J. R. Statist. Soc. B (2008)
70, *Part* 3, *pp.* 461–493

Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators

Jiashun Jin

Purdue University, West Lafayette, and Carnegie Mellon University, Pittsburgh, USA

[Received November 2006. Revised September 2007]

OnlineOpen: This article is available free online at www.blackwell-synergy.com

Summary. Since James and Stein's seminal work, the problem of *estimating n normal means* has received plenty of enthusiasm in the statistics community. Recently, driven by the fast expansion of the field of large-scale multiple testing, there has been a resurgence of research interest in the *n* normal means problem. The new interest, however, is more or less concentrated on *testing n normal means*: to determine simultaneously which means are 0 and which are not. In this setting, the *proportion* of the non-zero means plays a key role. Motivated by examples in genomics and astronomy, we are particularly interested in estimating the proportion of non-zero means, i.e. given *n* independent normal random variables with individual means $X_j \sim N(\mu_j, 1)$, $j = 1, \ldots, n$, to estimate the proportion $\varepsilon_n = (1/n)\#\{j : \mu_j \neq 0\}$. We propose a general approach to construct the universal oracle equivalence of the proportion. The construction is based on the underlying characteristic function. The oracle equivalence reduces the problem of estimating the proportion to the problem of estimating the oracle, which is relatively easier to handle. In fact, the oracle equivalence naturally yields a family of estimators for the proportion, which are consistent under mild conditions, uniformly across a wide class of parameters. The approach compares favourably with recent works by Meinshausen and Rice, and Genovese and Wasserman. In particular, the consistency is proved for an unprecedentedly broad class of situations; the class is almost the largest that can be hoped for without further constraints on the model. We also discuss various extensions of the approach, report results on simulation experiments and make connections between the approach and several recent procedures in large-scale multiple testing, including the false discovery rate approach and the local false discovery rate approach.

Keywords: Fourier transform; Oracle; Phase function; Uniform consistency

1. Introduction

Consider *n* independent normal random variables

$$X_j = \mu_j + z_j, \qquad j = 1, \ldots, n, \tag{1.1}$$

where $z_j \sim^{\mathrm{IID}} N(0, 1)$ and μ_j are unknown parameters. In the literature, the setting is referred to as the problem of *n normal means*. Frequently, a signal–noise scenario is used to describe the setting, where a data point is thought to contain a signal if the corresponding mean is non-zero and is regarded as pure noise otherwise (Abramovich *et al.*, 2006). Since James and Stein's seminal work on shrinkage estimation (James and Stein, 1961), the problem of *estimating n normal*

Address for correspondence: Jiashun Jin, Department of Statistics, Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
E-mail: jiashun@stat.cmu.edu

 1369–7412/08/70461

means has been extensively studied and well understood. Many modern procedures, e.g. wavelet thresholding procedures (Donoho and Johnstone, 1994) and the lasso (Tibshirani, 1996), are intellectually connected to the normal means problem. In these studies, the interest is more or less focused on the regime of relatively strong signals, and the small proportion of relatively large signals plays the key role.

Recently, there has been a resurgence of research interest in the field of large-scale multiple testing. The impetus is the need for sophisticated and implementable statistical tools to solve application problems in many scientific areas, e.g. genomics, astronomy, functional magnetic resonance imaging and image processing. In this field, a problem of major interest is *testing n normal means*: to determine simultaneously which means are 0 and which are not. In this context, the collection of 'moderately strong' or 'faint' signals plays the key role.

In the past few years, interest in the regime of faint signals has been steadily growing and many seemingly intractable problems have seen encouraging developments. The following three interconnected questions are of particular interest.

(a) Overall testing: is there any signal at all?
(b) Estimating the proportion: how many signals are there?
(c) Simultaneous testing: which are signals and which are noise?

The first question has been studied in Ingster (1997, 1999) and Donoho and Jin (2004). The third question has been studied in Benjamini and Hochberg (1995), Efron *et al.* (2001), Storey (2002, 2007), Efron (2004) and Genovese and Wasserman (2004). In this paper, we concentrate on the second question, i.e. estimating the proportion of signals or, equivalently, the proportion of non-zero means.

1.1. Estimating the proportion of non-zero means: motivations

Denote the mean vector by $\mu = (\mu_1, \ldots, \mu_n)$. We are interested in estimating the proportion of non-zero means:

$$\varepsilon_n = \varepsilon_n(\mu) = \frac{1}{n} \#\{j : \mu_j \neq 0\}. \tag{1.2}$$

Such a situation can be found in the following application examples.

1.1.1. Analysis of microarray data on breast cancer

In this example, on the basis of 15 patients who were diagnosed with breast cancer (seven with the BRCA1 mutation and eight with the BRCA2 mutation), microarray data were generated for the same set of 3226 genes. In this setting, the proportion of differentially expressed genes is of interest (Efron, 2004; Jin and Cai, 2007; Storey, 2007). For each gene, a p-value was first computed by using a two-sample t-test and then converted to a z-score. The z-scores can be modelled as $X_j \sim N(\mu_j, \sigma_0^2)$, where $\mu_j = \mu_0$ if and only if the corresponding gene is not differentially expressed, and μ_0 and σ_0 are called *null parameters* (Efron, 2004). After the null parameters have been estimated and the z-scores have been renormalized, the problem of estimating the proportion of differentially expressed genes reduces to the problem of estimating the proportion of non-zero normal means. The z-scores were kindly provided by Bradley Efron and can be downloaded from `http://www.stat.purdue.edu/~jinj/Research/software`. See section 5 of Jin and Cai (2007) and Efron (2004) for the assumptions on normality and homoscedasticity. Also, the assumption on independence would not be a serious issue in this example. The reason is that, although the main results of this paper are developed under the assumption of independence, they can be naturally generalized to handle many weakly dependent cases. See Section 7 for more discussion.

1.1.2. Kuiper Belt object

The Kuiper Belt refers to the region in the solar system that is beyond the orbit of Neptune. The Kuiper Belt contains a large unknown number of small objects (i.e. Kuiper Belt objects (KBOs)). The Taiwanese–American occultation survey is a recent project that studies the abundance of KBOs. In this project, one manipulates a very large number (10^{11}–10^{12}) of tests, but out of which only a small proportion is relevant to the KBOs. A major interest in this project is to estimate the proportion of tests that contains a KBO. Similarly, by first obtaining a p-value for each test and then converting it to a z-score, the resulting test statistics can be approximately modelled as normal random variables $X_j \sim N(\mu_j, 1)$, $j = 1, \ldots, n$, where $\mu_j \neq 0$ if and only if the jth test contains a KBO. In this example, X_j can be treated as independent. See Meinshausen and Rice (2006) for more details.

In addition to the above application examples, the proportion is also of interest for the following reason: the implementation of many recent procedures needs a reasonable estimate of the proportion. Among these procedures are the local false discovery rate (FDR) approach (Efron *et al.*, 2001), the B-statistic (Lönnstedt and Speed, 2002), the optimal discovery approach (Storey, 2007) and the adaptive FDR approach (Benjamini *et al.*, 2005). Hopefully, if a good estimate of the proportion is available, some of these procedures could be improved. See Section 3 for more discussion.

Estimating the proportion has long been known as a difficult problem. There have been some interesting developments recently, e.g. an approach by Meinshausen and Rice (2006) (see also Efron *et al.* (2001), Genovese and Wasserman (2004), Meinshausen and Bühlmann (2005) and Schweder and Spjotovoll (1982)), and an approach by Swanepoel (1999). Roughly, say, these approaches are only successful under a condition which Genovese and Wasserman (2004) called the 'purity'; see Section 4 for details. Unfortunately, the purity condition is difficult to check in practice and is also relatively stringent (see lemma 2). This motivates us to develop a different approach to estimating the proportion.

In this paper, we concentrate on the problem of estimating the proportion of non-zero normal means (see Cai *et al.* (2007), Jin and Cai (2007) and Jin *et al.* (2007) for related studies on other settings). We now begin by shedding some light on what could be an appropriate approach for this problem.

1.2. Ideas and preliminary oracle bounds for the proportion

A successful estimator needs to capture the essential features of the estimand. It seems that one of the unique features of the proportion is its *invariance to scaling*. For illustration, consider a scenario in which we can manipulate the data set by amplifying every signal component (i.e. μ_j) by an arbitrary non-zero factor but keeping the corresponding noise component (i.e. z_j) untouched. In this scenario, the proportion of non-zero means remains the same, although the data set has been dramatically changed. We call this the property of *scaling invariance*: the proportion of non-zero means remains the same if we multiply each entry of the mean vector by an arbitrary non-zero constant individually.

Unfortunately, the approaches that were introduced in Meinshausen and Rice (2006) and Swanepoel (1999) (and also those in Efron *et al.* (2001), Genovese and Wasserman (2004), Meinshausen and Bühlmann (2005) and Schweder and Spjotovoll (1982)) are based on the data tail or extreme values. Intuitively, as the data tail is not scaling invariant, these approaches are only successful for special cases, so we need to find somewhere other than the data tail to construct the estimators. Surprisingly, the right place to build scaling invariant statistics is *not* the spatial domain, but the frequency domain (Mallat, 1998). Consequently, we should use tools

that are based on Fourier transform coefficients, instead of moments or the data tail, for the estimation.

For illustration, suppose that out of n normal means a proportion of ε_0 has a common positive mean μ_0, and all others have mean 0. Denote $\mathrm{i} = \sqrt{-1}$ and introduce

$$\frac{1}{n}\sum_{j=1}^{n}\exp(\mathrm{i}tX_j)$$

which we call the *empirical characteristic function*. If we neglect stochastic fluctuations, the empirical characteristic function reduces to its own mean, which we call the *underlying characteristic function*. The underlying characteristic function is seen to be $\exp(-t^2/2)[1-\varepsilon_0\{1-\exp(\mathrm{i}t\mu_0)\}]$, which naturally factors into two components: the *amplitude* $A(t) = A(t;\varepsilon_0,\mu_0) \equiv \exp(-t^2/2)$ and the *phase* $\varphi(t) = \varphi(t;\varepsilon_0,\mu_0) \equiv 1-\varepsilon_0\{1-\exp(\mathrm{i}t\mu_0)\}$. Note that only the phase contains relevant information on ε_0, with μ_0 playing the role of a nuisance parameter. A convenient way to remove the nuisance parameter is to maximize the phase over all frequencies:

$$\frac{1}{2}\sup_t|\varphi(t;\varepsilon_0,\mu_0)-1| = \frac{\varepsilon_0}{2}\sup_t\{|1-\exp(\mathrm{i}t\mu_0)|\} \equiv \varepsilon_0,$$

which immediately gives a desired estimate of the proportion.

Inspired by this example, we introduce the *empirical phase function* and the *underlying phase function* (or *phase function* for short) for general normal means settings and denote them by $\varphi_n(t)$ and $\varphi(t)$ respectively:

$$\varphi_n(t) = \varphi_n(t;X_1,\ldots,X_n,n) = \frac{1}{n}\sum_{j=1}^{n}\left\{1-\exp\left(\frac{t^2}{2}\right)\cos(tX_j)\right\}, \tag{1.3}$$

$$\varphi(t) = \varphi(t;\mu,n) = \frac{1}{n}\sum_{j=1}^{n}\{1-\cos(t\mu_j)\}. \tag{1.4}$$

Here we use only the real parts of the phase functions. Because we shall see soon that the real parts alone yield desired estimators for the proportion, we drop the imaginary parts everywhere for simplicity. We call equations (1.3) and (1.4) the *cosinusoid construction for phase functions*. This construction conveniently yields oracle upper and lower bounds for the true proportion.

Theorem 1. With $\varepsilon_n(\mu)$ defined in equation (1.2) and $\varphi(t;\mu,n)$ defined in equation (1.4), for any μ and $n \geqslant 1$, we have

$$\frac{1}{2}\sup_{\{t\}}\{\varphi(t;\mu,n)\} \leqslant \varepsilon_n(\mu) \leqslant \sup_{\{t\}}\{\varphi(t;\mu,n)\}.$$

Theorem 1 is proved in Appendix A. We call the bounds 'oracle' bounds because they depend on the phase function, instead of on the data directly. However, replacing the phase function by the empirical phase function naturally yields data-driven bounds. We shall return to this point in Section 2.2.

Though the bounds hold for all mean vectors and are convenient to use, they are not tight, so they do not immediately yield consistent estimators. However, the result suggests that we are on the right track. In the next section, we show that, with careful refinements, the cosinusoid construction indeed yields an oracle equivalence of the proportion, that equals the true proportion for all n and all mean vectors μ (hence the terminology of *universal oracle equivalence*). In addition, the oracle equivalence naturally yields consistent estimators by replacing the phase function with its empirical counterpart—the empirical phase function.

1.3. Content of this paper

This paper is organized as follows. In Section 2, we first introduce an approach to constructing oracle equivalences for the proportion. We then use the oracle equivalence to construct a family of real estimators and show that the estimators are uniformly consistent for the true proportion for a wide class of parameters. We also introduce an approach for controlling the standard deviations of the estimators; the approach is especially useful in practice. We conclude the section by discussing some related work on estimating the proportion. Section 3 extends the results in Section 2 to a hierarchical model, which can be viewed as the Bayesian variant of model (1.1). This section also discusses the connection of our approach with the FDR approach of Benjamini and Hochberg (1995), as well as with several other recent approaches in large-scale multiple testing. Section 4 compares the approach proposed with that of Meinshausen and Rice (2006). Section 5 describes some simulation experiments. Section 6 extends the approach to estimating other functionals concerning the normal means, including the proportion of normal means that exceeds a given threshold and the average l^p-norm of the normal means vector. Section 7 discusses extensions to non-Gaussian data as well as data with dependent structures. Some concluding remarks are also made in this section. Appendix A contains proofs of the main theorems and corollaries in this paper. Proofs for theorems 12 and 13 and all lemmas have been omitted in this paper but can be found in sections 8 and 9 of Jin (2007).

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

```
http://www.blackwellpublishing.com/rss
```

2. Main results

In this section, we first introduce an approach to constructing oracle equivalence of the proportion. We then use the oracle equivalence to construct real estimators, study their consistency and discuss how to control their variations. Last, we comment on the connections between this paper and Cai *et al.* (2007), Jin and Cai (2007) and Jin *et al*. (2007).

2.1. The oracle equivalence of the proportion

We now develop the ideas that were introduced in Section 1 to construct oracle equivalences of the proportion. Consider the phase function in the form

$$\varphi(t) = \varphi(t;\mu,n) = \frac{1}{n}\sum_{j=1}^{n}\{1-\psi(\mu_j;t)\}, \tag{2.1}$$

where ψ is a function that we hope to construct such that,

(a) for any t, $\psi(0;t)=1$,
(b) for any fixed $u\neq 0$, $\lim_{t\to\infty}\{\psi(u;t)\}=0$ and
(c) $\psi(u;t)\geqslant 0$ for all u and t.

To distinguish from μ, we use u to denote a scalar quantity here.

In fact, for all fixed μ and n, it can be shown that, if both (a) and (b) are satisfied, then $\varepsilon_n(\mu)=\inf_{\{s>0\}}[\sup_{\{|t|>s\}}\{\varphi(t;\mu,n)\}]$, and the right-hand side provides an oracle equivalence of $\varepsilon_n(\mu)$. Moreover, if (c) is also satisfied, then the right-hand side has a simpler form and $\varepsilon_n(\mu)=\sup_{\{t\}}\{\varphi(t;\mu,n)\}$.

The intuition behind the construction is that, together, (a) and (b) ensure that the individual index function $\mathbf{1}_{\{\mu_j\neq 0\}}$ is well approximated by $1-\psi(\mu_j;t)$ with large t. Since the proportion is

the average of all individual index functions, it is then well approximated by the phase function, which is nothing other than the average of all individual functions $1-\psi(\mu_j;t)$.

We now describe the construction of ψ. Note that the cosinusoid construction $\psi(u;t)=\cos(ut)$ clearly does not satisfy condition (b), as the cosinusoid does not damp to 0 pointwisely. However, on a second thought, we note here that, though the cosinusoid does not damp to 0 pointwisely, it does damp to 0 'on average', according to the well-known Riemann–Lebesgue theorem (e.g. Mallat (1998), page 40).

