Proportion of Nonzero Normal Means:
Universal Oracle Equivalences and Uniformly Consistent
Estimators

Jiashun Jin
Department of Statistics, Purdue University
Department of Statistics, Carnegie Mellon University
September 3, 2007

Abstract
Since James and Stein’s seminal work (James and Stein, 1961), the problem of estimating \( n \) normal means has received a lot 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 zero and which are not. In this setting, the proportion of the nonzero means plays a key role.

In this paper, motivated by examples in genomics (Efron, 2004) and astronomy (Meinshausen and Rice, 2006), we are particularly interested in estimating the proportion of nonzero means. That is, given \( n \) independent normal random variables with individual means \( X_j \sim N(\mu_j, 1), \, j = 1, \ldots, n \), to estimate the proportion: \( \epsilon_n = \frac{1}{n} \cdot \#\{j : \mu_j \neq 0\} \).

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, while the latter 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 favorably with recent works by Meinshausen and Rice (2006) and Genovese and Wasserman (2004). 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 (Benjamini and Hochberg, 1995) and the local False Discovery Rate approach (Efron et al., 2001).

Keywords: Oracle, Fourier transform, Phase function, Uniform consistency.

1 Introduction

Consider $n$ independent normal random variables

$$X_j = \mu_j + z_j, \quad j = 1, \ldots, n,$$

(1.1)

where $z_j \sim i.d. N(0,1)$ and $\mu_j$ 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 nonzero, 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 means has been extensively studied and well understood. Many modern procedures, e.g. wavelet thresholding procedures (Donoho and Johnstone, 1994) and 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 (fMRI), and image processing. In this field, a problem of major interest is testing $n$ normal means: to determine simultaneously which means are zero 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:

1. Overall testing: Is there any signal at all?
2. Estimating the proportion: How many signals?
3. 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), Efron (2004), Genovese and Wasserman (2004) and Storey (2007). In this paper, we concentrate on the second question: that is, estimating the proportion of signals or, equivalently, the proportion of nonzero means.

1.1 Estimating the proportion of nonzero means: motivations

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

$$\epsilon_n = \epsilon_n(\mu) = \frac{1}{n} \cdot \# \{ j : \mu_j \neq 0 \}.$$

(1.2)

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

Analysis of microarray data on breast cancer. In this example, based on 15 patients diagnosed with breast cancer (7 with the BRCA1 mutation and 8 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 et al., 2004; Jin and Cai, 2007; Storey, 2007). For each gene, a $p$-value was first computed using a two-sample $t$-test and then converted to a z-score. The z-scores can be modeled 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 $\mu_0$ and $\sigma_0$ are called null parameters (Efron, 2004). After the null parameters are estimated and the z-scores are renormalized, the problem of estimating the proportion of differentially expressed genes reduces to the problem of estimating the proportion of nonzero 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, while 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.

**Kuiper Belt Object (KBO).** The Kuiper Belt refers to the region in the Solar system beyond the orbit of Neptune. The Kuiper Belt contains a large unknown number of small objects (i.e. KBOs). The Taiwanese-American Occultation Survey (TAOS) 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 modeled 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 \(j\)-th 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 that it is of interest in 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., 2002), 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, for example, an approach by Meinshausen and Rice (2006) (see also Efron et al. (2001), Genovese and Wasserman (2004), Meinshausen and Bühlmann (2005), 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 hard to check in practice, and is also relatively stringent (see Lemma 4.1). This motivate us to develop a different approach to estimating the proportion.

In this paper, we concentrate on the problem of estimating the proportion of nonzero 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 one is able to manipulate the data set by amplifying every signal component (i.e. \(\mu_j\)) by an arbitrary nonzero factor but keeping the corresponding noise component (i.e. \(z_j\)) untouched. In this scenario, the proportion of nonzero means remains the same, although the data set has been dramatically changed. We call this the property of *scaling invariance*: the proportion of nonzero means remains the same if we multiply each entry of the mean vector by an arbitrary nonzero constant individually.

Unfortunately, the approaches introduced in Meinshausen and Rice (2006) and Swanepoel (1999) (also those in Efron et al. (2001), Genovese and Wasserman (2004), Meinshausen and Bühlmann (2005), 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 based on Fourier transform coefficients, instead of moments or the data tail, for the estimation.

