Brown LD, Greenshtein E. 2009. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Stat.* 37:1685–704

Cai TT. 2017. Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annu. Rev. Stat. Appl.* 4:423–46

Cai TT, Guo Z. 2017. Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *Ann. Stat.* 45:615–46

Cai TT, Jeng XJ, Jin J. 2011. Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. B* 73:629–62

Cai TT, Jin J. 2010. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Stat.* 38:100–45

Cai TT, Jin J, Low MG. 2007. Estimation and confidence sets for sparse normal mixtures. *Ann. Stat.* 35:2421–49

Cai TT, Sun W. 2009. Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *J. Am. Stat. Assoc.* 104:1467–81

Cai TT, Sun W. 2016. Optimal screening and discovery of sparse signals with applications to multistage high-throughput studies. *J. R. Stat. Soc. B.* 79:197–223

Cai TT, Wu Y. 2014. Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inf. Theory* 60:2217–32

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

Chudik A, Kapetanios G, Pesaran MH. 2016. *A one-covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models*. Glob. Monet. Policy Inst. Work. Pap. 290, Fed. Reserve Bank Dallas, TX. **https://ssrn.com/abstract=2874165**

Clarke S, Hall P. 2009. Robustness of multiple testing procedures against dependence. *Ann. Stat.* 37:332–58

Dmitrienko A, Wiens BL, Tamhane AC, Wang X. 2007. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat. Med.* 26:2465–78

Donoho D, Jin J. 2004. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.* 32:962–94

Dudoit S, Shaffer JP, Boldrick JC. 2003. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* 18:71–103

Efron B. 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.* 99:96–104

Efron B. 2007a. Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.* 102:93–103

Efron B. 2007b. Size, power and false discovery rates. *Ann. Stat.* 35:1351–77

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

Efron B. 2011. Tweedie's formula and selection bias. *J. Am. Stat. Assoc.* 106:1602–14

Efron B, Tibshirani R, Storey JD, Tusher V. 2001. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* 96:1151–60

Fan J, Han X, Gu W. 2012. Estimating false discovery proportion under arbitrary covariance dependence. *J. Am. Stat. Assoc.* 107:1019–35

Ferkingstad E, Frigessi A, Rue H, Thorleifsson G, Kong A. 2008. Unsupervised empirical Bayesian multiple testing with external covariates. *Ann. Appl. Stat.* 2:714–35

Finner H, Dickhaus T, Roters M. 2007. Dependency and false discovery rate: asymptotics. *Ann. Stat.* 35:1432–55

Friguet C, Kloareg M, Causeur D. 2009. A factor model approach to multiple testing under dependence. *J. Am. Stat. Assoc.* 104:1406–15

Fryzlewicz P. 2014. Wild binary segmentation for multiple change-point detection. *Ann. Stat.* 42(6):2243–81

Genovese CR, Lazar NA, Nichols T. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–78

Genovese CR, Roeder K, Wasserman L. 2006. False discovery control with *p*-value weighting. *Biometrika* 93:509–24

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

Genovese CR, Wasserman L. 2004. A stochastic process approach to false discovery control. *Ann. Stat.* 32:1035–61

Genovese CR, Wasserman L. 2006. Exceedance control of the false discovery proportion. *J. Am. Stat. Assoc.* 101:1408–17

Goeman JJ, Mansmann U. 2008. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 24:537–44