Theorem 2 (Riemann–Lebesgue). If $\omega\in L^1(\mathbf{R})$, then $\lim_{t\to\infty}\{\int_{-\infty}^{\infty}\omega(\xi)\cos(t\xi)\,\mathrm{d}\xi\}=0$.

Inspired by this, we employ the Bayesian point of view and model the frequency t as random. As a result, the expected value of $\cos(ut)$ becomes the average of cosinusoids across different frequencies and is no longer tied to the cosinusoid pointwisely.

To elaborate, we choose a random variable Ξ on $(-1,1)$ that has a symmetric, bounded and continuous density function $\omega(\xi)$. Let

$$\psi(u;t)=E[\cos(u\Xi t)]=\int_{-1}^{1}\omega(\xi)\cos(u\xi t)\,\mathrm{d}\xi,\qquad\forall t,u. \tag{2.2}$$

By the Riemann–Lebesgue theorem, this construction satisfies both conditions (a) and (b). We point out that $\omega(\xi)$ does not have to be a density function, or continuous or bounded; we assume so only for convenience.

Next, we discuss under what conditions (c) holds. To do so, we introduce the following definitions.

Definition 1. We call a function f over $(0,1)$ *superadditive* if $f(\xi_1)+f(\xi_2)\leqslant f(\xi_1+\xi_2)$ for any $0<\xi_1,\xi_2<1$ and $\xi_1+\xi_2<1$. We call a density function ω over $(-1,1)$ *eligible* if it is symmetric, bounded and continuous. We call ω *good* if additionally $\omega(\xi)=g(1-\xi)$ for some convex and superadditive function g over (0,1).

It is proved in lemma 3 that condition (c) is satisfied if ω is good.

Finally, the only unfinished step is to find an empirical phase function that naturally connects to the phase function in equation (2.1) by taking the expectation. Comparing with equations (1.3) and (1.4), we define the empirical phase function by

$$\varphi_n(t;X_1,\ldots,X_n,n)=\frac{1}{n}\sum_{j=1}^{n}\{1-\kappa(X_j;t)\}, \tag{2.3}$$

where

$$\kappa(x,t)=\int_{-1}^{1}\omega(\xi)\exp(t^2\xi^2/2)\cos(t\xi x)\,\mathrm{d}\xi. \tag{2.4}$$

It is proved in lemma 3 that, when $X\sim N(u,1)$,

$$E[\kappa(X;t)]=\int_{-1}^{1}\omega(\xi)\cos(ut\xi)\,\mathrm{d}\xi\equiv\psi(u;t),$$

so κ naturally connects to ψ by taking the expectation. As a result, φ_n connects back to the phase function φ also by taking the expectation: $E[\varphi_n(t;X_1,\ldots,X_n,n)]=\varphi(t;\mu,n)$. This completes the construction.

We now reveal the intuition behind the construction of κ. Since the frequency t plays the role of a scaling parameter, we illustrate with $t=1$ and write $\psi(u)=\psi(u;1)$ and $\kappa(x)=\kappa(x;1)$ for short. Note that $\psi=\hat{\omega}$ and that $E[\kappa]=\kappa*\phi$; here ϕ is the density function of $N(0,1)$, the asterisk

denotes the usual convolution and $\hat{\omega}$ is the Fourier transform of ω. Under mild conditions, $\kappa * \phi = \psi$ is equivalent to $\hat{\kappa}\hat{\phi} = \hat{\psi} \equiv \omega$, so κ should be the inverse Fourier transform of $\omega/\hat{\phi}$, which is exactly the same as that in equation (2.4). We mention here that, although it seems that the choice of ω could be arbitrary, it is important to choose ω so that $\omega/\hat{\phi}$ is integrable, and its inverse Fourier transform exists. A convenient sufficient condition is that ω has a compact support.

The construction above indeed yields a family of oracle equivalences. The following theorem is proved in Appendix A.

Theorem 3. Fix n and $\mu \in \mathbf{R}^n$; let φ and ψ be defined as in equations (2.1) and (2.2) respectively. If ω is eligible, then $\varepsilon_n(\mu) = \inf_{\{s \geqslant 0\}}\{\sup_{\{|t|>s\}} \varphi(t;\mu,n)\}$. If additionally ω is good, then $\varepsilon_n(\mu) = \sup_{\{t\}}\{\varphi(t;\mu,n)\}$.

We conclude this section by giving some examples of ω.

2.1.1. Example A (triangle family)
$\omega(\xi) = \{(\alpha+1)/2\}\{(1-|\xi|)^+\}^\alpha$. When $\alpha = 1$, $\omega(\xi)$ is the well-known triangle density function: hence the name *triangle family*. Clearly, ω is eligible for all $\alpha > 0$ and is good for all $\alpha \geqslant 1$. Moreover,

$$\kappa(x;t) = (\alpha+1)\int_0^1 (1-\xi)^\alpha \exp(t^2\xi^2/2)\cos(tx\xi)\,\mathrm{d}\xi,$$

and

$$\psi(u;t) = (\alpha+1)\int_0^1 (1-\xi)^\alpha \cos(tu\xi)\,\mathrm{d}\xi.$$

In particular, $\psi(u;t) = 2\{1-\cos(tu)\}/(ut)^2$ when $\alpha = 1$, and $\psi(u;t) = 6\{ut - \sin(ut)\}/(ut)^3$ when $\alpha = 2$; here, the values of $\psi(0;t)$ are set to $\lim_{u\to 0}\{\psi(u,t)\}$. Fig. 1 shows the plot of $1-\psi(u;t)$ with $\alpha = 1, 2$. The plot illustrates that, for a moderately large t, the index function $\mathbf{1}_{\{u\neq 0\}}$ can be well approximated by $1-\psi(u;t)$.

2.1.2. Example B (uniform)
$\omega(\xi)$ is the uniform density function over $(-1, 1)$. In this example,

$$\kappa(x;t) = \int_0^1 \exp(t^2\xi^2/2)\cos(tx\xi)\,\mathrm{d}\xi,$$

and $\psi(u;t) = \sin(ut)/ut$. Similarly, the value of $\psi(0;t)$ is set to $\lim_{u\to 0}\{\psi(u;t)\}$. Note here that ω is eligible but is not good.

2.1.3. Example C (smooth)
$\omega(\xi) = c_0 \exp\{-1/(1-\xi^2)\}$ when $|\xi| < 1$ and $\omega(\xi) = 0$ otherwise. The coefficient $c_0 = \{\int_{-1}^1 \omega(\xi)\,\mathrm{d}\xi\}^{-1}$. Note that $\omega(\xi)$ is smooth over $(-\infty, \infty)$.

Interestingly, the cosinusoid construction (1.4) can also be thought of as a special case of our construction in equation (2.1), where the random variable Ξ does not have a density function. Instead, Ξ concentrates its mass equally on two points: 1 and -1.

2.2. Uniformly consistent estimation

We now consider empirical estimates of the proportion. The idea is to use the empirical phase function as the estimate, and to hope to choose an appropriate t such that

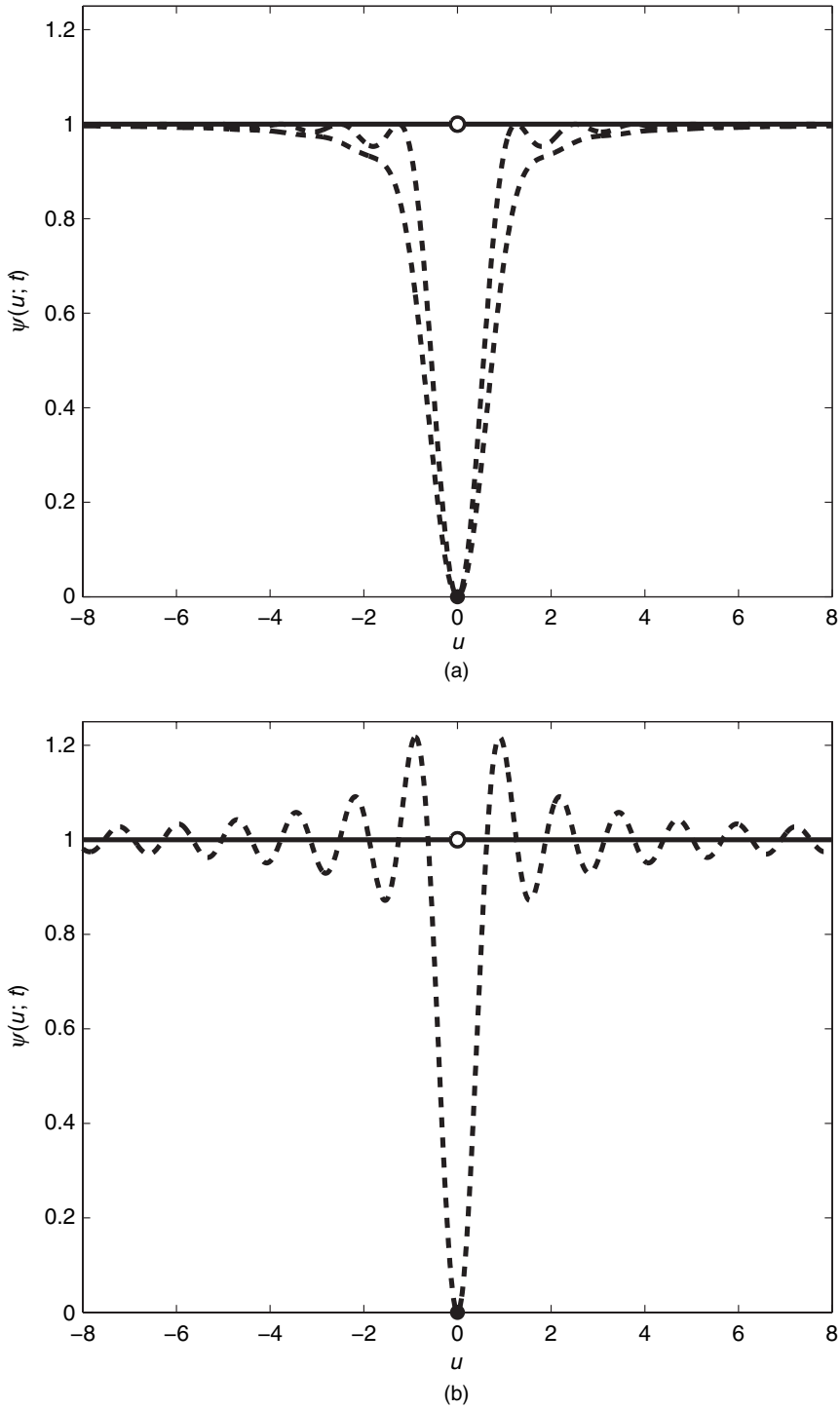

Fig. 1. Function $1 - \psi(u; t)$ with $t = 5$ (in both parts, together, the horizontal line (excluding the point at (0,1)) and the dot at the origin stand for the index function $\mathbf{1}_{\{u \neq 0\}}$): (a) densities of the triangle family with $\alpha = 1$ (upper) and $\alpha = 2$ (lower); (b) uniform density

$$\varphi_n(t; X_1, \ldots, X_n, n)/\varepsilon_n(\mu) \approx \varphi(t; \mu, n)/\varepsilon_n(\mu) \approx 1. \tag{2.5}$$

There is a trade-off in the choice of t. When t increases from 0 to ∞, the second approximation becomes increasingly accurate, but at the same time the variance of φ_n increases, so the first approximation becomes increasingly unstable. It turns out that the right choice of t is in the range of $O\{\sqrt{\log(n)}\}$ so, for convenience, we consider the family of estimators $\varphi_n\{t_n(\gamma); X_1, \ldots, X_n, n\}$, where

$$t_n(\gamma) = \sqrt{\{2\gamma \log(n)\}}, \qquad 0 < \gamma \leqslant \tfrac{1}{2}. \tag{2.6}$$

We now discuss when the approximations in expression (2.5) are accurate. Consider the second approximation first. For the approximation to be accurate, it is sufficient that

$$\operatorname*{Ave}_{\{j:\mu_j \neq 0\}} (\psi[\mu_j; \sqrt{\{2\gamma \log(n)\}}]) = o(1), \qquad \text{as } n \to \infty. \tag{2.7}$$

Recall that $\psi(u; t) \to 0$ whenever $|u|t \to \infty$; a sufficient condition for expression (2.7) is

$$\min_{\{j:\mu_j \neq 0\}} \{|\mu_j|\} \geqslant \frac{\log\{\log(n)\}}{\sqrt{\{2\log(n)\}}}. \tag{2.8}$$

In comparison, condition (2.8) is stronger than condition (2.7). However, for simplicity in the presentation, we use condition (2.8) below.

We now discuss when the first approximation in expression (2.5) is accurate. Here, two crucial factors are the magnitude of the stochastic fluctuation of φ_n and the magnitude of the true proportion. For the approximation to be accurate, it is necessary that the former is smaller than the latter. It turns out that the stochastic fluctuation of φ_n is of the order of $n^{\gamma - 1/2}$, on the basis of the following theorem which is proved in Appendix A.

Theorem 4. Let $\varphi(t; \mu, n)$ and $\varphi_n(t; X_1, \ldots, X_n, n)$ be constructed as in equations (2.1) and (2.3) with an eligible density ω. When $n \to \infty$, for any fixed $q > 3/2$ and $0 < \gamma \leqslant \frac{1}{2}$, there is a constant $C = C(r, q, \gamma, \omega)$ such that, except for an event having probability distributed as $2\log(n)^2 n^{-2q/3}$,

$$\sup_{\{\mu \in B_n^1(r)\}} \sup_{\{0 \leqslant t \leqslant \sqrt{\{2\gamma \log(n)\}}\}} |\varphi_n(t; X_1, \ldots, X_n, n) - \varphi(t; \mu, n)| \leqslant C \log(n)^{-1/2} n^{\gamma - 1/2},$$

where $B_n^1(r) = \{\mu \in \mathbf{R}^n : (1/n)\Sigma_{j=1}^n |\mu_j| \leqslant r\}$ is the l^1-ball in $\mathbf{R}^n$ with radius $r > 0$.

The accuracy of the first approximation in expression (2.5) now depends on the magnitude of the true proportion. In the literature, the magnitude of the proportion is modelled through the concept of *sparsity* (e.g. Abramovich *et al.* (2006)). We list three different regimes of sparsity.

(a) *Relatively dense regime*: the proportion is small (e.g. $\varepsilon_n = 10\%$) but does not tend to 0 as $n \to \infty$. See Efron *et al.* (2001) and Genovese and Wasserman (2004).
(b) *Moderately sparse regime*: the proportion tends to 0 as $n \to \infty$ but does so slower than $1/n^{1/2}$ does, e.g. $\varepsilon_n = n^{-\beta}$ with $0 < \beta < \frac{1}{2}$. See for example section 3.1 of Meinshausen and Rice (2006).
(c) *Very sparse regime*: the proportion tends to 0 faster than $1/n^{1/2}$ does, e.g. $\varepsilon_n = n^{-\beta}$ with $\frac{1}{2} < \beta < 1$. This is the most challenging case, with very few known results; see Donoho and Jin (2004), Abramovich *et al.* (2006), Meinshausen and Rice (2006) and Cai *et al.* (2007).

Now, so that the first approximation in expression (2.5) is accurate, it is necessary that the

situation is either relatively dense or moderately sparse, but not very sparse (see Section 2.4 for more discussion on the very sparse case). More precisely, it is necessary that

$$\varepsilon_n(\mu) \geqslant n^{\gamma-1/2}. \tag{2.9}$$

In summary, together, conditions (2.8) and (2.9) give a sufficient condition for the consistency of the estimators proposed. Inspired by this, we introduce the following set of parameters:

$$\Theta_n(\gamma, r) = \left\{ \mu \in B_n^1(r),\ \min_{\{j:\mu_j \neq 0\}} \{|\mu_j|\} \geqslant \frac{\log\{\log(n)\}}{\sqrt{\{2\log(n)\}}}, \varepsilon_n(\mu) \geqslant n^{\gamma-1/2} \right\}, \qquad r > 0. \tag{2.10}$$

It turns out that, as stated in the following theorem, the estimators proposed are uniformly consistent for all parameters in $\Theta_n(\gamma, r)$.