For illustration, suppose out of \(n\) normal means, a proportion of \(\epsilon_0\) has a common positive mean \(\mu_0\), and all others have mean 0. Denote \(i = \sqrt{-1}\), introduce \(\frac{1}{n} \sum_{j=1}^{n} e^{itX_j}\) which we call the empirical characteristic function. If we neglect stochastic fluctuations, the empirical character-
istic function reduces to its own mean which we call the *underlying characteristic function*. The underlying characteristic function is seen to be \( e^{-t^2/2} [1 - e^{it\mu_0}] \), which naturally factors into two components: the *amplitude* \( A(t) = A(t; \epsilon_0, \mu_0) \equiv e^{-t^2/2} \) and the *phase* \( \phi(t) = \phi(t; \epsilon_0, \mu_0) \equiv [1 - e^{it\mu_0}] \). Notice that only the phase contains relevant information of \( \epsilon_0 \), with \( \mu_0 \) playing the role of a nuisance parameter. A convenient way to get rid of the nuisance parameter is to maximize the phase over all frequencies: \( \frac{1}{2} \sup_t |\phi(t; \epsilon_0, \mu_0) - 1| = \frac{\epsilon_0}{2} \sup_t \{1 - e^{it\mu_0}\} \equiv \epsilon_0 \), 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 \( \phi_n(t) \) and \( \phi(t) \), respectively:

\[
\phi_n(t) = \phi_n(t; X_1, \ldots, X_n, n) = \frac{1}{n} \sum_{j=1}^{n} [1 - e^{itX_j}], \\
\phi(t) = \phi(t; \mu, n) = \frac{1}{n} \sum_{j=1}^{n} (1 - \cos(t\mu_j)).
\]

(1.3)

Here we only use the real parts of the phase functions. Because we will see soon that the real parts alone yield desired estimators for the proportion, we drop the imaginary parts everywhere for simplicity. We call (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 1.1** With \( \epsilon_n(\mu) \) defined in (1.2) and \( \phi(t; \mu, n) \) defined in (1.4), for any \( \mu \) and \( n \geq 1 \), we have \( \frac{1}{2} \sup_{\{t\}} \{\phi(t; \mu, n)\} \leq \epsilon_n(\mu) \leq \sup_{\{t\}} \{\phi(t; \mu, n)\} \).

Theorem 1.1 is proved in Section 8. 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 will 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 \( \mu \) (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 by Benjamini and Hochberg (1995), as well as with several other recent approaches in large-scale multiple testing. Section 4 compares the proposed approach with that by 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 \( p \)-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. Section 8 contains proofs of main theorems and corollaries in this paper. Proofs for Theorems 6.1–6.2 and all lemmas are omitted in this paper but can be found in Sections 8–9 of Jin (2007).
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 introduced in Section 1 to construct oracle equivalences of the proportion. Consider the phase function in the following form:

\[ \varphi(t) = \varphi(t; \mu, n) = \frac{1}{n} \sum_{j=1}^{n} [1 - \psi(\mu_j; t)], \quad (2.5) \]

where \( \psi \) is a function we hope to construct such that (a) for any \( t \), \( \psi(0; t) = 1 \); (b) for any fixed \( u \neq 0 \), \( \lim_{u \to -\infty} \psi(u; t) = 0 \); and (c) \( \psi(u; t) \geq 0 \) for all \( u \) and \( t \). To distinguish from \( \mu \), we use \( u \) to denote a scalar quantity here.

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

The intuition behind the construction is that, together, (a) and (b) ensure that the individual index function \( 1_{(\mu_j \neq 0)} \) 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 else but the average of all individual functions \( 1 - \psi(\mu_j; t) \).

We now describe the construction of \( \psi \). Notice that the cosinusoid construction \( \psi(u; t) = \cos(ut) \) clearly does not satisfy (b), as the cosinusoid does not damp to 0 point-wisely. However, on a second thought, we notice 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 2.1 (Riemann-Lebesgue).** If \( \omega \in L^1(\mathbb{R}) \), then \( \lim_{u \to -\infty} \left( \int_{-\infty}^{\infty} \omega(\xi) \cos(t\xi) \, d\xi \right) = 0 \).

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 \( \Xi \) on \((-1, 1)\) that has a symmetric, bounded, and continuous density function \( \omega(\xi) \). Let

\[ \psi(u; t) = E[\cos(u\Xi t)] = \int_{-1}^{1} \omega(\xi) \cos(u\xi t) \, d\xi, \quad \forall \ t, \ u. \quad (2.6) \]

By the Riemann-Lebesgue Theorem, this construction satisfies both (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 \((0, 1)\) **super-additive** if \( f(\xi_1) + f(\xi_2) \leq f(\xi_1 + \xi_2) \) for any \( 0 < \xi_1, \xi_2 < 1 \) and \( \xi_1 + \xi_2 < 1 \). We call a density function \( \omega \) over \((-1, 1)\) **eligible** if it is symmetric, bounded, and continuous. We call \( \omega \) **good** if additionally \( \omega(\xi) = g(1 - \xi) \) for some convex and super-additive function \( g \) over \((0, 1)\).

It is proved in Lemma 8.1 that (c) is satisfied if \( \omega \) is good.