Goeman JJ, Solari A. 2010. The sequential rejection principle of familywise error control. *Ann. Stat.* 38:3782–810
Goeman JJ, Solari A, Stijnen T. 2010. Three-sided hypothesis testing: simultaneous testing of superiority, equivalence and inferiority. *Stat. Med.* 29:2117–25
Green PJ, Richardson S. 2002. Hidden Markov models and disease mapping. *J. Am. Stat. Assoc.* 97:1055–70
Guo W, Sarkar SK, Peddada SD. 2010. Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics* 66:485–92
Hall P, Jin J. 2010. Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Stat.* 38:1686–732
Harvey CR, Liu Y. 2015. Backtesting. *J. Portf. Manag.* 42:13–28
Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y. 2006. Cluster-based analysis of fMRI data. *Neuroimage* 33:599–608
Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–2
Hochberg Y, Tamhane AC. 2009. *Multiple Comparison Procedures*. Hoboken, NJ: Wiley
Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6:65–70
Hommel G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75:383–86
Hu JX, Zhao H, Zhou HH. 2012. False discovery rate control with groups. *J. Am. Stat. Assoc.* 105:1215–27
Hwang JG, Zhao Z. 2013. Empirical Bayes confidence intervals for selected parameters in high-dimensional data. *J. Am. Stat. Assoc.* 108:607–18
Ingster YI. 1998. Minimax detection of a signal for l^n-balls. *Math. Methods Stat.* 7:401–28
Jager L, Wellner JA. 2007. Goodness-of-fit tests via phi-divergences. *Ann. Stat.* 35:2018–53
Javanmard A, Montanari A. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* 15:2869–909
Jin J. 2008. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. B* 70:461–93
Jin J, Cai TT. 2007. Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Am. Stat. Assoc.* 102:495–506
Langaas M, Lindqvist BH, Ferkingstad E. 2005. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. B* 67:555–72
Lee JD, Sun DL, Sun Y, Taylor JE. 2016. Exact post-selection inference, with application to the lasso. *Ann. Stat.* 44:907–27
Leeb H, Pötscher BM. 2005. Model selection and inference: facts and fiction. *Econom. Theory* 21(1):21–59
Leek JT, Storey JD. 2008. A general framework for multiple testing dependence. *PNAS* 105:18718–23
Lehmann EL, Romano JP. 2005a. Generalizations of the familywise error rate. *Ann. Stat.* 33:1138–54
Lehmann EL, Romano JP. 2005b. *Testing Statistical Hypotheses*. New York: Springer. 3rd ed.
Leroux BG. 1992. Maximum-likelihood estimation for hidden Markov models. *Stoch. Process. Appl.* 40:127–43
Liang K, Nettleton D. 2010. A hidden Markov model approach to testing multiple hypotheses on a tree-transformed gene ontology graph. *J. Am. Stat. Assoc.* 105:1444–54
Lin DY. 2006. Evaluating statistical significance in two-stage genomewide association studies. *Am. J. Hum. Genet.* 78:505–9
Liu W. 2013. Gaussian graphical model estimation with false discovery rate control. *Ann. Stat.* 41(6):2948–78
Liu W, Luo S. 2014. *Hypothesis testing for high-dimensional regression models*. Tech. Rep., Shanghai Jiao Tong Univ., Shanghai
Lo AW, MacKinlay AC. 1990. Data-snooping biases in tests of financial asset pricing models. *Rev. Financ. Stud.* 3:431–67
Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. 2014. A significance test for the lasso. *Ann. Stat.* 42:413–68
Lumsdaine RL, Papell DH. 1997. Multiple trend breaks and the unit-root hypothesis. *Rev. Econ. Stat.* 79(2):212–18
Meinshausen N. 2008. Hierarchical testing of variable importance. *Biometrika* 95:265–78
Meinshausen N, Rice J. 2006. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Stat.* 34:373–93

Newton MA, Noueiry A, Sarkar D, Ahlquist P. 2004. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5:155–76
Owen AB. 2005. Variance of the number of false discoveries. *J. R. Stat. Soc. B* 67:411–26
Pacifico M, Genovese C, Verdinelli I, Wasserman L. 2004. False discovery control for random fields. *J. Am. Stat. Assoc.* 99:1002–14
Posch M, Zehetmayer S, Bauer P. 2009. Hunting for significance with the false discovery rate. *J. Am. Stat. Assoc.* 104:832–40
Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–86
Robbins H. 1951. Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Berkeley Symp. Math. Stat. Probab., 2nd, Berkeley*, pp. 131–48. Berkeley, CA: Univ. Calif. Press
Roeder K, Wasserman L. 2009. Genome-wide significance levels and weighted hypothesis testing. *Stat. Sci.* 24:398–413
Rogosa D. 2003. *Accuracy of API index and school base report elements: 2003 Academic Performance Index, California Department of Education*. Work. Pap., Stanford Univ., Stanford, CA
Romano JP, Shaikh AM. 2006. Stepup procedures for control of generalizations of the familywise error rate. *Ann. Stat.* 34:1850–73
Roquain E, Van De Wiel MA. 2009. Optimal weighting for false discovery rate control. *Electron. J. Stat.* 3:678–711
Roquain E, Villers F. 2011. Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Stat.* 39:584–612
Sarkar SK. 2002. Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Stat.* 30:239–57
Sarkar SK. 2004. FDR-controlling stepwise procedures and their false negatives rates. *J. Stat. Plan. Inference* 125:119–37
Sarkar SK. 2007. Stepup procedures controlling generalized FWER and generalized FDR. *Ann. Stat.* 35:2405–20
Sarkar SK, Chen J, Guo W. 2013. Controlling the false discovery rate in two-stage combination tests for multiple endpoints. *J. Am. Stat. Assoc.* 108:1385–401
Schwartzman A, Dougherty RF, Taylor JE. 2008. False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Stat.* 2:153–75
Schwartzman A, Lin X. 2011. The effect of correlation in false discovery rate estimation. *Biometrika* 98:199–214
Schweder T, Spjøtvoll E. 1982. Plots of *p*-values to evaluate many tests simultaneously. *Biometrika* 69:493–502
Shaffer JP. 1995. Multiple hypothesis testing. *Annu. Rev. Psychol.* 46:561–84
Shorack GR, Wellner JA. 2009. *Empirical Processes with Applications to Statistics*. Philadelphia: SIAM
Silverman BW. 1986. *Density Estimation for Statistics and Data Analysis*. Boca Raton, FL: CRC Press
Spjøtvoll E. 1972. On the optimality of some multiple comparison procedures. *Ann. Math. Stat.* 43:398–411
Stock JH, Watson MW. 2012. Generalized shrinkage methods for forecasting using many predictors. *J. Bus. Econ. Stat.* 30(4):481–93
Storey JD. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. B* 64:479–98
Storey JD. 2003. The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Stat.* 31:2013–35
Storey JD. 2007. The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Stat. Soc. B* 69:347–68
Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B* 66:187–205
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *PNAS* 100:9440–45
Stoye J. 2009. More on confidence intervals for partially identified parameters. *Econometrica* 77(4):1299–315
Sun W, Cai TT. 2007. Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.* 102:901–12
Sun W, Cai TT. 2009. Large-scale multiple testing under dependence. *J. R. Stat. Soc. B* 71:393–424
Sun W, McLain A. 2012. Multiple testing of composite null hypotheses in heteroscedastic models. *J. Am. Stat. Assoc.* 107:673–87