Theorem 5. Let $\Theta_n(\gamma, r)$ be defined as in expression (2.10) and $\varphi_n(t; X_1, \ldots, X_n, n)$ be defined as in equation (2.3) where the density ω is eligible. When $n \to \infty$, for any fixed $0 < \gamma \leqslant \frac{1}{2}$, except for an event with algebraically small probability,

$$\lim_{n\to\infty} \left(\sup_{\{\Theta_n(\gamma,r)\}} \left| \frac{\varphi_n[\sqrt{\{2\gamma\log(n)\}}; X_1, \ldots, X_n, n]}{\varepsilon_n(\mu)} - 1 \right| \right) = 0.$$

Here, we say that a probability is algebraically small if it is bounded by Cn^{-a} for some constants $C = C(\gamma, r) > 0$ and $a = a(\gamma, r) > 0$. Theorem 5 is proved in Appendix A. We mention that theorem 5 is closely related to theorem 5 of Jin and Cai (2007). In fact, if we take ω to be the triangle density, then theorem 5 can be thought of as a special case of theorem 5 in Jin and Cai (2007).

Additionally, if ω is a good density, then $\sup_{\{0 \leqslant t \leqslant \sqrt{\{2\gamma\log(n)\}}\}} \{\varphi_n(t; X_1, \ldots, X_n, n)\}$ are also consistent. This is the following corollary, which is proved in Appendix A.

Corollary 1. Let $\Theta_n(\gamma, r)$ be defined as in expression (2.10), and $\varphi_n(t; X_1, \ldots, X_n, n)$ be defined as in equation (2.3) with a good ω. When $n \to \infty$, for any $0 < \gamma \leqslant \frac{1}{2}$, except for an event with algebraically small probability,

$$\lim_{n\to\infty} \left[\sup_{\{\Theta_n(\gamma,r)\}} \left| \frac{\sup_{\{0 \leqslant t \leqslant \sqrt{\{2\gamma\log(n)\}}\}} \varphi_n(t; X_1, \ldots, X_n, n)}{\varepsilon_n(\mu)} - 1 \right| \right] = 0.$$

2.3. Adaptive control on the standard deviations of the estimators

In practice, it is of interest to know how to select the 'best' t for a given data set and a given ω. To do so, a useful strategy is to preselect a tolerance parameter α_n, and to pick the largest t such that the standard deviation of the estimator is no larger than α_n (recall that, the larger the t, the smaller the bias and the larger the variance). In this section, we introduce an approach to realize this strategy. The approach is adaptive for different n and ω; and, as a bonus, the resulting t is non-random and can be conveniently calculated.

The approach is based on the following simple observation: for any fixed $t > 0$ and $z \sim N(0, 1)$, the second moment of $\kappa(u + z; t)$, as a function of u, reaches its maximum at $u = 0$. This leads to the following lemma, which is proved in section 9 of Jin (2007).

Lemma 1. Fix $u, t > 0$ and $z \sim N(0, 1)$, $E[\kappa(z + u; t)]^2 \leqslant E[\kappa(z; t)]^2$. As a result, with $\varphi(t; \mu, n)$ defined as in equation (2.1) and ω being an eligible density function, for any fixed n and μ,

$$\frac{1 - \varepsilon_n(\mu)}{n} \mathrm{var}\{\kappa(z; t)\} \leqslant \mathrm{var}\{\varphi_n(t; X_1, \ldots, X_n, n)\} \leqslant \frac{1}{n}[\mathrm{var}\{\kappa(z; t)\} + 1].$$

In many applications, $\mathrm{var}\{\kappa(z;t)\} \gg 1$ for t in the range of interest. So the lower bound differs from the upper bound only by a factor of $1-\varepsilon_n(\mu)$. In particular, for the sparse case where $\varepsilon_n(\mu) \approx 0$, the bounds are tight.

By lemma 1, the variance of $\varphi_n(t)$ is no greater than $(1/n)[\mathrm{var}\{\kappa(z;t)\}+1]$, which can be conveniently calculated once ω is given. Consequently, if we set $t = t_n^*(\alpha_n;\omega)$, where

$$t_n^*(\alpha_n;\omega) = \max\left(t : \frac{1}{n}[\mathrm{var}\{\kappa(z;t)\}+1] \leqslant \alpha_n^2\right), \tag{2.11}$$

then the standard deviation of $\varphi_n(t)$ is no greater than α_n. This is the following theorem, which follows directly from lemma 1.

Theorem 6. Let $\varphi_n(t; X_1, \ldots, X_n, n)$ be defined as in equation (2.3) and $t_n^*(\alpha_n;\omega)$ be defined as in equation (2.11), where ω is eligible. We have $\sup_{\{\mu\in\mathbf{R}^n\}}(\mathrm{var}[\varphi_n\{t_n^*(\alpha_n;\omega); X_1, \ldots, X_n\}]) \leqslant \alpha_n^2$.

We note here that $t_n^*(\alpha_n;\omega)$ is non-random and can be conveniently calculated. We have tabulated the standard deviations of $\kappa(z;t)$ for t in the range 1–5, and for ω being the uniform, triangle or smooth density as introduced in Section 2.1. The table can be downloaded from `www.stat.purdue.edu/~jinj/Research/software/PropOracle`. Using the table, the values of $t_n^*(\alpha_n;\omega)$ can be easily obtained for a wide range of α_n.

Next, note that, the faster that $\alpha_n \to 0$, the slower that $t_n^*(\alpha_n;\omega) \to \infty$, and the larger the bias. So, to ensure the consistency of $\varphi_n\{t_n^*(\alpha_n;\omega); X_1, \ldots, X_n\}$, a necessary condition is that $\alpha_n \to 0$ *sufficiently slowly*. For example, to ensure the uniform consistency for all $\mu \in \Theta_n(r,\gamma)$, we need that $\alpha_n \to 0$ sufficiently slowly that $t_n^*(\alpha_n;\omega) \geqslant c_0\sqrt{\log(n)}$ for some constant $c_0 > 0$. In practice, since the value of $t_n^*(\alpha_n;\omega)$ is non-random and is convenient to obtain, the condition $t_n^*(\alpha_n;\omega) \geqslant c_0\sqrt{\log(n)}$ can be checked before we implement the procedure. The proof of the following theorem is similar to that of theorem 5 and so has been omitted.

Theorem 7. Fix a constant $c_0 > 0$; let $\varphi_n(t; X_1, \ldots, X_n, n)$ be defined as in equation (2.3) and $t_n^*(\alpha_n;\omega)$ be defined as in equation (2.11), where ω is eligible. When $n \to \infty$, if $\alpha_n \to 0$ sufficiently slowly such that $t_n^*(\alpha_n;\omega) \geqslant c_0\sqrt{\log(n)}$, then, uniformly for all $\mu \in \Theta_n(\gamma, r)$, $\varphi_n\{t_n^*(\alpha_n;\omega), X_1, \ldots, X_n\}/\varepsilon_n(\mu)$ converges to 1 in probability.

We mention that the main contribution of the adaptive procedure is that it offers a non-asymptotic approach for controlling the standard deviations of the estimators and consequently provides a useful guideline for choosing t. Simulations show that the control on the standard deviations is usually tight; see Section 5 for more discussion.

2.4. Recent work on estimating the proportion

We briefly review some closely related work that we have done. Part of the work concerns the generalization to heteroscedastic Gaussian models, part of it concerns the very sparse case and part of it concerns the situation that, in model (1.1), the variances of X_j are unknown.

First, we discuss the generalization to heteroscedastic Gaussian models. In this setting, we model each X_j as a normal random variable with individual mean μ_j and variance σ_j^2. In the terminology of multiple testing, we assume that $(\mu_j, \sigma_j) = (0,1)$ if X_j is a *null effect* and $(\mu_j, \sigma_j) \neq (0,1)$ if X_j is a *non-null effect*. The proportion of signals is then the proportion of non-null effects: $\varepsilon_n = \#\{j : (\mu_j, \sigma_j) \neq (0,1)\}/n$. Clearly, this is an extension of the setting of n normal means and is a more realistic model for applications. In Jin and Cai (2007) and Jin *et al.* (2007), we have successfully extended the ideas in previous sections to construct a new family of estimators. We show that, under mild identifiability conditions, these estimators are uniformly

consistent for the proportion. We have also implemented these estimators in the analysis of microarray data on breast cancer (Efron, 2004; Jin and Cai, 2007) and the analysis of comparative genomic hybridization data on lung cancer (Jin *et al.*, 2007). The new approaches compare favourably with existing approaches both in theory and in applications (Jin and Cai, 2007; Jin *et al.*, 2007).

Next, we discuss the very sparse case. Since the estimators that were proposed in previous sections generally have a standard deviation no less than $1/n^{1/2}$, we should not expect them to be consistent in the very sparse case, where the true proportion is much smaller than $1/n^{1/2}$. The subtlety of the sparse case has been addressed in detail in Ingster (1997, 1999), Donoho and Jin (2004) and Cai *et al.* (2007). It is surprising that the proportion may not be estimable even when all non-zero μ_js tend to ∞ uniformly. In fact, Donoho and Jin (2004) considered a setting where X_j are modelled as samples from a two-component Gaussian mixture $(1-\varepsilon_n)\,N(0,1)+\varepsilon_n\,N(\mu_n,1)$, where $\varepsilon_n = n^{-\beta}$ with $\beta \in (\frac{1}{2}, 1)$, and $\mu_n = \sqrt{\{2r\log(n)\}}$ with $0 < r < 1$. Clearly, this is a very sparse case. Define a function $\rho^*(\beta)$ which equals $\beta - \frac{1}{2}$ when $\beta \in (\frac{1}{2}, \frac{3}{4}]$ and equals $\{1-\sqrt{(1-\beta)}\}^2$ when $\beta \in (\frac{3}{4}, 1)$. It was shown in Donoho and Jin (2004) (see also Ingster (1997, 1999)) that, if $r < \rho^*(\beta)$, then no test could reliably tell whether ε_n equals 0 or not (i.e. any test would have a sum of type I and type II errors that tends to 1). Consequently, no estimator could be consistent for the proportion. This shows that, in the very sparse case, the proportion may not be estimable even when all signals tend to ∞.

Although the very sparse case is much more challenging than the relatively dense case and the moderately sparse case, interesting progress is still possible. Cai *et al.* (2007) developed a family of new estimators called the *Cai–Jin–Low lower bounds*. At any specified level $\alpha \in (0, 1)$, the Cai–Jin–Low lower bound provides an honest confidence lower bound for the proportion, which holds uniformly for all one-sided Gaussian shift mixtures. Additionally, when applied to the two-component Gaussian mixture model above, the lower bound is also optimal in two senses: it is consistent for the true proportion whenever consistent estimators exist, and it obtains the suboptimal rate of convergence. Interesting progress was also made in Meinshausen and Rice (2006).

Last, we discuss the case of unknown variance. A direct generalization of model (1.1) is that we assume that X_j have a common *unknown* variance σ^2. This setting can be viewed as a special case of that studied in Section 3 of Jin and Cai (2007) if we set μ_0 to 0 and assume homoscedasticity; see details therein. We remark that, although theorem 6 of Jin and Cai (2007) has been focused on the special case where ω is the triangle density, it can be generalized to handle the cases where ω is only assumed to be eligible. For brevity, we skip further discussion.

3. Bayesian hierarchical model

In this section, we extend the results in Section 2.2 to the Gaussian hierarchical model. We also use the hierarchical model to discuss the connections of the proposed procedures to some recent procedures in large-scale multiple testing.

The Gaussian hierarchical model (e.g. Genovese and Wasserman (2004)) is the Bayesian variant of model (1.1). It can be thought of as follows. Pick $\varepsilon \in (0, 1)$ and a marginal cumulative distribution function (CDF) F with no mass at 0. For each $j = 1, \ldots, n$, we flip a coin with probability ε of landing heads. When the coin lands tails, we draw an observation X_j from $N(0, 1)$. When the coin lands heads, we draw an observation μ_j from F and then an observation X_j from $N(\mu_j, 1)$. As a result, the marginal density of X_j can be written as a mixture of two components, one being the standard normal and the other being a Gaussian shift mixture where F is the mixing CDF:

$$(1-\varepsilon)\,\phi(x)+\varepsilon\int\phi(x-u)\,\mathrm{d}F(u). \tag{3.1}$$

Here ϕ is the density function of $N(0,1)$; ε can be thought of as the proportion of non-zero normal means. We assume that $P_F\{u\neq 0\}=1$.

We now extend the results in Section 2.2 to model (3.1). First, we discuss the moderately sparse case by calibrating ε_n with $\varepsilon_n=n^{-\beta}$. The following theorem is proved in Appendix A.

Theorem 8. Fix F, $0<\beta<\frac{1}{2}$ and $0<\gamma\leqslant\frac{1}{2}-\beta$, and let ε_n be $n^{-\beta}$ and $\varphi_n(t;X_1,\ldots,X_n,n)$ be defined as in equation (2.3). When $n\to\infty$, if F has a finite second moment and ω is eligible, then

$$\frac{\varphi_n[\sqrt{\{2\gamma\log(n)\}},X_1,\ldots,X_n,n]}{\varepsilon_n}\to 1$$

in probability. If ω is also good, then

$$\frac{\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}\{\varphi_n(t,X_1,\ldots,X_n,n)\}}{\varepsilon_n}\to 1$$

in probability.

Second, we consider the relatively dense case by calibrating ε as a fixed constant. In this case, the estimators are consistent for all $\gamma\in(0,\frac{1}{2})$. This is the following corollary, the proof of which is similar to that of theorem 8 and is omitted.

Corollary 2. Fix F, $0<\varepsilon<1$ and $0<\gamma\leqslant\frac{1}{2}$, and let $\varphi_n(t;X_1,\ldots,X_n,n)$ be defined as in equation (2.3). When $n\to\infty$, if F has a finite second moment and ω is eligible, then

$$\frac{\varphi_n[\sqrt{\{2\gamma\log(n)\}},X_1,\ldots,X_n,n]}{\varepsilon}\to 1$$

in probability. If in addition ω is good, then

$$\frac{\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}\{\varphi_n(t,X_1,\ldots,X_n,n)\}}{\varepsilon}\to 1$$

in probability.

The conditions in theorem 8 and corollary 2 can be relaxed. However, as the Bayesian model is not very different from the frequentist model, we feel that it is unnecessary to duplicate theorem 5 and corollary 1 completely. The main point here is that the results under the Bayesian model are stronger and cleaner.

From time to time, we may worry about the Gaussian assumption for the non-null effects. We note here that theorem 8 can be extended to the following non-Gaussian case.

Theorem 9. Fix $0<\beta<\frac{1}{2}$ and $0<\gamma\leqslant\frac{1}{2}-\beta$ and let ε_n be $n^{-\beta}$ and $\varphi_n(t;X_1,\ldots,X_n,n)$ be defined as in equation (2.3), where ω is eligible. Suppose that

$$X_j\overset{\mathrm{IID}}{\sim}(1-\varepsilon_n)\,\phi(x)+\varepsilon_n\,g(x),$$

where $g(x)$ is a density function which has a finite second moment and satisfies that

$$\lim_{t\to\infty}[\mathrm{Re}\{\hat{g}(t)\}/\hat{\phi}(t)]=0.$$

Then, as $n\to\infty$,

$$\frac{\varphi_n[\sqrt{\{2\gamma\log(n)\}},X_1,\ldots,X_n,n]}{\varepsilon_n}\to 1$$

in probability.

Here, $\hat{g}(t)$ is the Fourier transform of $g(x)$ and $\mathrm{Re}\{\hat{g}(t)\}$ denotes its real part. We note here that no Gaussian mixture assumption is made on g. Theorem 9 is proved in Appendix A.

Next, we discuss the connection between the approach proposed and some recent procedures in the field of large-scale multiple testing.

3.1. Connection with false discovery rate controlling procedures

A strategy in microarray analysis is first to identify a subset of genes for follow-up study (Smyth, 2004), and then to focus on this subset in subsequent experiments. In the current setting, a natural approach to the problem is to find the largest threshold $\hat{t}_n = \hat{t}_n(X_1, \ldots, X_n)$ such that the subset $\{X_j : |X_j| \geqslant \hat{t}_n\}$ contains at least $n\varepsilon_n\alpha$ signals, where $0 < \alpha < 1$ (e.g. $\alpha = 95\%$). Note here that the total number of signals in the data set is $n\varepsilon_n$. By combining the estimators proposed with the recent FDR procedure by Benjamini and Hochberg (1995), we can give an interesting approach to setting the threshold.