Finally, the only unfinished step is to find an empirical phase function that naturally connects to the phase function in (2.5) by taking the expectation. Comparing with (1.3)–(1.4), we define
the empirical phase function by:

\[ \varphi_n(t; X_1, \ldots, X_n, n) = \frac{1}{n} \sum_{j=1}^{n} [1 - \kappa(X_j; t)], \quad (2.7) \]

where

\[ \kappa(x, t) = \int_{-1}^{1} \omega(\xi)e^{i \frac{\alpha^2}{2} \cos(tx)}d\xi. \quad (2.8) \]

It is proved in Lemma 8.1 that, when \( X \sim N(u, 1), E[\kappa(X; t)] = \int_{-1}^{1} \omega(\xi)\cos(ut\xi) d\xi \equiv \psi(u; t), \) so \( \kappa \) naturally connects to \( \psi \) by taking the expectation. As a result, \( \varphi_n \) connects back to the phase function \( \varphi \) also by taking the expectation: \( E[\varphi_n(t; X_1, \ldots, X_n, n)] = \psi(t; \mu, n) \). This completes the construction.

We now reveal the intuition behind the construction of \( \kappa \). 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. Notice that \( \psi = \tilde{\omega} \) and that \( E[\kappa] = \kappa * \phi \), where \( \phi \) is the density function of \( N(0, 1) \), \( * \) denotes the usual convolution, and \( \tilde{\omega} \) is the Fourier transform of \( \omega \). Under mild conditions, \( \kappa * \phi = \psi \) is equivalent to \( \kappa \tilde{\phi} = \psi \equiv \omega \), so \( \kappa \) should be the inverse Fourier transform of \( \omega/\tilde{\phi} \), which is exactly the same as that in (2.8). We mention here that, while it seems that the choice of \( \omega \) could be arbitrary, it is important to choose \( \omega \) so that \( \omega/\tilde{\phi} \) is integrable, and its inverse Fourier transform exists. A convenient sufficient condition is that \( \omega \) has a compact support.

The construction above indeed yields a family of oracle equivalences. The following theorem is proved in Section 8.

**Theorem 2.2** Fix \( n \) and \( \mu \in \mathbb{R}^n \), let \( \varphi \) and \( \psi \) be defined as in (2.5) and (2.6), respectively.

If \( \omega \) is eligible, then \( \epsilon_n(\mu) = \inf_{\{s \geq 0\}} \{\sup_{\{|u| > s\}} \varphi(t; \mu, n)\} \). If additionally \( \omega \) is good, then \( \epsilon_n(\mu) = \sup_{\{|t| \leq 1\}} \{\varphi(t; \mu, n)\} \).

We conclude this section by giving some examples of \( \omega \).

**Example A.** (Triangle family). \( \omega(\xi) = \frac{\alpha^\alpha}{\Gamma(\alpha+1)} (1 - |\xi|)^{\alpha} \). When \( \alpha = 1 \), \( \omega(\xi) \) is the well-known triangle density function, hence the name triangle family. Clearly, \( \omega \) is eligible for all \( \alpha > 0 \) and is good for all \( \alpha \geq 1 \). Moreover, \( \kappa(x; t) = (\alpha + 1) \int_{0}^{1} (1 - \xi)^{\alpha} e^{-\frac{\alpha^2}{2} \cos(tx)} d\xi, \) and \( \psi(u; t) = (\alpha + 1) \int_{0}^{1} (1 - \xi)^{\alpha} \cos(tu) d\xi \). 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 \( \lim_{u \to 0} \psi(u, t) \). See Figure 1 for the plot of \( 1 - \psi(u; t) \) with \( \alpha = 1, 2 \). The plot illustrates that, for a moderately large \( t \), the index function \( 1_{\{u \neq 0\}} \) can be well approximated by \( 1 - \psi(u; t) \).

**Example B.** (Uniform). \( \omega(\xi) \) is the uniform density function over \((-1, 1)\). In this example, \( \kappa(x; t) = \int_{0}^{1} e^{-\frac{\alpha^2}{2} \cos(tx)} d\xi, \) and \( \psi(u; t) = \sin(tu)/(tu) \). Similarly, the value of \( \psi(0; t) \) is set to \( \lim_{u \to 0} \psi(u, t) \). Notice here that \( \omega \) is eligible but is not good.