Sun W, Reich BJ, Cai TT, Guindani M, Schwartzman A. 2015. False discovery control in large-scale spatial multiple testing. *J. R. Stat. Soc. B* 77:59–83

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

Sun W, Wei Z. 2015. Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika* 32:1823–31

Taylor J, Tibshirani RJ. 2015. Statistical learning and selective inference. *PNAS* 112:7629–34

Taylor J, Tibshirani RJ, Efron B. 2005. The 'miss rate' for the analysis of gene expression data. *Biostatistics* 6:111–17

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98:5116–21

Van de Geer S, Bühlmann P, Ritov Y, Dezeure R. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.* 42:1166–202

van der Laan MJ, Dudoit S, Pollard KS. 2004. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* 3:1–25

Wei Z, Li H. 2007. A Markov random field model for network-based analysis of genomic data. *Bioinformatics* 23:1537–44

Westfall PH, Young SS. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Hoboken, NJ: Wiley

White H. 2000. A reality check for data snooping. *Econometrica* 68:1097–126

Wu WB. 2008. On false discovery control under dependence. *Ann. Stat.* 36:364–80

Wu Z, Zhou HH. 2013. Model selection and sharp asymptotic minimaxity. *Probab. Theory Relat. Fields* 156(1–2):165–91

Xia Y, Cai T, Cai TT. 2017. Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Stat. Sin.* In press

Yekutieli D. 2008. Hierarchical false discovery rate–controlling methodology. *J. Am. Stat. Assoc.* 103:309–16

Yekutieli D. 2012. Adjusted Bayesian inference for selected parameters. *J. R. Stat. Soc. B* 74:515–41

Zhang C-H, Zhang SS. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. B* 76:217–42

Annual Review of Economics

Volume 9, 2017

Contents

Tony Atkinson on Poverty, Inequality, and Public Policy: The Work and Life of a Great Economist
Anthony Barnes Atkinson and Nicholas Stern 1

Quantitative Spatial Economics
Stephen J. Redding and Esteban Rossi-Hansberg 21

Trade and the Environment: New Methods, Measurements, and Results
Jevan Cherniwchan, Brian R. Copeland, and M. Scott Taylor 59

Bestseller Lists and the Economics of Product Discovery
Alan T. Sorensen 87

Set Identification, Moment Restrictions, and Inference
Christian Bontemps and Thierry Magnac 103

Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure
Leonardo Bursztyn and Robert Jensen 131

Quantile Regression: 40 Years On
Roger Koenker 155

Globalization and Labor Market Dynamics
John McLaren 177

High-Skilled Migration and Agglomeration
Sari Pekkala Kerr, William Kerr, Çağlar Özden, and Christopher Parsons 201

Agricultural Insurance and Economic Development
Shawn A. Cole and Wentao Xiong 235

Conflict and Development
Debraj Ray and Joan Esteban 263

Quantitative Trade Models: Developments and Challenges
Timothy J. Kehoe, Pau S. Pujolàs, and Jack Rossbach 295

The Economics of Nonmarital Childbearing and the Marriage Premium for Children
Melissa S. Kearney and Phillip B. Levine 327

The Formation of Consumer Brand Preferences
Bart J. Bronnenberg and Jean-Pierre Dubé 353

Health, Health Insurance, and Retirement: A Survey
Eric French and John Bailey Jones 383

Large-Scale Global and Simultaneous Inference: Estimation and Testing in Very High Dimensions
T. Tony Cai and Wenguang Sun 411

How Do Patents Affect Research Investments?
Heidi L. Williams 441

Nonlinear Panel Data Methods for Dynamic Heterogeneous Agent Models
Manuel Arellano and Stéphane Bonhomme 471

Mobile Money
Tavneet Suri 497

Nonparametric Welfare Analysis
Jerry A. Hausman and Whitney K. Newey 521

The History and Economics of Safe Assets
Gary Gorton 547

Global Liquidity: A Selective Review
Benjamin H. Cohen, Dietrich Domanski, Ingo Fender, and Hyun Song Shin 587

Indexes

Cumulative Index of Contributing Authors, Volumes 5–9 613

Cumulative Index of Article Titles, Volumes 5–9 616

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://www.annualreviews.org/errata/economics