To implement Benjamini and Hochberg's FDR procedure in the current setting, we view model (3.1) as testing n independent null hypotheses $H_j : \mu_j = 0$, $j = 1, \ldots, n$. For any parameter $0 < q < 1$, the procedure picks a threshold $t_q = t_q^{\mathrm{FDR}}(X_1, \ldots, X_n)$, rejects all those hypotheses with $|X_j|$ exceeding the threshold and accepts all others. If we call any case a 'discovery' when H_j is rejected, then a 'false discovery' is a situation where H_j is falsely rejected. Benjamini and Hochberg's procedure controls the FDR, which is the expected value of the *false discovery proportion* FDP (Genovese and Wasserman, 2004):

$$\mathrm{FDP}_q = \frac{\#\{\text{falsely discoveries}\}_q}{\#\{\text{total discoveries}\}_q}.$$

The following theorem is proved in Benjamini and Yekutieli (2005) and Ferreira and Zwinderman (2006).

Theorem 10. For any $0 < q < 1$, let FDP_q be the false discovery proportion that is obtained by implementing Benjamini and Hochberg's FDR procedure to model (3.1); then for any μ and $n \geqslant 1$, $E[\mathrm{FDP}_q] = (1 - \varepsilon_n)q$.

We now combine the approach proposed with Benjamini and Hochberg's FDR procedure to tackle the problem that was mentioned earlier in this subsection. Viewing theorem 10 from a different perspective, we have

$$\#\{\text{true discoveries}\}_q \approx \#\{\text{total discoveries}\}_q\{1 - (1 - \varepsilon_n)q\}.$$

Note that the number of total discoveries is observable, so, to obtain $\alpha n\varepsilon_n$ true discoveries out of all discoveries, we should pick q such that

$$\#\{\text{total discoveries}\}_q\{1 - (1 - \varepsilon_n)q\} \approx \alpha n\varepsilon_n.$$

This suggests the following procedure.

Step 1: estimate ε_n by any of the procedures proposed, denote the estimation by $\hat{\varepsilon}_n$ (e.g. $\hat{\varepsilon}_n = \varphi_n\{\sqrt{\log(n)}; X_1, \ldots, X_n, n\}$ or $\hat{\varepsilon}_n = \sup_{\{0 \leqslant t \leqslant \sqrt{\log(n)}\}}\{\varphi_n(t; X_1, \ldots, X_n, n)\}$ when ω is good).

Step 2: solve for q from the equation $\#\{\text{total discoveries}\}_q = \alpha n\hat{\varepsilon}_n / \{1 - (1 - \hat{\varepsilon}_n)q\}$, where, for any $0 < q < 1$, $\#\{\text{total discoveries}\}_q$ is obtained by implementing Benjamini and Hoch-

berg's FDR procedure. Denote the solution by $\hat{q}$ (pick any if there are more than one).
Step 3: implement Benjamini and Hochberg's FDR procedure with $q = \hat{q}$.

Apply the procedure to model (3.1) and denote the resulting true positive discoveries (i.e. H_j that are correctly rejected) by

$$\hat{T}_n(\alpha, \hat{\varepsilon}_n) = \hat{T}_n(\alpha, \hat{\varepsilon}_n; X_1, \ldots, X_n). \tag{3.2}$$

Though $\hat{T}_n(\alpha, \hat{\varepsilon}_n)$ is a random quantity that is not directly observable and does not have an explicit formula, it equals $n\varepsilon_n\alpha$ approximately for large n. Consequently, in the resulting set of discoveries in step 3, about $n\varepsilon_n\alpha$ discoveries are true positive discoveries. Take $\alpha = 0.95$ for example; the resulting set contains about 95% of all true positive in the original set! This is the following theorem, which is proved in Appendix A.

Theorem 11. Fix $0 < \alpha < 1$, $0 < \varepsilon < 1$ and F, and let $\hat{T}_n(\alpha, \hat{\varepsilon}_n)$ be defined as in equation (3.2) where $\hat{\varepsilon}_n$ is consistent for ε. When $n \to \infty$, $\hat{T}_n(\alpha, \hat{\varepsilon}_n)/n\varepsilon \to \alpha$ in probability.

3.2. Connection with other procedures in large-scale multiple testing

The approach proposed is also connected with several other recent procedures in the field of large-scale multiple testing.

The procedure proposed is intellectually connected with the optimal discovery approach of Storey (2007), as well as the local FDR approach of Efron *et al.* (2001). Storey noticed that, by controlling the expected fraction of false positive discoveries, the optimal approach to obtaining the largest expected number of true positive discoveries is to utilize an oracle which he called the *optimal discovery function*. Under the current model, the optimal discovery function can be written as

$$\mathrm{OD}(x) = 1 - \frac{(1-\varepsilon)\,\phi(x)}{(1-\varepsilon)\,\phi(x) + \varepsilon \int \phi(x-u)\,\mathrm{d}F(u)}.$$

Note here that the denominator is the marginal density of test statistics X_j and can be estimated by many density estimation methods, e.g. kernel methods (Wasserman, 2006), so the problem of estimating the optimal discovery function reduces to the problem of estimating the proportion $1-\varepsilon$. We thus expect to see better results by combining the approach proposed with the optimal discovery approach.

The approach proposed is also intellectually connected with the B-statistic of Lönnstedt and Speed (2002), which was proposed for analysing microarray data. As mentioned in Lönnstedt and Speed (2002), the implementation of the B-statistic depends on knowledge of ε_n:

> '... one drawback in using B is that we need a value for the prior proportion of differentially expressed genes...'.

Combining the approach proposed with the B-statistic, we expect to see better results in many applications.

To conclude this section, we mention that there are many other procedures that depend more or less on the proportion, e.g. Benjamini *et al.* (2005). We expect the estimated proportion to be helpful in implementing these procedures.

4. Comparison with Meinshausen and Rice's estimator

Recently, there have been some interesting approaches to the problem of estimating the proportion; among them are the work by Meinshausen and Rice (2006) (see also Efron *et al.* (2001), Genovese and Wasserman (2004) and Meinshausen and Bühlmann (2005)). These procedures

are intellectually connected with each other, so we discuss only that in Meinshausen and Rice (2006).

Meinshausen and Rice considered a setting in which they tested n uncorrelated null hypotheses H_j, $j = 1, \ldots, n$. Associated with the jth hypothesis is a p-value p_j, which has a uniform distribution—$U(0, 1)$—when H_j is true and some other distribution otherwise. It is of interest to estimate the proportion of non-null effects (i.e. untrue hypotheses). Meinshausen and Rice proposed the estimator

$$\varepsilon_n^{\mathrm{MR}} = \sup_{0<t<1} \left\{ \frac{F_n(t) - t - \beta_{n,\alpha}\delta(t)}{1-t} \right\},$$

where $F_n(t)$ is the empirical CDF of the p-values and $\beta_{n,\alpha}\,\delta(t)$ is the so-called *bounding function* (Meinshausen and Rice, 2006). They have studied various aspects of the estimator including its consistency. In fact, by modelling the p-values as samples from a two-component mixture: $p_j \sim^{\mathrm{IID}} (1-\varepsilon)\,U(0,1) + \varepsilon h$, $j = 1, \ldots, n$, they found that, for the estimator to be consistent, it is necessary that

$$\operatorname*{essinf}_{\{0<p<1\}} \{h(p)\} = 0. \tag{4.1}$$

Here ε is the proportion of non-null effects, $U(0, 1)$ is the marginal density of p_j when H_j is true and h is the marginal density when H_j is untrue. We remark here that a similar result can be found in Genovese and Wasserman (2004), who referred to densities satisfying condition (4.1) as *pure* densities. Also, Swanepoel (1999) proposed a different estimator using spacings, but the consistency of the estimator is also limited to the case where h is pure.

Unfortunately, the purity condition is generally not satisfied in the n normal means setting. To elaborate, we translate the previous model from the p-scale to the z-scale through the transformation $X_j = \bar{\Phi}^{-1}(p_j)$, $j = 1, \ldots, n$. It follows that X_j are samples from the density $(1-\varepsilon)\,\phi(x) + \varepsilon\, g(x)$, where $g(x) = h\{\bar{\Phi}(x)\}\phi(x)$. Here, $\bar{\Phi}$ and ϕ denote the survival function and the density function of the standard normal respectively. Accordingly, the purity condition (4.1) is equivalent to

$$\operatorname*{essinf}_{\{-\infty<x<\infty\}} \{g(x)/\phi(x)\} = 0, \tag{4.2}$$

which says that $g(x)$ has a thinner tail than that of the standard normal, either to the left or to the right. The following lemma says that the purity condition is generally not statisfied for Gaussian location mixtures.

Lemma 2. Suppose that $g(x) = \int \phi(x-u)\,\mathrm{d}F(u)$ for some distribution function F. If $P_F\{u < 0\} \neq 0$ and $P_F\{u > 0\} \neq 0$, then $\operatorname{essinf}_{\{-\infty<x<\infty\}}\{g(x)/\phi(x)\} > 0$. If F is also symmetric, then

$$\operatorname*{essinf}_{\{-\infty<x<\infty\}} \{g(x)/\phi(x)\} = \int \exp(-u^2/2)\,\mathrm{d}F(u) > 0.$$

The proof of lemma 2 is elementary so we skip it. Lemma 2 implies that Meinshausen and Rice's estimator (and also those in Genovese and Wasserman (2004), Efron *et al.* (2001), Meinshausen and Bühlmann (2005) and Swanepoel (1999)) is generally *not* consistent. In Section 5, we shall further compare the approach proposed with Meinshausen and Rice's estimator and show that the latter is generally unsatisfactory for the present setting. However, we mention here that one advantage of Meinshausen and Rice's estimator—which we like—is that it provides an honest confidence lower bound for the proportion even without the Gaussian model for the non-null effects; see Meinshausen and Rice (2006) for details.

Recall that the approaches proposed are consistent if the following condition holds (theorem 9):

$$\lim_{t\to\infty}[\mathrm{Re}\{\hat{g}(t)\}/\hat{\phi}(t)]=0. \tag{4.3}$$

It is interesting that condition (4.3) is highly similar to the purity condition (4.2), and the only major difference is that the former concerns g and ϕ themselves, and the latter concerns their Fourier transforms. In a sense, our findings in this paper complement those in Genovese and Wasserman (2004) and Meinshausen and Rice (2006). First, we mirror the purity condition that was originally defined in the spatial domain to its cousin in the frequency domain—the purity condition based on the Fourier transform. Second, we develop a class of new estimators and show them to be consistent for the true proportion when the Fourier transform purity condition holds (but the original purity condition may be violated). We mention that the idea here can be largely generalized; see Jin *et al.* (2007) for the detail.

To conclude this section, we mention that, for an approach to be consistent, *some* constraint on g is necessary, as otherwise the proportion would be unidentifiable (e.g. Genovese and Wasserman (2004)). In application problems where we cannot make *any* assumption on h (e.g. the purity condition or the Fourier transform purity condition), it is argued in Genovese and Wasserman (2004) (see also Meinshausen and Rice (2006)) that all that we can hope to estimate consistently is the quantity

$$\bar{\varepsilon}=\varepsilon\left[1-\operatorname*{essinf}_x\left\{\frac{g(x)}{\phi(x)}\right\}\right], \tag{4.4}$$

which we call the Genovese–Wasserman *lower bound*. As pessimistic as it may seem, in many applications, some reasonable assumptions on g can be made. See for example Efron (2004), Jin and Cai (2007) and Jin *et al.* (2007).

5. Simulation study

We have conducted a small-scale empirical study. The idea is to choose a few interesting cases and to investigate the performance of the approach proposed for different choices of ω, signals and parameters. Let ω be any one of the uniform, triangle or smooth density as introduced in Section 2.1, and denote the estimators that were proposed in Section 2.2 and Section 2.3 by

$$\begin{aligned}\hat{\varepsilon}_n(\gamma)&\equiv\varphi_n[\sqrt{\{2\gamma\log(n)\}};X_1,\ldots,X_n,n],\\ \hat{\varepsilon}_n^*(\alpha_n)&\equiv\varphi_n\{t_n^*(\omega;\alpha_n);X_1,\ldots,X_n,n\}\end{aligned} \tag{5.1}$$

respectively. The simulation experiment contains six parts, which we now describe. For clarification, we note that, except for the last part of the experiment, the variances of the X_j are assumed to be known and equal 1.

5.1. Experiment (a)

We investigate the effect of γ over $\hat{\varepsilon}_n(\gamma)$. Set $n=10^5$, $\varepsilon_n=0.2$ and $\mu_0=0.5,0.75,1,1.25$. For each μ_0, we generate $n(1-\varepsilon_n)$ samples from $N(0,1)$, and $n\varepsilon_n$ samples from $N(\mu_j,1)$; here $|\mu_j|$ are randomly generated from $U(\mu_0,\mu_0+1)$, and $\mathrm{sgn}(\mu_j)$ are randomly generated from $\{-1,1\}$ with equal probabilities, where U denotes the uniform distribution, and sgn denotes the usual sign function. As a result, the Genovese–Wasserman lower bound (see equation (4.4)) equals 0.079, 0.107, 0.132 and 0.153 correspondingly. Next, we pick 50 different γs so that $\sqrt{(2\gamma)}$ ranges from 0.02 to 1 with an increment of 0.02. We then apply $\hat{\varepsilon}_n(\gamma)$ to the whole sample for each γ

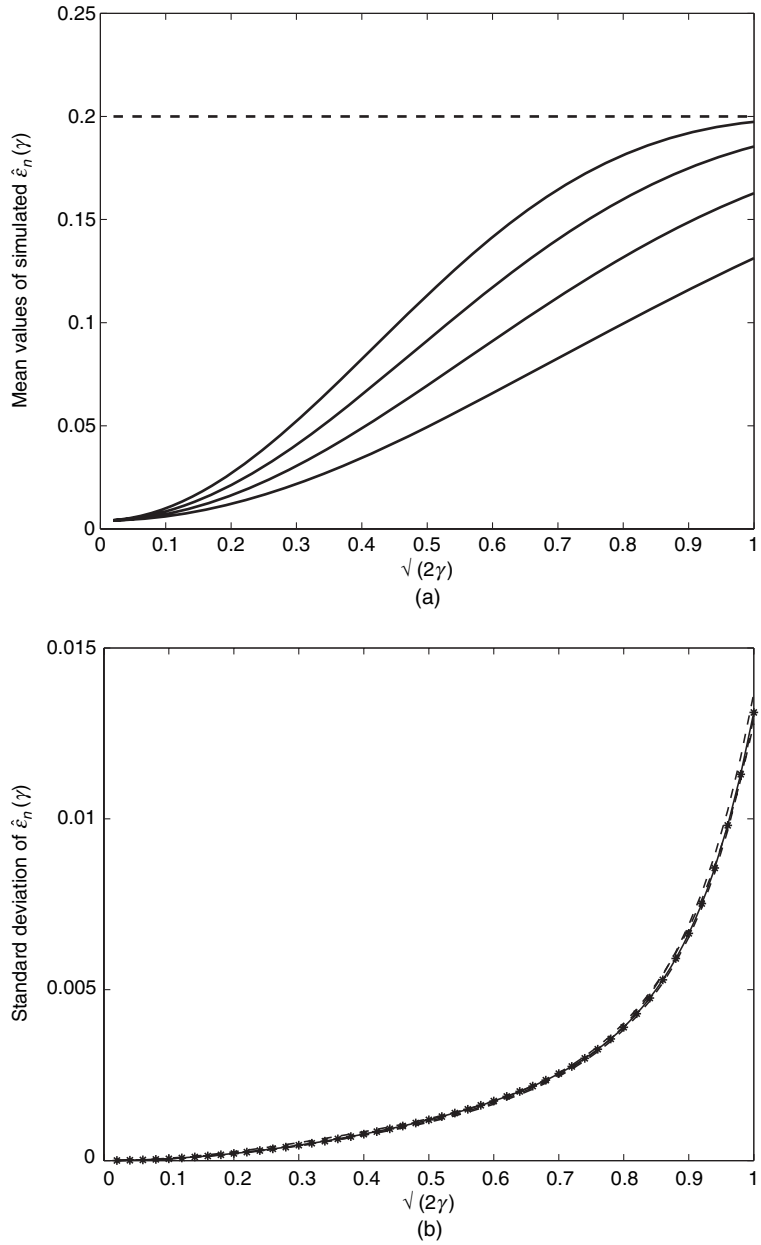

Fig. 2. Effect of γ on $\hat{\varepsilon}_n(\gamma)$ (see experiment (a) for details): (a) four curves corresponding, from bottom to top, to $\mu_0 = 0.5, 0.75, 1, 1.25$; (b) four curves (which happen to be very close to each other corresponding, from bottom to top, to $\mu_0 = 0.5, 0.75, 1, 1.25$

and each ω. Last, we repeat the whole process independently 100 times. For later reference, we refer to samples that are generated in this way as *samples with signals uniformly distributed with parameters* $(n, \varepsilon_n, \mu_0)$.