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

Interestingly, the cosine rotating family (1.4) can also be thought of as a special case of our construction in (2.5), where the random variable \( \Xi \) does not have a density function. Instead, \( \Xi \) 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 hope to choose an appropriate \( t \) such that

\[ \varphi_n(t; X_1, \ldots, X_n, n)/\epsilon_n(\mu) \approx \varphi(t; \mu, n)/\epsilon_n(\mu) \approx 1. \quad (2.9) \]

There is a tradeoff in the choice of \( t \). When \( t \) increases from 0 to \( \infty \), the second approximation becomes increasingly accurate, but at the same time the variance of \( \varphi_n \) increases, so the first approximation becomes increasingly unstable. It turns out that the right choice of \( t \)
Figure 1: Function $1 - \psi(u; t)$ with $t = 5$. Left: $\omega$ are densities of the triangle family with $\alpha = 1$ (top) and $\alpha = 2$ (bottom) respectively. Right: $\omega$ is the uniform density. In both panels, together, the horizontal line (excluding the point at $(0,1)$) and the dot at the origin stand for the index function $1_{\{u \neq 0\}}$.

is in the range of $O(\sqrt{\log n})$, so for convenience, we consider the following family of estimators: $\varphi_n(t_n(\gamma); X_1, \ldots, X_n, n)$, where

$$t_n(\gamma) = \sqrt{2\gamma \log n}, \quad 0 < \gamma \leq 1/2.$$  \hspace{1cm} (2.10)

We now discuss when the approximations in (2.9) are accurate. Consider the second one first. In order 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 $\varphi_n$ is of the order of $n^{-1/2}$, based on the following theorem which is proved in Section 8.

Theorem 2.3 Let $\varphi(t; \mu, n)$ and $\varphi_n(t; X_1, \ldots, X_n, n)$ be constructed as in (2.5) and (2.7) with an eligible density $\omega$. When $n \to \infty$, for any fixed $q > 3/2$ and $0 < \gamma \leq \frac{1}{3}$, there is a constant $C = C(q, \gamma, \omega)$ such that, except for an event having probability $\sim 2 \cdot \log^2(n) \cdot n^{-2q/3}$,

$$\sup_{\{\mu \in B^1_n(r)\}} \sup_{0 \leq t \leq \sqrt{2\gamma \log n}} \left| \varphi_n(t; X_1, \ldots, X_n, n) - \varphi(t; \mu, n) \right| \leq C \cdot \log^\frac{1}{2} (n) \cdot n^{\gamma - 1/2},$$

where $B^1_n(r) = \{\mu \in \mathbb{R}^n : \frac{1}{n} \sum_{j=1}^n |\mu_j| \leq r\}$ is the $\ell^1$-ball in $\mathbb{R}^n$ with radius $r > 0$.

The accuracy of the first approximation in (2.9) now depends on the magnitude of the true proportion. In the literature, the magnitude of the proportion is modeled through the concept of sparsity (e.g. Abramovich et al. (2006)). We list three different regimes of sparsity:

Relatively dense regime. The proportion is small (e.g. $\epsilon_n = 10\%$), but does not tend to 0 as $n \to \infty$. See Efron et al. (2001) and Genovese and Wasserman (2004).

Moderately sparse regime. The proportion tends to 0 as $n \to \infty$, but does so slower than $1/\sqrt{n}$ does, e.g. $\epsilon_n = n^{-\beta}$ with $0 < \beta < 1/2$. See for example Section 3.1 of Meinshausen and Rice (2006).
Very sparse regime. The proportion tends to 0 faster than $1/\sqrt{n}$ does, e.g. $\epsilon_n = n^{-\beta}$ with $1/2 < \beta < 1$. 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, in order that the first approximation in (2.9) 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

$$\epsilon_n(\mu) \geq n^{\gamma-1/2}. \quad (2.13)$$

In summary, together, (2.12) and (2.13) give a sufficient condition for the consistency of the proposed estimators. Inspired by this, we introduce the following set of parameters:

$$\Theta_n(\gamma, r) = \{\mu \in B^1_n(r), \min_{j: \mu_j \neq 0} \{|\mu_j|\} \geq \frac{\log \log n}{\sqrt{2 \log n}}, \epsilon_n(\mu) \geq n^{\gamma-1/2}, \quad r > 0. \quad (2.14)$$

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

**Theorem 2.4** Let $\Theta_n(\gamma, r)$ be defined as in (2.14), and $\varphi_n(t; X_1, \ldots, X_n, n)$ be defined as in (2.7) where the density $\omega$ is eligible. When $n \to \infty$, for any fixed $0 < \gamma \leq 1/2$, except for an event with algebraically small probability, $\lim_{n \to \infty} \{\sup_{\Theta_n(\gamma, r)} |\varphi_n(\sqrt{\frac{2\log n}{\log \log n}}, X_1, \ldots, X_n, n) - 1|\} = 0$.

Here, we say that a probability is algebraically small if it is bounded by $C \cdot n^{-a}$ for some constants $C = C(\gamma, r) > 0$ and $a = a(\gamma, r) > 0$. Theorem 2.4 is proved in Section 8. We mention that Theorem 2.4 is closely related to Theorem 5 of Jin and Cai (2007). In fact, if we take $\omega$ to be the triangle density, then Theorem 2.4 can be thought as a special case of Theorem 5 in Jin and Cai (2007).

Additionally, if $\omega$ is a good density, then $\sup_{\Theta_n(\gamma, r)} |\varphi_n(t; X_1, \ldots, X_n, n)|$ are also consistent. This is the following corollary, which is proved in Section 8.

**Corollary 2.1** Let $\Theta_n(\gamma, r)$ be defined as in (2.14), and $\varphi_n(t; X_1, \ldots, X_n, n)$ be defined as in (2.7) with a good $\omega$. When $n \to \infty$, for any $0 < \gamma \leq 1/2$, except for an event with algebraically small probability, $\lim_{n \to \infty} \{\sup_{\Theta_n(\gamma, r)} |\varphi_n(t; X_1, \ldots, X_n, n) - 1|\} = 0$.

### 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 $\omega$. To do so, a useful strategy is to pre-select a tolerance parameter $\alpha_n$, and pick the largest $t$ such that the standard deviation of the estimator is no larger than $\alpha_n$ (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 $\omega$; and as a bonus, the resulted $t$ 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 2.1** Fix $u$, $t > 0$, and $z \sim N(0, 1)$, $E[\kappa(z + u; t)]^2 \leq E[\kappa(z; t)]^2$. As a result, with $\varphi(t; \mu, n)$ defined as in (2.5) and $\omega$ being an eligible density function, for any fixed $n$ and $\mu$, $\frac{1}{n} \frac{\epsilon_n(\mu)}{\epsilon_n(\mu)} \frac{\Var(\kappa(z; t))}{\Var(\varphi_n(t; X_1, \ldots, X_n, n))} \leq \frac{1}{\epsilon_n(\mu)} \frac{\Var(\kappa(z; t))}{\Var(\varphi_n(t; X_1, \ldots, X_n, n))}$.

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

By Lemma 2.1, the variance of $\varphi_n(t)$ is no greater than $\frac{1}{n} (\Var(\kappa(z; t)) + 1)$, which can be conveniently calculated once $\omega$ is given. Consequently, if we set $t = t_n^\star(\alpha_n; \omega)$, where

$$t_n^\star(\alpha_n; \omega) = \max\{t : \frac{1}{n} \cdot (\Var(\kappa(z; t)) + 1) \leq \alpha_n^2\}, \quad (2.15)$$

then the standard deviation of $\varphi_n(t)$ is no greater than $\alpha_n$. This is the following theorem, which follows directly from Lemma 2.1.
Theorem 2.5 Let \( \varphi_n(t; X_1, \ldots, X_n, n) \) be defined as in (2.7) and \( t_n^*(\alpha_n; \omega) \) be defined as in (2.15), where \( \omega \) is eligible. We have \( \sup_{\mu \in \mathbb{R}^r} \{ \text{Var}(\varphi_n(t_n^*(\alpha_n; \omega); X_1, \ldots, X_n)) \} \leq \alpha_n^2. \)

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

Next, note that the faster that \( \alpha_n \) tends to 0, the slower that \( t_n^*(\alpha_n; \omega) \) tends 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 \( \alpha_n \) tends to 0 *slowly enough*. For example, to ensure the uniform consistency for all \( \mu \in \Theta_n(\gamma, r) \), one needs that \( \alpha_n \) tends to 0 slowly enough so that \( t_n^*(\alpha_n; \omega) \geq c_0 \sqrt{\log n} \) 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 \( t_n^*(\alpha_n; \omega) \geq c_0 \sqrt{\log n} \) can be checked before we implement the procedure. The proof of the following theorem is similar to that of Theorem 2.4 so is omitted.

Theorem 2.6 Fix a constant \( c_0 > 0 \), let \( \varphi_n(t; X_1, \ldots, X_n, n) \) be defined as in (2.7) and \( t_n^*(\alpha_n; \omega) \) be defined as in (2.15), where \( \omega \) is eligible. When \( n \to \infty \), if \( \alpha_n \to 0 \) slowly enough such that \( t_n^*(\alpha_n; \omega) \geq c_0 \sqrt{\log n} \), then uniformly for all \( \mu \in \Theta_n(\gamma, r) \), \( \varphi_n(t_n^*(\alpha_n; \omega); X_1, \ldots, X_n)/\epsilon_n(\mu) \) 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 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 \( \mu_j \) and variance \( \sigma_j^2 \). In the terminology of multiple testing, we assume that \( (\mu_j, \sigma_j) = (0, 1) \) if \( X_j \) is a *null effect* and \( \neq (0, 1) \) if \( X_j \) is a *non-null effect*. The proportion of signals is then the proportion of non-null effects:

\[
\epsilon_n = \frac{\# \{ j : (\mu_j, \sigma_j) \neq (0, 1) \} }{n}.
\]