The results are summarized in Fig. 2. To be brief, we report the results corresponding to the triangle density only. The results suggest the following. First, $\hat{\varepsilon}_n(\gamma)$ monotonely increases as γ

increases. The estimator is generally conservative and underestimates the true proportion, but it becomes increasingly closer to the true proportion as γ approaches $\frac{1}{2}$. This, together with more empirical studies, suggests that the best choice in this family is $\hat{\varepsilon}_n(\frac{1}{2})$, and also that the difference between the two estimators $\hat{\varepsilon}_n(\frac{1}{2})$ and $\sup_{\{0<\gamma\leqslant 1/2\}}\{\hat{\varepsilon}_n(\gamma)\}$ is generally negligible. Second, if we fix γ and let μ_0 increase (so that the strength of the signal increases) then $\hat{\varepsilon}_n(\gamma)$ becomes increasingly accurate and becomes fairly close to the true proportion when $\mu_0 \geqslant 1$. Third, when γ increases, the standard deviations (SDs) increase as well, which implies that the estimator becomes increasingly unstable. However, the SDs remain of a smaller magnitude than that of the biases, so the stochastic fluctuation of the estimator is generally negligible. It is interesting that the SDs do not respond much to the strength of the signals; they remain almost the same when the signals range from faint to strong.

5.2. Experiment (b)

We compare the performances of $\hat{\varepsilon}_n(\frac{1}{2})$ and $\hat{\varepsilon}_n^*(\alpha_n)$. Especially, we investigate how well the SD of $\hat{\varepsilon}_n^*(\alpha_n)$ is controlled. First, for each of $n = 0.5\times 10^4, 1\times 10^4, 2\times 10^4, 4\times 10^4, 8\times 10^4$, fix $\varepsilon_n = 0.2$ and generate samples with signals uniformly distributed with parameters $(n, 0.2, 1)$. Next, we apply $\hat{\varepsilon}_n(\frac{1}{2})$ and $\hat{\varepsilon}_n^*(\alpha_n)$ to the sample, with $\alpha_n = 0.015, 0.020, 0.025$ and ω being the uniform, triangle or smooth density. We repeat the whole process 100 times and report the results in Table 1. Here, the Genovese–Wasserman lower bound does not depend on n and equals 0.132.

The results suggest the following. Firstly, the adaptive approach—$\hat{\varepsilon}_n^*(\alpha_n)$—gives tight control

Table 1. Comparison of SDs and RMSEs of $\hat{\varepsilon}_n(\frac{1}{2})$ and $\hat{\varepsilon}_n^*(\alpha_n)$†

Density	*Function*	α_n	*Parameter*	*Results for the following values of n:*				
				0.5×10^4	10^4	2×10^4	4×10^4	8×10^4
Uniform	$\hat{\varepsilon}_n(\frac{1}{2})$		SD	0.0878	0.0837	0.0671	0.0682	0.0699
			RMSE	0.0906	0.0841	0.0710	0.0725	0.0713
	$\hat{\varepsilon}_n^*(\alpha_n)$	0.015	SD	0.0083	0.0105	0.0127	0.0120	0.0155
			RMSE	0.0785	0.0283	0.0128	0.0192	0.0253
		0.020	SD	0.0149	0.0164	0.0172	0.0167	0.0208
			RMSE	0.0350	0.0167	0.0212	0.0263	0.0302
		0.025	SD	0.0214	0.0220	0.0214	0.0214	0.0259
			RMSE	0.0231	0.0230	0.0275	0.0312	0.0341
Triangle	$\hat{\varepsilon}_n(\frac{1}{2})$		SD	0.0248	0.0206	0.0158	0.0139	0.0144
			RMSE	0.0423	0.0391	0.0309	0.0261	0.0253
	$\hat{\varepsilon}_n^*(\alpha_n)$	0.015	SD	0.0054	0.0105	0.0118	0.0122	0.0150
			RMSE	0.1137	0.0529	0.0360	0.0271	0.0253
		0.020	SD	0.0145	0.0165	0.0160	0.0171	0.0200
			RMSE	0.0566	0.0413	0.0308	0.0257	0.0267
		0.025	SD	0.0205	0.0220	0.0201	0.0218	0.0249
			RMSE	0.0451	0.0390	0.0300	0.0272	0.0297
Smooth	$\hat{\varepsilon}_n(\frac{1}{2})$		SD	0.0199	0.0152	0.0121	0.0092	0.0095
			RMSE	0.0334	0.0281	0.0210	0.0149	0.0131
	$\hat{\varepsilon}_n^*(\alpha_n)$	0.015	SD	0.0055	0.0104	0.0121	0.0121	0.0151
			RMSE	0.1105	0.0401	0.0210	0.0133	0.0152
		0.020	SD	0.0147	0.0164	0.0163	0.0169	0.0202
			RMSE	0.0451	0.0268	0.0182	0.0169	0.0202
		0.025	SD	0.0209	0.0220	0.0204	0.0217	0.0251
			RMSE	0.0324	0.0260	0.0206	0.0219	0.0253

†See experiment (b) for details.

on the empirical SD; this property is not assumed by $\hat{\varepsilon}_n(\frac{1}{2})$. In fact, the empirical SD of $\hat{\varepsilon}_n^*(\alpha_n)$ seldom exceeds α_n and, if so, only by a very small amount. This suggests that, as predicted by theorem 6, the empirical SD of $\hat{\varepsilon}_n^*(\alpha_n)$ is nicely bounded by α_n. Additionally, the bound is tight and the empirical SDs do not change much for different n and ω: except for a few cases, the empirical SDs fall between $0.7\alpha_n$ and α_n. In contrast, the empirical SD of $\hat{\varepsilon}_n(\frac{1}{2})$ may fluctuate for more than seven times across different ω, and for more than two times across different n. Secondly, in terms of the root-mean-squared errors (RMSEs), the performance of $\hat{\varepsilon}_n^*(\alpha_n)$ is mainly determined by α_n, and different choices of n and ω do not have prominent effects. Lastly, all estimators yield a reasonably good estimate for the true proportion.

In practice, we might want to know how to set α_n (the tolerance parameter). Ideally, if we have good knowledge of both the *variances* and the *biases* of $\hat{\varepsilon}_n^*(\alpha_n)$ across a wide range of α_n, then we know how to select the best α_n. Unfortunately, although sometimes it is possible to estimate the variances (i.e. by using the bootstrap method), it is frequently impossible to estimate the biases. Still, we propose the following *ad hoc* two-stage procedure for selecting α_n. First, we pick $\alpha_n = 0.015$ and obtain $\hat{\varepsilon}_n^*(0.015)$. Second, we select an α_n that is much smaller than $\varepsilon_n^*(0.015)$ and use $\hat{\varepsilon}_n(\alpha_n)$ as the final estimate of the proportion. By doing so, we hope that the stochastic fluctuation of $\hat{\varepsilon}_n^*(\alpha_n)$ has a smaller magnitude than that of the true proportion.

5.3. Experiment (c)

We compare $\hat{\varepsilon}_n^*(\alpha_n)$ with Meinshausen and Rice's (2006) estimator (equation (5) of Meinshausen and Rice (2006)), which we denote by $\hat{\varepsilon}_n^{\mathrm{MR}}$. The bounding function $\beta_{n,\alpha}\,\delta(t)$ is set to $\sqrt{[2t(1-t)\log\{\log(n)\}/n]}$. Fix $n = 80\,000$, $\varepsilon_n = 0.2$ and $\alpha_n = 0.015$, and pick $\mu_0 = 0.5, 0.75, 1, 1.25$. Correspondingly, the Genovese–Wasserman lower bound equals 0.079, 0.107, 0.132 and 0.153. For each μ_0, we generate samples with signals uniformly distributed with parameters $(8 \times 10^4, 0.2, \mu_0)$. We then apply $\hat{\varepsilon}_n^*(\alpha_n)$ and $\hat{\varepsilon}_n^{\mathrm{MR}}$ to the sample, with ω being the triangle density and the smooth density (for brevity, we omit the case for the uniform density). We repeat the whole process 100 times. The results are displayed in Fig. 3; they are also summarized in terms of the SD and RMSE in Table 2. The results suggest that the performance of $\hat{\varepsilon}_n^{\mathrm{MR}}$ is generally unsatisfactory, and $\hat{\varepsilon}_n^*(\alpha_n)$ behaves much better. In fact, $\hat{\varepsilon}_n^*(\alpha_n)$ is encouragingly accurate when μ_0 is greater than 1.

5.4. Experiment (d)

We continue the study in experiment (c) by letting the proportion vary from the dense regime to the moderately sparse regime. Fixing $n = 10^6$ and $\alpha_n = 0.002$, for each of $\varepsilon_n = n^{-0.1}, n^{-0.2}, n^{-0.3}, n^{-0.4}$ ($n^{-0.1} = 0.25$, $n^{-0.2} = 0.063$, $n^{-0.3} = 0.016$ and $n^{-0.4} = 0.004$), we generate samples with signals uniformly distributed with parameters $(10^6, \varepsilon_n, 1.25)$. Correspondingly, the Genovese–Wasserman lower bound equals 0.192, 0.048, 0.012 and 0.003. We apply both $\hat{\varepsilon}_n^*(\alpha_n)$ and $\hat{\varepsilon}_n^{\mathrm{MR}}$ to the whole sample, with ω being the triangle density and the smooth density. We then repeat the whole process 100 times. The results are summarized in Table 3, where we tabulated the RMSE of $\hat{\varepsilon}_n^*(\alpha_n)/\varepsilon_n$ and $\hat{\varepsilon}_n^{\mathrm{MR}}/\varepsilon_n$. The results suggest that $\hat{\varepsilon}_n^*(\alpha_n)$ continues to perform well for the moderately sparse case and continues to outperform the procedure of Meinshausen and Rice (2006).

5.5. Experiment (e)

We continue the study in experiment (c), but with a different ε_n and a different way to generate non-zero μ_js. Fix $n = 80\,000$, $\varepsilon_n = 0.1$ and $\alpha_n = 0.015$, and pick $\sigma_0 = 1, 1.25, 1.50, 1.75$. For each σ_0, we generate $n(1-\varepsilon_n)$ samples from $N(0,1)$, and $n\varepsilon_n$ samples from $N(\mu_j, 1)$, where μ_j are sampled from $N(0, \sigma_0^2)$. Correspondingly, the Genovese–Wasserman lower bound equals

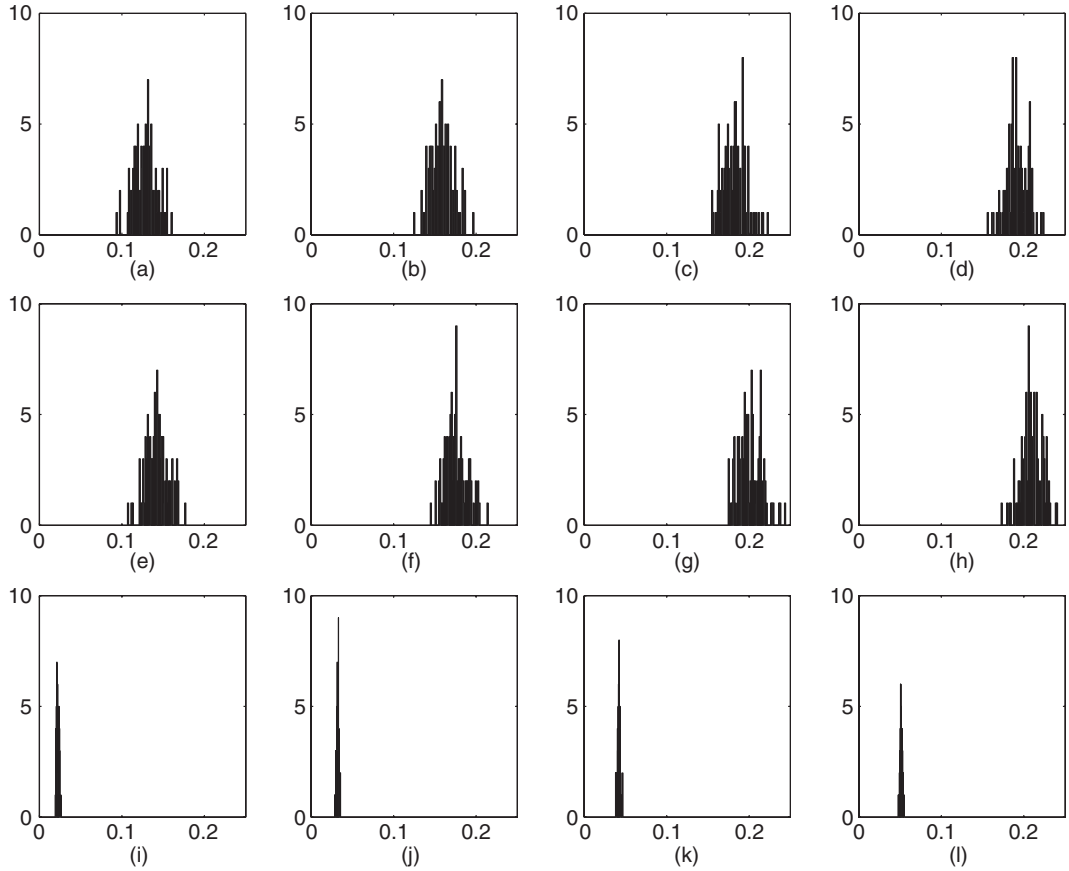

Fig. 3. Histogram comparisons of (a)–(h) $\hat{\varepsilon}_n^*(\alpha_n)$ (with ω being the triangle density and smooth density) with (i)–(l) $\hat{\varepsilon}_n^{\mathrm{MR}}$ (the true proportion is 0.2 and $\alpha_n = 0.015$; see experiment (c) for details): (a), (e), (i) $\mu_0 = 0.5$; (b), (f), (j) $\mu_0 = 0.75$; (c), (g), (k) $\mu_0 = 1$; (d), (h), (l) $\mu_0 = 1.25$

Table 2. RMSEs of $\hat{\varepsilon}_n^*(\alpha_n)/\varepsilon_n$ and $\hat{\varepsilon}_n^{\mathrm{MR}}/\varepsilon_n$ for various signal strengths†

Estimator	*Results for the following signal strengths:*			
	$\mu_0 = 0.5$	$\mu_0 = 0.75$	$\mu_0 = 1$	$\mu_0 = 1.25$
$\hat{\varepsilon}_n^*(\alpha_n)$ ($\alpha_n = 0.015$, triangle)	0.3649	0.2194	0.1139	0.0784
$\hat{\varepsilon}_n^*(\alpha_n)$ ($\alpha_n = 0.015$, smooth)	0.2957	0.1375	0.0792	0.0821
Meinshausen and Rice $\hat{\varepsilon}_n^{\mathrm{MR}}$	0.8882	0.8383	0.7903	0.7435

†See experiment (c) for details.