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 (CGH) data on lung cancer (Jin et al., 2007). The new approaches compare favorably 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 proposed in previous sections generally have a standard deviation no less than \( 1/\sqrt{n} \), one should not expect them to be consistent in the very sparse case, where the true proportion is much smaller than \( 1/\sqrt{n} \). 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 nonzero \( \mu_j \)’s tend to \( \infty \) uniformly. In fact, Donoho and Jin (2004) considered a setting where \( X_j \) are modeled as samples from a two-component Gaussian mixture \( (1-\epsilon_n)N(0, 1) + \epsilon_nN(\mu_n, 1) \), where \( \epsilon_n = n^{-\beta} \) with \( \beta \in (1/2, 1) \), 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 \( \beta - 1/2 \) when \( \beta \in (1/2, 3/4] \) and equals \( (1 - \sqrt{1 - \beta})^2 \) when \( \beta \in (3/4, 1) \). 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 \( \epsilon_n \) equals 0 or not (i.e. any test would have a sum of Type I and Type II errors tends to 1). Consequently, no estimator could
be consistent for the proportion. This shows that, in the very sparse case, the proportion may be not estimable even when all signals tend to \( \infty \).

While 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) develop a family of new estimators called the CJL lower bounds. At any specified level \( \alpha \in (0, 1) \), the CJL 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 sub-optimal rate of convergence. Interesting progress is 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 \( X_j \) has a common unknown variance \( \sigma^2 \). This setting can be viewed as a special case of that studied in Section 3 of Jin and Cai (2007) if we set \( \mu_0 \) to 0 and assume homoscedasticity; see details therein. We remark that, while Theorem 6 of Jin and Cai (2007) has been focused on the special case where \( \omega \) is the triangle density, it can be generalized to handle the cases where \( \omega \) is only assumed to be eligible. For reasons of space, 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 the following. Pick \( \epsilon \in (0, 1) \) and a marginal cdf \( F \) with no mass at 0. For each \( j = 1, \ldots, n \), we flip a coin with probability \( \epsilon \) of landing head. When the coin lands tail, we draw an observation \( X_j \) from \( N(0, 1) \). When the coin lands head, we draw an observation \( \mu_j \) 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:

\[
(1 - \epsilon)\phi(x) + \epsilon \int \phi(x - u)\,dF(u).
\]

Here \( \phi \) is the density function of \( N(0, 1) \), \( \epsilon \) can be thought of as the proportion of nonzero normal means. We assume that \( F_F \{ u \neq 0 \} = 1 \).

We now extend the results in Section 2.2 to Model (3.1). First, we discuss the moderately sparse case by calibrating \( \epsilon_n \) with \( \epsilon_n = n^{-\beta} \). The following theorem is proved in Section 8.

**Theorem 3.1** Fix \( F \), \( 0 < \beta < \frac{1}{2} \), and \( 0 < \gamma \leq \frac{1}{2} - \beta \), let \( \epsilon_n \) be \( n^{-\beta} \) and \( \varphi_n(t; X_1, \ldots, X_n, n) \) be defined as in (2.7). When \( n \to \infty \), if \( F \) has a finite second moment and \( \omega \) is eligible, then

\[
\sup_{\epsilon_n, \epsilon \in (0, 1)} \left\{ \epsilon_n \varphi_n(t; X_1, \ldots, X_n, n) \right\} \to 1 \quad \text{in probability.}
\]

Second, we consider the relatively dense case by calibrating \( \epsilon \) as a fixed constant. In this case, the estimators are consistent for all \( \gamma \in (0, 1/2) \). This is the following corollary, the proof of which is similar to that of Theorem 3.1 and is omitted.

**Corollary 3.1** Fix \( F \), \( 0 < \epsilon < 1 \), and \( 0 < \gamma \leq 1/2 \), let \( \varphi_n(t; X_1, \ldots, X_n, n) \) be defined as in (2.7). When \( n \to \infty \), if \( F \) has a finite second moment and \( \omega \) is eligible, then

\[
\sup_{\epsilon} \left\{ \epsilon \varphi_n(t; X_1, \ldots, X_n, n) \right\} \to 1 \quad \text{in probability.}
\]

The conditions in Theorem 3.1 and Corollary 3.1 can be relaxed. However, as the Bayesian model is not very different from the frequentist model, we feel it is unnecessary to completely duplicate Theorem 2.4 and Corollary 2.1. The main point here is that, the results under the Bayesian model are stronger and cleaner.
From time to time, one may worry about the Gaussian assumption for the non-null effects. We note here that Theorem 3.1 can be extended to the following non-Gaussian case.