0.029, 0.038, 0.045 and 0.050. We apply both $\hat{\varepsilon}_n^*(\alpha_n)$ and $\hat{\varepsilon}_n^{\mathrm{MR}}$ to the whole sample, with ω being the triangle density and the smooth density. We then repeat the whole process 100 times. The results are shown in Table 4 in terms of SD and RMSE. In comparison, the non-zero μ_js in experiment (c) are bounded away from 0 by a distance μ_0 but, in the current case, a certain fraction of non-zero μ_js is concentrated around 0. We thus expect that the proportion is more

Table 3. RMSEs of $\hat{\varepsilon}_n^*(\alpha_n)/\varepsilon_n$ and $\hat{\varepsilon}_n^{\mathrm{MR}}/\varepsilon_n$ for various sparsity levels†

Estimator	*Results for the following sparsity levels:*			
	$\varepsilon_n = n^{-0.1}$	$\varepsilon_n = n^{-0.2}$	$\varepsilon_n = n^{-0.3}$	$\varepsilon_n = n^{-0.4}$
$\hat{\varepsilon}_n^*(\alpha_n)$ ($\alpha_n = 0.0015$, triangle)	0.1409	0.1391	0.1533	0.2854
$\hat{\varepsilon}_n^*(\alpha_n)$ ($\alpha_n = 0.0015$, smooth)	0.0611	0.0612	0.0903	0.2613
Meinshausen and Rice $\hat{\varepsilon}_n^{\mathrm{MR}}$	0.7341	0.7421	0.7697	0.8335

†$n = 10^6$. See experiment (d) for details.

Table 4. RMSEs of $\hat{\varepsilon}_n^*(\alpha_n)/\varepsilon_n$ and $\hat{\varepsilon}_n^{\mathrm{MR}}/\varepsilon_n$†

Estimator	*Results for the following signal strengths:*			
	$\sigma_0 = 1$	$\sigma_0 = 1.25$	$\sigma_0 = 1.5$	$\sigma_0 = 1.75$
$\hat{\varepsilon}_n^*(\alpha_n)$ ($\alpha_n = 0.015$, triangle)	0.5804	0.5122	0.4532	0.3998
$\hat{\varepsilon}_n^*(\alpha_n)$ ($\alpha_n = 0.015$, smooth)	0.5389	0.4674	0.4078	0.3521
Meinshausen and Rice $\hat{\varepsilon}_n^{\mathrm{MR}}$	0.9247	0.8943	0.8671	0.8371

†The non-zero μ_j are Gaussian distributed; see experiment (e) for details.

difficult to estimate in the current situation. The differences can be seen in more detail by comparing Table 2 and Table 4. In both cases, the approaches proposed compare favourably with that of Meinshausen and Rice (2006). We mention that the unsatisfactory behaviour of $\hat{\varepsilon}_n^{\mathrm{MR}}$ is mainly due to its inconsistency in the current setting; tuning the bounding function would not be very helpful.

5.6. Experiment (f)

We now study an example to obtain a feeling of how the estimators behave in the cases where the normality assumption is violated. Fix $n = 10^4$, $\varepsilon_n = 0.2$ and $\alpha_n = 0.015$, and pick $\lambda = 1, 2, 3, 4$. For each λ, we generate $n(1-\varepsilon_n)$ samples from $N(0,1)$, and $n\varepsilon_n$ samples from $\mathrm{DE}(\lambda)$, where $\mathrm{DE}(\lambda)$ denotes the double-exponential distribution with mean 0 and standard deviation $\lambda\sqrt{2}$. Correspondingly, the Genovese–Wasserman lower bound equals $0.048, 0.089, 0.121$ and 0.139. We apply both $\hat{\varepsilon}_n^*(\alpha_n)$ and $\hat{\varepsilon}_n^{\mathrm{MR}}$ to the whole sample, with ω being the triangle density and the smooth density. We then repeat the whole process 100 times. The results are reported in Fig. 4. In this example, despite the violation of the normality assumption, the estimator proposed behaves well and compares favourably with that of Meinshausen and Rice (2006).

6. Extensions

The approach proposed can be extended to estimating many other functionals of the normal mean vector. Below are some functionals which are of interest in theory and applications.

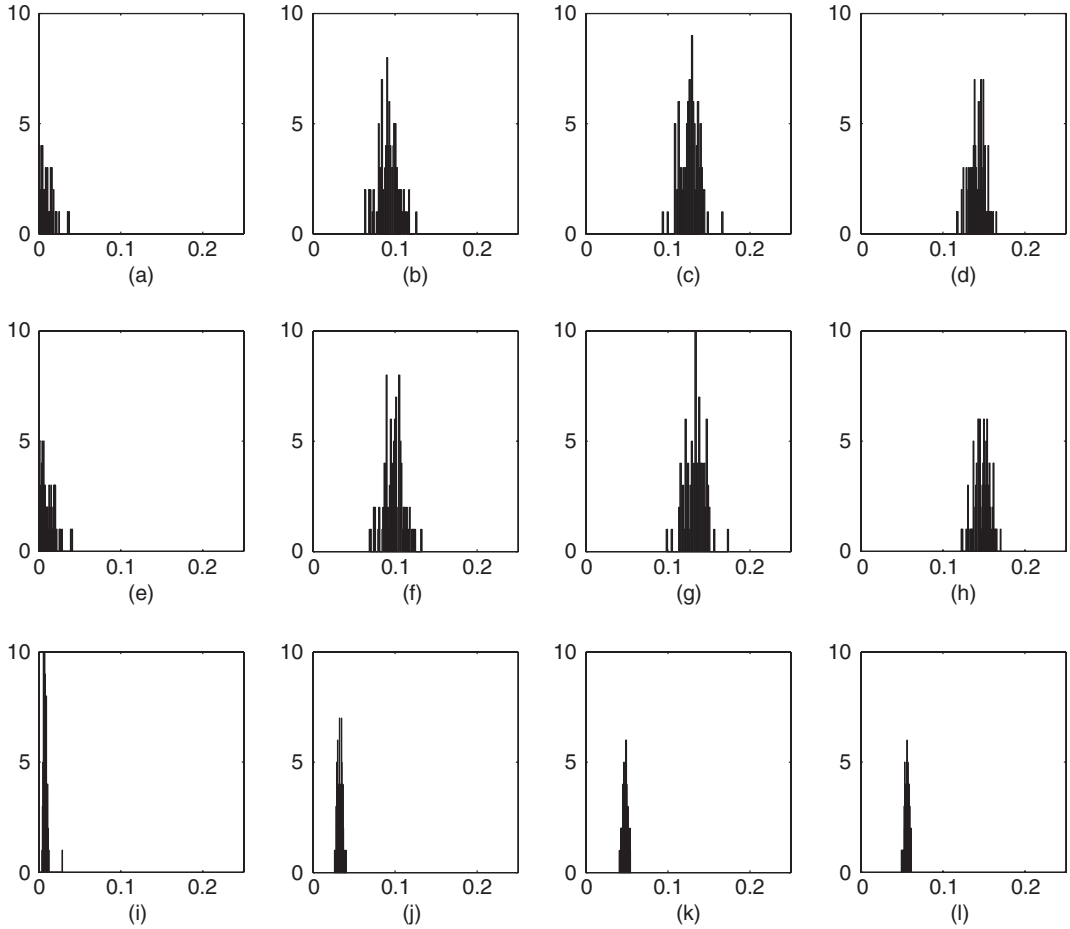

Fig. 4. Histogram comparison of the behaviour of (a)–(h) $\hat{\varepsilon}_n(\alpha_n)$ (with ω being the triangle density and smooth density) with (i)–(l) $\hat{\varepsilon}_n^{\mathrm{MR}}$ when the normality assumption is violated (the true proportion is 0.2 and the non-null effects were generated from the double-exponential distribution; see experiment (f) for details): (a), (e), (i) $\lambda = 1$; (b), (f), (j) $\lambda = 2$; (c), (g), (k) $\lambda = 3$; (d), (h), (l) $\lambda = 4$

6.1. Example I

In many applications in designing downstream experiments (see Yang *et al.* (2002) as well as

`www.niams.nih.gov/rtbc/labs_branches/ost/core_facility/biodata/strategy.htm`),

only signals with a magnitude exceeding a given threshold are of interest. This motivates a careful study on estimating the proportion of normal means that exceeds a given threshold.

6.2. Example II

The level of sparsity of the mean vector plays an important role in many inference problems. There are many models for the level of sparsity, and the model where the level of sparsity is defined as the average l^p-norm is particularly well known (e.g. Abramovich *et al.* (2006)). A successful estimation for the average l^p-norm $(1/n)\Sigma_{j=1}^n |\mu_j|^p$ has potential applications.

6.3. Example III

A variant of the functional in example II is $(1/n)\Sigma_{j=1}^{n} \min\{|\mu_j|^p, a^p\}$, where $a>0$ is a constant which may depend on n but not j. This variant is easier to handle but can still capture the essence of that in example II. In fact, if we take $a=\sqrt{\{2\log(n)\}}$, then example II can be viewed as a special case of example III. The reason is that, since the extreme value of n standard normals is approximately $\sqrt{\{2\log(n)\}}$ (Shorack and Wellner, 1986), any signals with a magnitude that is larger than $\sqrt{\{2\log(n)\}}$ can be easily estimated individually, so it makes sense to assume that the magnitude of each μ_j does not exceed $\sqrt{\{2\log(n)\}}$. Consequently, the functional in example II reduces to the current functional.

Motivated by these examples, we introduce a univariate function $\pi=\pi(u;a)$ over $\mathbf{R}$ which satisfies

(a) $\pi(u;a)=0$ when $|u|>a$,
(b) $\pi(u;a)$ is symmetric and continuous over $[-a,a]$ and
(c) $0\leqslant\pi(u;a)\leqslant\pi_0$ and $\pi_0>0$, where $\pi_0=\pi(0;a)$.

We are interested in estimating the functional

$$\Pi_n(\mu;a)=\frac{1}{n}\sum_{\{j:\mu_j\neq\pm a\}}\Big[\pi_0-\pi(\mu_j;a)+\frac{2\pi_0-\pi_a}{2}\,\#\{j:\mu_j=\pm a\}\Big], \tag{6.1}$$

where $\pi_a=\pi(a;a)$. Owing to the possible discontinuity of π at $\pm a$, we use a randomized rule at $\pm a$ (i.e. $\pi(a;a)$ equal to the value of $\lim_{u\to a+}\{\pi(u;a)\}$ and the value of $\lim_{u\to a-}\{\pi(u;a)\}$ with 0.5 probability each, and similarly for $\pi(-a;a)$). When π is continuous at $\pm a$, the functional reduces to $\Pi_n(\mu;a)=(1/n)\Sigma_{j=1}^{n}\{\pi_0-\pi(\mu_j;a)\}$. The functional includes examples I–III as special cases. In fact, in example I, $\pi(u;a)=\mathbf{1}_{\{|u|\leqslant a\}}$, and

$$\Pi_n(\mu;a)=\frac{1}{n}\#\{j:|\mu_j|>a\}+\frac{\#\{j:|\mu_j|=a\}}{2}.$$

In example III, $\pi(u;a)=(a^p-|u|^p)^+$ and $\Pi_n(\mu;a)=(1/n)\Sigma_{j=1}^{n}\min(|\mu_j|^p,a^p)$.

The idea that we introduced in Section 2 can be extended to estimating $\Pi_n(\mu;a)$ (note that $\pi_0-\pi(u;a)$ plays a similar role to that of $\mathbf{1}_{\{u>0\}}$). Similarly, we hope to construct a function $\psi=\psi(u;t,a)$ such that, for any fixed u, $\lim_{t\to\infty}\{\psi(u;t,a)\}=0$, $\pi_a/2$ and $\pi(u;a)$ according to $|u|>a$, $|u|=a$ and $|u|<a$. Once such a ψ has been constructed, we let the phase function be

$$\varphi(t;\mu,n,a)=\frac{1}{n}\sum_{j=1}^{n}\{\pi_0-\psi(\mu_j;t,a)\}. \tag{6.2}$$

It follows that, for any fixed n and μ, $\lim_{t\to\infty}\{\varphi(t;\mu,n,a)\}=\Pi_n(\mu;a)$, and we expect that consistent estimators of $\Pi_n(\mu;a)$ can be constructed in a similar fashion to that in Section 2.

To do so, we pick an eligible density $\omega(\xi)$ and define $K(u)\equiv\hat{\omega}(u)=\int_{-1}^{1}\omega(\xi)\cos(u\xi)\,\mathrm{d}\xi$. Denote $A(\omega)=\int\{\int_{-1}^{1}\omega(\xi)\cos(u\xi)\,\mathrm{d}\xi\}\,\mathrm{d}u\equiv\int K(u)\,\mathrm{d}u$. When $A(\omega)\neq0$, we introduce the *kernel function* $K_t(u)=\{t/A(\omega)\}\,K(tu)$, where $t>0$. Note that $\int K_t(u)\,\mathrm{d}u=1$. We then construct $\psi(\cdot;t,a)$ as the convolution of K_t and $\pi(u;a)$:

$$\psi(u;t,a)\equiv K_t(u)*\pi(u;a)=\int_{-a}^{a}K_t(u-y)\,\pi(y;a)\,\mathrm{d}y. \tag{6.3}$$

It is shown in lemma 8.4 of Jin (2007) that the function ψ can be equivalently written as

$$\psi(u;t,a)=\frac{t}{A(\omega)}\int_{-1}^{1}\omega(\xi)\,\hat{\pi}(t\xi;a)\cos(tu\xi)\,\mathrm{d}\xi$$

and has the property that was desired above. As a result, we have the following theorem, which is proved in section 8 of Jin (2007).

Theorem 12. Fix $a>0$ and let $\Pi_n(\mu;a)$ be defined as in equation (6.1) and φ be defined as in equation (6.2), where the density ω is eligible. If $A(\omega)\neq 0$, then, for any fixed n and μ, $\lim_{t\to\infty}\{\varphi(t;\mu,n,a)\}=\Pi_n(\mu;a)$.

We now construct the empirical phase function. Similarly, the key is to construct a function $\kappa(x;t,a)$ that connects to $\psi(t;u,a)$ by taking the expectation. Define

$$\kappa(x;t,a)=\frac{t}{A(\omega)}\int_{-1}^{1}\omega(\xi)\,\hat{\pi}(t\xi;a)\exp\Big(\frac{t^2\xi^2}{2}\Big)\cos(tx\xi)\,\mathrm{d}\xi. \tag{6.4}$$

It is proved in lemma 8.4 of Jin (2007) that

$$E[\kappa(X;t,a)]=\psi(u;t,a), \qquad t>0, \quad X\sim N(u,1). \tag{6.5}$$

Thus, if we let the empirical phase function be

$$\varphi_n(t;X_1,\ldots,X_n,n,a)=\frac{1}{n}\sum_{j=1}^{n}\{\pi_0-\kappa(X_j;t,a)\}, \tag{6.6}$$

then, through the equality $E[\varphi_n(t;X_1,\ldots,X_n,n,a)]\equiv\varphi(t;\mu,n,a)$, the empirical function naturally connects to the phase function.

We are now ready for the main claim of this section. When $\pi(\cdot;a)$ is discontinuous at $\pm a$, similarly to $\Theta_n(r,\gamma)$ (see expression (2.10)), we define the following set of parameters:

$$\Theta_n^a(r)=\Big\{\mu\in B_n^1(r),\ \min_{1\leqslant j\leqslant n}\{||\mu_j|-a|\}\geqslant\frac{\log\{\log(n)\}}{\sqrt{\{2\log(n)\}}}\Big\}, \tag{6.7}$$

where, as before, $B_n^1(r)$ is the l^1-ball in $\mathbf{R}^n$ with radius r. The following theorem is proved in Section 8 of Jin (2007).

Theorem 13. Fix $a>0$, $r>0$ and $0<\gamma\leqslant\frac{1}{2}$, let $\varphi_n(t;X_1,\ldots,X_n,n,a)$ be defined as in equation (6.6), where the density ω is eligible with $A(\omega)\neq 0$, and suppose that π is absolutely continuous over $[-a,a]$. When $n\to\infty$, $\sup_{\{\Theta_n^a(r)\}}(|\varphi_n[\sqrt{\{2\gamma\log(n)\}};X_1,\ldots,X_n,n,a]-\Pi_n(\mu;a)|)\to 0$ in probability. If additionally π is continuous everywhere, then $\sup_{\{B_n^1(r)\}}(|\varphi_n[\sqrt{\{2\gamma\log(n)\}};X_1,\ldots,X_n,n,a]-\Pi_n(\mu;a)|)\to 0$ in probability.

Again, the condition that all μ_j are bounded away from $\pm a$ by an amount

$$\log\{\log(n)\}/\sqrt{\{2\log(n)\}}$$

can be largely relaxed; we choose $\Theta_n^a(r)$ only to make the presentation cleaner.

We now continue the discussion of examples I–III. Theorems 12 and 13 directly apply to examples I and III. Moreover, it can be shown that theorems 12 and 13 continue to hold if we take $a=\sqrt{\{2\log(n)\}}$ in example III, so these theorems apply to example II as well. Also, we note that some explicit formulae are available for these examples. In fact, in example I, $\hat{\pi}(\xi,a)=2\sin(a\xi)/\xi$,

$$\psi(u;t,a)=\frac{t}{A(\omega)}\int_{-1}^{1}\omega(\xi)\frac{2\sin(at\xi)}{t\xi}\cos(tu\xi)\,\mathrm{d}\xi$$

and

$$\kappa(x;t,a)=\frac{t}{A(\omega)}\int_{-1}^{1}\omega(\xi)\frac{2\sin(at\xi)}{t\xi}\exp\Big(\frac{t^2\xi^2}{2}\Big)\cos(tx\xi)\,\mathrm{d}\xi.$$

In example III, when $p=1$, $\hat{\pi}(\xi;a)=2\{1-\cos(a\xi)\}/\xi^2$,

$$\psi(u;t,a)=\frac{t}{A(\omega)}\int_{-1}^{1}\omega(\xi)\frac{2\{1-\cos(at\xi)\}}{t^2\xi^2}\cos(tu\xi)\,\mathrm{d}\xi,$$

and

$$\kappa(x;t,a)=\frac{t}{A(\omega)}\int_{-1}^{1}\omega(\xi)\frac{2\{1-\cos(at\xi)\}}{t^2\xi^2}\exp\Big(\frac{t^2\xi^2}{2}\Big)\cos(tx\xi)\,\mathrm{d}\xi.$$

In practice, it is convenient to pick ω as either the uniform density or the triangle density, for in both cases $\hat{\omega}$ has an explicit formula. For example, when ω is the uniform density, $K(u)\equiv\hat{\omega}(u)=\sin(u)/u$, $A(\omega)=\pi$ and ψ can be written as

$$\frac{t}{\pi}\int\frac{\sin\{t(u-y)\}}{t(u-y)}\pi(y;a)\,\mathrm{d}y$$

(here $\pi\approx 3.14$ is the Archimedes constant). In Fig. 5, let ω be the uniform density (Figs 5(a) and 5(b)) and the triangle density (Figs 5(c) and 5(d)); we have plotted the function $\pi_0-\psi(u;t,a)$ in example I (Figs 5(a) and 5(b)) and example III (Figs 5(c) and 5(d)). Fig. 5 shows that, with a relatively large t, $\pi_0-\psi$ well approximates the function $\pi_0-\pi(t;a)$.

We conclude this section by commenting on the case where $\Pi_n(\mu;a)\to 0$ algebraically fast (i.e. $\Pi_n(\mu;a)\leqslant O(n^{-c})$ for some constant $c>0$). The theorems above do not apply to this case as $\Pi_n(\mu;a)$ is very small. The difficulty is that, to ensure consistency, we need to construct ψ such that, for all $|u|>a$ and $t>0$, $\psi(u;t,a)\equiv\pi(u;t,a)$. Generally, such a construction is challenging. Take example I for instance; the construction requires that $\psi(u;t,a)\equiv 1$, which implies that $\hat{\psi}$ does not have a compact support (the Heisenberg uncertainty principle (e.g. page 32 of Mallat (1998)). However, the existence of the function κ critically depends on the condition that $\hat{\psi}$ has a compact support. In fact, expression (6.5) can be interpreted as $\kappa*\phi=\psi$, which is equivalent to $\hat{\kappa}=\exp(\xi^2/2)\hat{\psi}$ (recall that the asterisk denotes the usual convolution and that ϕ denotes the density function of $N(0,1)$). Without the compact support of $\hat{\psi}$, the integrability of $\hat{\kappa}$ is difficult to ensure, and so is the existence of κ.

7. Discussion

In this section, we briefly mention the generalization of the approach proposed to non-Gaussian data and data with dependent structures. We also make several concluding remarks.

7.1. Generalization

The approach can be conveniently extended to general location shift families. In fact, consider n independent observations $X_j=\mu_j+\varepsilon_j$, $j=1,\ldots,n$, where $\varepsilon_j\sim^{\mathrm{IID}} f_0$ and all except a small proportion of μ_j are 0; we are interested in estimating this proportion.

Let $A_0(t)$ be the characteristic function that is associated with f_0; then the underlying characteristic function that is associated with the model equals

$$A_0(t)\frac{1}{n}\sum_{j=1}^{n}\exp(\mathrm{i}\mu_j t),$$

which, in a similar fashion, factors into two terms: the amplitude $A_0(t)$ and the (underlying)

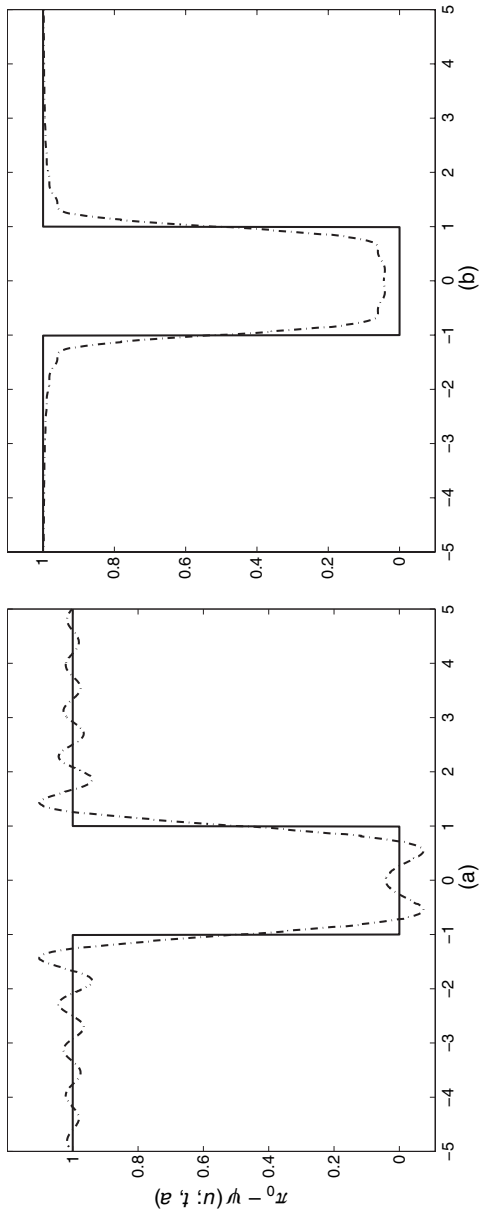

Fig. 5. Constructed functions $\pi_0 - \psi(u; t, a)$ (———, functions $\pi_0 - \pi(u; a)$): (a) uniform density, example I, $a = 1$, $t = 7.5$; (b) uniform density, example I, $a = 1$, $t = 15$; (c) triangle density, example III, $a = p = 1$, $t = 7.5$; (d) triangle density, example III, $a = p = 1$, $t = 15$

phase function $(1/n)\Sigma_{j=1}^{n}\exp(\mathrm{i}\mu_j t)$. Surprisingly, the phase function does not depend on f_0 and is uniquely determined by the mean vector $\mu=\{\mu_1,\ldots,\mu_n\}$. Since the phase function is the key to the approach proposed, we expect that results that are presented in this paper can be extended to general location shift families.

An interesting special case is the Laplace location shift family, in which $f_0(x)=\frac{1}{2}\exp(-|x|)$, and $A_0(t)=1/(1+t^2)$. Similarly, if we define the empirical phase function as

$$\varphi_n(t)=\varphi_n(t;X_1,\ldots,X_n,n)=\int_{-1}^{1}\omega(\xi)\,(1+t^2\xi^2)\,\frac{1}{n}\sum_{j=1}^{n}\cos(tX_j\xi)\,\mathrm{d}\xi,$$

then the empirical phase function and the phase function connect to each other through $E[\varphi_n(t)]=\varphi(t)$. Compared with the Gaussian case, the term $\exp(t^2\xi^2/2)$ is replaced by $1+t^2\xi^2$. When $t\to\infty$, the latter tends to ∞ much more slowly; consequently, the empirical phase function that corresponds to the Laplace family converges to the phase function much faster. In a sense, the Gaussian case is the most difficult case, as the term $\exp(t^2\xi^2/2)$ largely undermines the convergence rate of the empirical phase function.

Our approach can also be conveniently generalized to data with weakly dependent structures. As we mentioned in Section 2.2, the key for the approach proposed to be successful is that $\varphi_n(t;X_1,\ldots,X_n,n)/\varepsilon_n(\mu)\approx\varphi(t;\mu,n)/\varepsilon_n(\mu)$ and $\varphi(t;\mu,n)/\varepsilon_n(\mu)\approx 1$. Note that, first, the second approximation will not be affected by dependence and, second, the accuracy of the first approximation is based on the central limit theorem. Since the central limit theorem holds for many weakly dependent structures, we expect that both approximations continue to be accurate under various weakly dependent structures, and so do the key results in this paper.

7.2. Concluding remarks

We have proposed a general approach to constructing the oracle equivalent to the proportion of non-zero normal means. The oracle equivalent equals the true proportion universally for all dimensions and all normal mean vectors. The construction of the oracle equivalent reduces the problem of estimating the proportion to that of estimating the oracle equivalent. By replacing the underlying phase function with the empirical phase function in the oracle equivalent, we formed a family of estimators. Under mild conditions, these estimators are consistent for the true proportion; uniformly so for a wide class of parameters. The ideas and methods that were presented in this paper can be extended to handle more complicated models. The estimators were also successfully applied to the analysis of microarray data on breast cancer and comparative genomic hybridization data on lung cancer. See Jin and Cai (2007) and Jin *et al.* (2007) for details.

The approach proposed appears to provide new solutions and new opportunities in the field of large-scale multiple testing. As many procedures critically depend on knowledge of the proportion (e.g. the local FDR procedure (Efron *et al.*, 2001), *B*-statistic (Lönnstedt and Speed, 2002), optimal discovery approach (Storey, 2007) and the adaptive FDR approach (Benjamini *et al.*, 2005)), we expect to have better results by combining the estimated proportion with these procedures. Moreover, the approach suggests that Fourier analysis could be a useful tool for solving problems in large-scale multiple testing. In the literature, Fourier analysis has been repeatedly shown to be useful for statistical inference. One example can be found in Fan (1991) and Zhang (1990), where Fourier analysis is shown to be useful in density estimation. Another example can be found in Tang and Zhang (2006, 2007), where Fourier analysis is used to derive FDR controlling procedures (in a way, our approach is related to that in Fan (1991), Zhang (1990) and Tang and Zhang (2006, 2007)). Still another example can be found in Kendall (1974). It is tempting to think that many other seemingly intractable statistical problems can be tackled

by Fourier analysis. We call this *the temptation of the Fourier kingdom* (Mallat, 1998), a kingdom with many sophisticated tools that are ready for use.

Acknowledgements

The author thanks Tony Cai, Bradley Efron, Jianqing Fan, Christopher Genovese, Chong Gu, Mark Low, Jeffrey Scargle, Larry Wasserman, Cun-Hui Zhang, Zepu Zhang and the referees for useful comments and discussion. The author is partially supported by National Science Foundation grants DMS-0505423 and DMS-0639980.

Appendix A: Proofs

A.1. Proof of theorem 1

The first inequality follows directly from that, for all fixed t, μ and n,

$$\varphi(t;\mu,n) \leqslant \frac{1}{n}\sum_{j:\mu_j\neq 0}|1-\cos(\mu_j t)| \leqslant 2\,\varepsilon_n(\mu).$$

For the second inequality, write $\varphi(t;\mu,n)=\varepsilon_n(\mu)\,\mathrm{Ave}_{\{j:\mu_j\neq 0\}}\{1-\cos(t\mu_j)\}$, so it is sufficient to show that, for any $k\geqslant 1$ and $u=(u_1,\ldots,u_k)$, when all entries of u are non-zero, $\sup_t[(1/k)\Sigma_{k=1}^k\{1-\cos(u_j t)\}]\geqslant 1$. To show this, note that, by symmetry and scaling invariance, we can assume that $u_k\geqslant u_{k-1}\geqslant\ldots\geqslant u_1=1$ without loss of generality. Observe that, for any $x>1$,

$$\int_0^x \frac{1}{k}\sum_{k=1}^k\{1-\cos(u_j t)\}\,\mathrm{d}t = x-\frac{1}{k}\sum_{j=1}^k\frac{\sin(u_j x)}{u_j}\geqslant x-1,$$

so

$$\max_{\{0\leqslant t\leqslant x\}}\left[\frac{1}{k}\sum_{k=1}^k\{1-\cos(u_j t)\}\right]\geqslant 1-\frac{1}{x},$$

and the claim follows directly by letting $x\to\infty$.

A.2. Proof of theorem 3

The following lemma is proved in Section 9 of Jin (2007).

Lemma 3. With ψ and κ as defined in equations (2.2) and (2.4) respectively, where ω is eligible, we have

(a) $\psi(0;t)\equiv 0$,
(b) for any t and $X\sim N(u,1)$, $E[\kappa(X;t)]=\psi(u;t)$ and
(c) if additionally ω is good, then, for any t and u, $0\leqslant\psi(u;t)\leqslant 1$.

We now prove theorem 3. Write $\varphi(t;\mu,n)=\varepsilon_n(\mu)\,\mathrm{Ave}_{\{j:\mu_j\neq 0\}}\{1-\psi(\mu_j;t)\}$. For the first claim, by lemma 3, $\lim_{t\to\infty}\{\psi(u;t)\}=0$ for any $u\neq 0$, so $\lim_{s\to\infty}(\sup_{\{|t|>s\}}[\mathrm{Ave}_{\{j:\mu_j\neq 0\}}\{1-\psi(\mu_j;t)\}])=1$, and the first claim follows directly. For the second claim, again by lemma 3, $0\leqslant\psi\leqslant 1$, so we can strengthen the claim of $\lim_{s\to\infty}(\sup_{\{|t|>s\}}[\mathrm{Ave}_{\{j:\mu_j\neq 0\}}\{1-\psi(\mu_j;t)\}])=1$ into the claim of $\sup_{\{|t|>s\}}[\mathrm{Ave}_{\{j:\mu_j\neq 0\}}\{1-\psi(\mu_j;t)\}]=1$ for all $s\geqslant 0$; taking $s=0$ yields the second claim.

A.3. Proof of theorem 4

The key for the proof is the following lemma, which is proved in Section 9 of Jin (2007).

Lemma 4. Consider n independent random variables $X_j=\mu_j+z_j$ where $z_j\sim^{\mathrm{IID}}N(0,1)$; suppose that $(1/n)\Sigma_{j=1}^n|\mu_j|\leqslant r$ for some constant $r>0$. When $n\to\infty$, for any fixed $q>3/2$,

$$P\left[\max_{\{0\leqslant t\leqslant\log(n)\}}\left|\frac{1}{n}\sum_{j=1}^n\{\cos(tX_j)-E[\cos(tX_j)]\}\right|\geqslant\frac{\sqrt{\{2q\log(n)\}}}{n^{1/2}}\right]\leqslant 2\log(n)^2 n^{-2q/3}\{1+o(1)\}.$$

We now proceed to prove theorem 4. For short, write $\varphi_n(t) = \varphi(t; X_1, \ldots, X_n, n)$ and $\varphi(t) = \varphi(t; \mu, n)$. Note that $E[\cos(tX_j)] = \exp(-t^2/2)\cos(t\mu_j)$, so by definition, for $t > 0$,

$$|\varphi_n(t) - \varphi(t)| \leqslant 2\int_0^1 \omega(\xi)\exp\left(\frac{t^2\xi^2}{2}\right)\left|\frac{1}{n}\{\cos(t\xi X_j) - E[\cos(t\xi X_j)]\}\right| d\xi.$$

By lemma 3, for any fixed $q > 3/2$, except for an event with a probability that is distributed as $2\log(n)^2 n^{-2q/3}$,

$$\sup_{\{\mu\in B_n^1(r)\}}\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}|\varphi_n(t) - \varphi(t)| \leqslant \frac{2\sqrt{\{2q\log(n)\}}}{n^{1/2}}\int_0^1 \omega(\xi)\exp\{\gamma\log(n)\xi^2\}\,d\xi.$$

Now, denote $A = \sup_{\{0<\xi<1\}}\{\omega(\xi)\}$ and write $\gamma_n = \gamma\log(n)$ for short. By elementary calculus,

$$\int_0^1 \omega(\xi)\exp\{\gamma\log(n)\xi^2\}\,d\xi \leqslant A\int_0^1 \exp\{\gamma\log(n)\xi^2\}\,d\xi = \frac{A}{2\gamma_n}n^{\gamma}\{1 + o(1)\}.$$

Combining these gives the theorem.