**Theorem 3.2** Fix $0 < \beta < \frac{1}{2}$ and $0 < \gamma \leq \frac{1}{2} - \beta$, let $\epsilon_n$ be $n^{-\beta}$ and $\varphi_n(t; X_1, \ldots, X_n, n)$ be defined as in (2.7), where $\omega$ is eligible. Suppose $X_j \overset{iid}{\sim} (1 - \epsilon_n)\phi(x) + \epsilon_n g(x)$, where $g(x)$ is a density function which has a finite second moment and satisfies that $\lim_{n \to \infty} \text{Re}(\hat{g}(t))/\phi(t) = 0$. Then as $n$ tends to $\infty$, $\frac{\varphi_n(\sqrt{2\pi \log n}; X_1, \ldots, X_n, n)}{\epsilon_n}$ tends to 1 in probability.

Here, $\hat{g}(t)$ is the Fourier transform of $g(x)$ and $\text{Re}(\hat{g}(t))$ denotes its real part. We note here that no Gaussian mixture assumption is made on $g$. Theorem 3.2 is proved in Section 8.

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

### 3.1 Connection with FDR-controlling procedures

A strategy in microarray analysis is to first identify a subset of genes for follow-up study (Smyth, 2004), and then focus on this subset in subsequent experiments. In the current setting, a natural approach to the problem is to find the largest threshold $t_n = t_n(X_1, \ldots, X_n)$ such that the subset $\{X_j : |X_j| \geq t_n\}$ contains at least $n\epsilon_n\alpha$ signals, where $0 < \alpha < 1$ (e.g. $\alpha = 95\%$). Note here that the total number of signals in the data set is $n\epsilon_n$. By combining the proposed estimators with the recent False Discovery Rate (FDR) procedure by Benjamini and Hochberg (1995), we are able to 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 $t_q = t_q^{FDP}(X_1, \ldots, X_n)$, 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 FDR procedure controls the false discovery rate (FDR), which is the expected value of the false discovery proportion (FDP) (Genovese and Wasserman, 2004): $FDP_q = \frac{\#(\text{Falsely Discoveries})_q}{\#(\text{Total Discoveries})_q}$. The following theorem is proved in Benjamini and Yekutieli (2005) and Ferreira and Zwinderman (2006).

**Theorem 3.3** For any $0 < q < 1$, let $FDP_q$ be the false discovery proportion obtained by implementing the Benjamini and Hochberg’s FDR procedure to Model (3.1), then for any $\mu$ and $n \geq 1$, $E[FDP_q] = (1 - \epsilon_n)q$.

We now combine the proposed approach with Benjamini and Hochberg’s FDR procedure to tackle the problem mentioned earlier in this subsection. Viewing Theorem 3.3 from a different perspective, we have $\#(\text{True Discoveries})_q \approx \#(\text{Total Discoveries})_q \cdot [1 - (1 - \epsilon_n)q]$.

Notice that the number of total discoveries is observable, so in order to obtain $n\epsilon_n\alpha$ true discoveries out of all discoveries, we should pick $q$ such that $\#(\text{Total Discoveries})_q \cdot [1 - (1 - \epsilon_n)q] \approx n\epsilon_n\alpha$. This suggests the following procedure:

**Step 1.** Estimate $\epsilon_n$ by any of the proposed procedures, denote the estimation by $\hat{\epsilon}_n$ (e.g. $\hat{\epsilon}_n = \varphi_n(\sqrt{\log n}; X_1, \ldots, X_n, n)$ or $\hat{\epsilon}_n = \sup_{0 \leq t \leq \sqrt{\log n}} \varphi_n(t; X_1, \ldots, X_n, n)$ when $\omega$ is good).

**Step 2.** Solve for $q$ from the equation $\#(\text{Total Discoveries})_q = \alpha \cdot n \cdot \hat{\epsilon}_n/[1 - (1 - \epsilon_n)q]$, where for any $0 < q < 1$, $\#(\text{Total Discoveries})_q$ is obtained by implementing the Benjamini and Hochberg’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 resulted true positives (i.e. $H_j$ that are correctly rejected) by

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

(3.2)

Though $\hat{T}_n(\alpha, \hat{\epsilon}_n)$ is a random quantity that is not directly observable and does not have an explicit formula, it equals $n\epsilon_n\alpha$ approximately for large $n$. Consequently, in the resulted set of discoveries in Step 3, about $n\epsilon_n\alpha$ discoveries are true positives. Take $\alpha = 0.95$ for example, the resulted set contains about 95% of all true positive in the original set! This is the following theorem, which is proved in Section 8.
Theorem 3.4 Fix $0 < \alpha < 1$, $0 < \epsilon < 1$, and $F$, let $\hat{T}_n(\alpha, \hat{\epsilon}_n)$ be defined as in (3.2) where $\hat{\epsilon}_n$ is consistent for $\epsilon$. When $n \to \infty$, $\frac{\hat{T}_n(\alpha, \hat{\epsilon}_n)}{n \epsilon} \to \alpha$ in probability.