A.4. Proof of theorem 5

For short, write $t_n = \sqrt{\{2\gamma\log(n)\}}$, $\varphi_n(t) = \varphi(t; X_1, \ldots, X_n, n)$, $\varphi(t) = \varphi(t; \mu, n)$ and $\varepsilon_n = \varepsilon_n(\mu)$. Observe that, for any t, we have the triangle inequality

$$\left|\frac{\varphi_n(t)}{\varepsilon_n} - 1\right| \leqslant \left|\frac{\varphi_n(t) - \varphi(t)}{\varepsilon_n}\right| + \left|\frac{\varphi(t)}{\varepsilon_n} - 1\right|.$$

Now, first, using theorem 4, when $n \to \infty$, for any fixed $q > 3/2$, except for an event with an algebraically small probability, there is a generic constant $C = C(q, r; \omega)$ such that

$$\sup_{\{\Theta_n(\gamma,r)\}}\left|\frac{\varphi_n(t_n) - \varphi(t_n)}{\varepsilon_n}\right| \leqslant \sup_{\{\Theta_n(\gamma,r)\}}\left\{\frac{C}{\varepsilon_n\sqrt{\log(n)}n^{1/2-\gamma}}\right\} \leqslant \frac{C}{\sqrt{\log(n)}}.$$

Second, by the definition of φ and ψ,

$$|\varphi(t_n)/\varepsilon_n - 1| = |\mathrm{Ave}_{\{j:\mu_j\neq 0\}}\{\psi(\mu_j; t_n)\}| \leqslant \sup_{\{tu\geqslant\gamma^{1/2}\log\{\log(n)\}\}}|\psi(u; t)|,$$

uniformly for all $\mu \in \Theta_n(\gamma, r)$; note that, by the way that ψ is constructed, the right-hand side of the inequality tends to 0. Plugging these into the triangle inequality gives the theorem.

A.5. Proof of corollary 1

For short, write $\varepsilon_n = \varepsilon_n(\mu)$, $\varphi_n(t) = \varphi(t; X_1, \ldots, X_n)$, $\varphi(t) = \varphi(t; \mu, n)$ and $\Theta_n = \Theta_n(\gamma, r)$. Note that

$$\left|\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}\{\varphi_n(t)\} - \sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}\{\varphi(t)\}\right| \leqslant \sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}|\varphi_n(t) - \varphi(t)|,$$

so it is sufficient to show

(a) $\sup_{\{\Theta_n\}}\{\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}|\varphi_n(t) - \varphi(t)|/\varepsilon_n\} \to 0$ in probability and
(b) $\lim_{n\to\infty}[\sup_{\{\Theta_n\}}|\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}\{\varphi(t)\}/\varepsilon_n - 1|] = 0$.

First, for (a), using theorem 5, when $n \to \infty$, for any fixed $q > 3/2$, except for an event with an algebraically small probability, there is a generic constant $C = C(q, r; \omega)$ such that

$$\sup_{\{\Theta_n\}}\left\{\frac{\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}|\varphi_n(t) - \varphi(t)|}{\varepsilon_n}\right\} \leqslant C\sup_{\{\Theta_n\}}\left[\frac{1}{\varepsilon_n\sqrt{\log(n)}n^{1/2-\gamma}}\right];$$

(a) follows directly. Second, for (b), by theorem 1 and symmetry, $\varphi[\sqrt{\{2\gamma\log(n)\}}] \leqslant \sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}\{\varphi(t)\} \leqslant \varepsilon_n$; hence

$$\sup_{\{\Theta_n\}}\left|\frac{\sup_{\{0\leqslant t\leqslant\sqrt{\{2\gamma\log(n)\}}\}}\{\varphi(t)\}}{\varepsilon_n}-1\right|\leqslant\sup_{\{\Theta_n\}}\left|\frac{\varphi[\sqrt{\{2\gamma\log(n)\}}]}{\varepsilon_n}-1\right|,$$

and (b) follows by similar proofs to that in theorem 5.

A.6. Proof of theorem 6

The proof of the following lemma is similar to that of lemma 4 so we skip it.

Lemma 5. Consider n independent samples $X_j\sim F$ with $E_F|X|^2<\infty$. When $n\to\infty$, there is a constant $C>0$ such that, with overwhelming probability,

$$\max_{\{0\leqslant t\leqslant\log(n)\}}\left|\frac{1}{n}\sum_{j=1}^{n}[\cos(tX_j)-E\{\cos(tX_j)\}]\right|\leqslant C\frac{\sqrt{\log(n)}}{n^{1/2}}.$$

We now proceed to prove theorem 8. As the proofs are similar, we prove only the first claim. Define $\varphi(t;\mu,n,F)=E_F[\varphi_n(t;X_1,\ldots,X_n,n)]$. Using the Fubini theorem,

$$\varphi(t;\mu,n,F)=\int_{-1}^{1}\omega(\xi)\,E\left[1-\exp\left(\frac{t^2\xi^2}{2}\right)\cos(t\xi X_1)\right]\mathrm{d}\xi=\varepsilon_n\int_{-1}^{1}\omega(\xi)\left[\int\{1-\cos(tu\xi)\}\,\mathrm{d}F\right]\mathrm{d}\xi.$$

For short, write $t_n=\sqrt{\{2\gamma\log(n)\}}$, $\varphi_n(t)=\varphi(t;X_1,\ldots,X_n,n)$ and $\varphi(t)=\varphi(t;\mu,n,F)$; by the Fubini theorem, $\varphi(t_n)/\varepsilon_n=\int[\int_{-1}^{1}\omega(\xi)\{1-\cos(t_nu\xi)\}\,\mathrm{d}\xi]\,\mathrm{d}F(u)$. By theorem 2, $\lim_{n\to\infty}[\int_{-1}^{1}\omega(\xi)\{1-\cos(t_nu\xi)\}\,\mathrm{d}\xi]=1$ for any $u\neq0$; using dominant convergence theorem gives $\lim_{n\to\infty}|\varphi(t_n)/\varepsilon_n-1|=0$.

At the same time, by lemma 5 and similar arguments to that in the proof of theorem 4, when $n\to\infty$, there is a constant $C=C(\gamma,\omega,F)$ such that, with overwhelming probability,

$$\left|\frac{\varphi_n(t)-\varphi(t)}{\varepsilon_n}\right|\leqslant n^{\beta}\int_{-1}^{1}\omega(\xi)\exp\left(\frac{t^2\xi^2}{2}\right)\left|\frac{1}{n}\{\cos(t\xi X_j)-E[\cos(t\xi X_j)]\}\right|\leqslant C\frac{n^{\beta+\gamma-1/2}}{\sqrt{\log(n)}}.$$

Since $\gamma+\beta\leqslant\frac{1}{2}$, combining this with $\lim_{n\to\infty}|\varphi(t_n)/\varepsilon_n-1|=0$ gives the theorem.

A.7. Proof of theorem 7

For short, write $t=\sqrt{\{2\gamma\log(n)\}}$, $\varphi_n(t)=\varphi_n(t;X_1,\ldots,X_n)$ and $\varphi(t)=E\{\varphi_n(t)\}$. At the same time, we write $\mathrm{Re}\{\hat{g}(s)\}=\hat{\phi}(s)\,h(s)$. Note that $h(s)$ is a bounded function which tends to 0 as $s\to\infty$. On one hand, by similar arguments to that in the proof of theorem 6, there is a constant $C=C(\gamma,\omega,g)$ such that, with overwhelming probability, $|\{\varphi_n(t)-\varphi(t)\}/\varepsilon_n|\leqslant C/\sqrt{\log(n)}$. On the other hand, direct calculation shows that $|\varphi(t)/\varepsilon_n-1|=|\int\omega(\xi)\,h(t\xi)\,\mathrm{d}\xi|$, where the right-hand side tends to 0 as $n\to\infty$. Since $|\varphi_n(t)/\varepsilon_n-1|\leqslant|\{\varphi_n(t)-\varphi(t)\}/\varepsilon_n|+|\varphi(t)/\varepsilon_n-1|$, the claim follows directly.

A.8. Proof of theorem 8

We employ the theory on the FDR functional that was developed in Donoho and Jin (2006) for the proof. The FDR functional $T_q(\cdot)$ is defined as $T_q(G)=\inf\{t:\bar{G}(t)\geqslant(1/q)\bar{G}_0(t)\}$, where $\bar{G}=1-G$ is any survival function and $\bar{G}_0$ is the survival function of $|N(0,1)|$. Particularly, we have $T_q(G_n)$ and $T_q(G)$, where G_n and G denote the empirical CDF and underlying CDF for $|X_1|,\ldots,|X_n|$ respectively. For any constant $0<c_0<\frac{1}{2}$, corollary 4.2 in Donoho and Jin (2006) can be extended to the current situation and we have $\sup_{\{c_0\leqslant q\leqslant1-c_0\}}|T_q(G)-T_q(G_n)|=O_p(1/n^{1/2})$.

$G(t)$ can be written in the form of $G(t)=(1-\varepsilon)\,G_0(t)+\varepsilon\,H(t)$, where $H(t)$ is the marginal CDF that is associated with the non-null effects. Let q_α be the unique solution of $\alpha=\bar{H}\{T_q(G)\}$. Note that, when $X_j\sim H$, the probability that X_j exceeds $T_{q_\alpha}(G)$ is α. View $T_{q_\alpha}(G)$ as a non-stochastic oracle threshold, and treat X_j as a discovery if and only if it exceeds the threshold; then the resulting total number of true positive discoveries is distributed as binomial$(n\varepsilon,\alpha)$. As a result, the proportion of signals exceeding the threshold $T_q(G)$ tends to α in probability.

At the same time, note that the stochastic threshold in the procedure proposed equals $T_{\hat{q}}(G_n)$. So, to show the theorem, it is sufficient to show that the stochastic threshold converges to the non-stochastic oracle threshold:

$$T_{\hat{q}}(G_n)\to T_{q_\alpha}(G),\qquad\text{in probability.}\tag{A.1}$$

We now show expression (A.1). For short, write $t_0 = T_{q_\alpha}(G)$ and $\hat{t}_n = T_{\hat{q}}(G_n)$. Introduce a bridging quantity $\hat{t}_n^* = T_{\hat{q}}(G)$. By the definition of the FDR functional, it is not difficult to show that there is a constant $c = c(F) \in (0, \frac{1}{2})$ such that, with overwhelming probability, $\hat{q}$ falls in the interval $[c, 1-c]$. Recall that $\sup_{\{c \leqslant q \leqslant 1-c\}} |T_q(G) - T_q(G_n)| = O_p(1/n^{1/2})$; hence

$$|\hat{t}_n^* - \hat{t}_n| \to 0, \qquad \text{in probability.} \tag{A.2}$$

Now, by the way that the procedure is designed, $\#\{\text{total discoveries}\}_{\hat{q}} = n\,\bar{G}_n(\hat{t}_n)$, so $\hat{\varepsilon}_n \alpha = \bar{G}_n(\hat{t}_n)\{1 - (1 - \hat{\varepsilon}_n)\hat{q}_n\}$. Note that

(a) $\sup_t |G_n(t) - G(t)| = O_p(1/n^{1/2})$ by the Dvoretzky–Kiefer–Wolfowitz theorem (Shorack and Wellner, 1986) and
(b) $\hat{\varepsilon}_n / \varepsilon \to 1$ in probability;

combining these with expression (A.2) yields

$$\varepsilon\alpha\{1 + o_p(1)\} = \bar{G}(\hat{t}_n^*)\{1 - (1-\varepsilon)\hat{q}_n\}. \tag{A.3}$$

In addition, observe that, for any $0 < q < 1$ and $t \equiv T_q(G)$, $(1/q)\,\bar{G}_0(t) = \bar{G}(t) = (1-\varepsilon)\,\bar{G}_0(t) + \varepsilon\,\bar{H}(t)$, so $\varepsilon\,\bar{H}(\hat{t}_n^*) = \bar{G}(\hat{t}_n^*)\,\{1 - (1-\varepsilon)\hat{q}\}$; plugging this into equation (A.3) gives $\alpha\{1 + o_p(1)\} = \bar{H}(\hat{t}_n^*)$. Now, comparing $\alpha\{1 + o_p(1)\} = \bar{H}(\hat{t}_n^*)$ with the definition of q_α gives that $|\hat{t}_n^* - t_0| \to 0$ in probability, which, together with expression (A.2), gives expression (A.1).

References

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2006) Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34**, 584–653.

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

Benjamini, Y., Krieger, A. and Yekutieli, D. (2005) Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, **93**, 491–507.

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

Cai, T., Jin, J. and Low, M. (2007) Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, to be published.

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

Donoho, D. and Jin, J. (2006) Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.*, **34**, 2980–3018.

Donoho, D. and Johnstone, I. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

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

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

Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, **19**, 1257–1272.

Ferreira, J. A. and Zwinderman, A. H. (2006) On the Benjamini-Hochberg method. *Ann. Statist.*, **34**, 1827–1849.

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

Ingster, Y. I. (1997) Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Meth. Statist.*, **6**, 47–69.

Ingster, Y. I. (1999) Minimax detection of a signal for l_n^p-balls. *Math. Meth. Statist.*, **7**, 401–428.

James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability* (ed. J. Neyman), vol. 1, pp. 361–379. Berkeley: University of California Press.

Jin, J. (2007) Proportion of nonzero normal means: universal oracle equivalences and uniformly consistent estimators. *Technical Report*. Department of Statistics, Purdue University, West Lafayette. (Available from `www.stat.purdue.edu/~jinj/Research/PNNM-full.pdf`.)

Jin, J. and Cai, T. (2007) Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *J. Am. Statist. Ass.*, **102**, 496–506.

Jin, J., Peng, J. and Wang, P. (2007) Estimating the proportion of non-null effects, with applications to CGH lung cancer data. *Manuscript*.

Kendall, D. G. (1974) Hunting quanta. *Phil. Trans. R. Soc. Lond.* A, **276**, 195–230.

Lönnstedt, I. and Speed, T. (2002) Replicated microarray data. *Statist. Sin.*, **12**, 31–46.

Mallat, S. (1998) *A Wavelet Tour of Signal Processing*, 2nd edn. New York: Academic Press.
Meinshausen, N. and Bühlmann, P. (2005) Lower bounds for the number of false null hypotheses for multiple testing of associations. *Biometrika*, **92**, 893–907.
Meinshausen, M. and Rice, J. (2006) Estimating the proportion of false null hypothesis among a large number of independent tested hypotheses. *Ann. Statist.*, **34**, 373–393.
Schweder, T. and Spjotovoll, E. (1982) Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.
Shorack, G. R. and Wellner, J. A. (1986) Empirical processes with applications to statistics. New York: Wiley.
Smyth, G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Molec. Biol.*, **3**, article 3.
Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.
Storey, J. D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc.* B, **69**, 347–368.
Swanepoel, J. W. H. (1999) The limiting behavior of a modified maximal symmetric 2s-spacing with applications. *Ann. Statist.*, **27**, 24–35.
Tang, W. and Zhang, C.-H. (2006) Bayes and empirical Bayes approaches to controlling the false discovery rate. *Technical Report*. Department of Statistics, Rutgers University, Piscataway.
Tang, W. and Zhang, C.-H. (2007) Empirical Bayes methods for controlling the false discovery rate with dependent data. *IMS Monogr. Ser.*, **54**, 151–160.
Tibshirani, R. (1996) Regression selection and shrinkage via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Wasserman, L. (2006) *All of Nonparametric Statistics*. New York: Springer.
Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J., Lee, N. H., Yeatman, T. J. and Quackenbush, J. (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.*, **3**, 1–12.
Zhang, C.-H. (1990) Fourier methods for estimating mixing densities and distributions. *Ann. Statist.*, **18**, 806–831.