3.2 Connection with other procedures in large-scale multiple testing

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

The proposed procedure is intellectually connected with the Optimal Discovery approach by Storey (2007), as well as the local FDR approach by Efron et al. (2001). Storey noticed that by controlling the expected fraction of false positives, the optimal approach to obtaining the largest expected number of true positives is to utilize an oracle which he called the optimal discovery function. Under the current model, the optimal discovery function can be written as $\text{OD}(x) = 1 - \frac{(1-\epsilon)\phi(x)}{(1-\epsilon)\phi(x) + \epsilon \phi(x-\epsilon)DF(x)}$. Notice 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 $1 - \epsilon$. We thus expect to see better results by combining the proposed approach with the Optimal Discovery approach.

The proposed approach is also intellectually connected with the B-statistic by Lönnstedt and Speed (2002), which was proposed for analyzing microarray data. As mentioned in Lönnstedt and Speed (2002), the implementation of the B-statistic depends on the knowledge of $\epsilon_n$: “... one drawback in using B is that we need a value for the prior proportion of differentially expressed genes ...”. Combining the proposed approach 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 more or less depend on the proportion, e.g. Benjamini et al. (2005). We expect the estimated proportion to be helpful in implementing these procedures.

4 Comparison to Meinshausen and Rice’s estimator

Recently, there have been a number of interesting approaches to the problem of estimating the proportion: among them are the work by Meinshausen and Rice (2004) (see also Efron et al. (2001), Genovese and Wasserman (2004), Meinshausen and Bühlmann (2005)). These procedures are intellectually connected with each other, so we only discuss that in Meinshausen and Rice (2006).

Meinshausen and Rice consider a setting in which they test $n$ uncorrelated null hypotheses $H_j$, $j = 1, \ldots, n$. Associated with the $j$-th 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 following estimator: $\epsilon_n^{MR} = \frac{\sup_{0 \leq t < 1} \frac{F_n(t) - t - \beta_n \delta(t)}{1-t}}{\epsilon \delta(t)}$, where $F_n(t)$ is the empirical cdf of the $p$-values and $\beta_n \delta(t)$ is the so-called bounding function (Meinshausen and Rice, 2006). They have studied various aspects of the estimator including its consistency. In fact, by modeling the $p$-values as samples from a two-component mixture: $p_j^{\text{id}} \sim (1 - \epsilon) \cdot U(0, 1) + \epsilon \cdot h$, $j = 1, \ldots, n$, they found that, in order for the estimator to be consistent, it is necessary that

$$\text{essinf}_{0 < p < 1} h(p) = 0.$$  \hfill (4.1)

Here $\epsilon$ 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), where the authors referred to densities satisfying (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 $X_j = \Phi^{-1}(p_j)$, $j = 1, \ldots, n$. It follows that $X_j$ are samples from the density
$(1 - \epsilon)\phi(x) + \epsilon g(x)$, where $g(x) = h(\hat{\Phi}(x))\phi(x)$. Here, $\hat{\Phi}$ and $\phi$ denote the survival function and the density function of the standard normal, respectively. Accordingly, the purity condition (4.1) is equivalent to:
\[
\text{essinf}_{-\infty < x < \infty} \left\{ g(x)/\phi(x) \right\} = 0,
\]
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 satisfied for Gaussian location mixtures.

**Lemma 4.1** Suppose $g(x) = \int \phi(x - u) dF(u)$ for some distribution function $F$. If $P_F\{u < 0\} \neq 0$ and $P_F\{u > 0\} \neq 0$, then $\text{essinf}_{-\infty < x < \infty} \left\{ g(x)/\phi(x) \right\} > 0$. If $F$ is also symmetric, then $\text{essinf}_{-\infty < x < \infty} \left\{ g(x)/\phi(x) \right\} = \int e^{-\frac{x^2}{2}} dF(u) > 0$.

The proof of Lemma 4.1 is elementary so we skip it. Lemma 4.1 implies that Meinshausen and Rice’s estimator (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 will further compare the proposed approach 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 do 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 proposed approaches are consistent if the following holds (Theorem 3.2):
\[
\lim_{t \to \infty} \left( \text{Re}(\hat{g}(t))/\hat{\phi}(t) \right) = 0.
\]
It is interesting to notice that the condition in (4.3) is highly similar to the purity condition in (4.2), and the only major difference is that, the former concerns $g$ and $\phi$ themselves, and the latter concerns their Fourier transforms. In a sense, our findings in this paper complement with those in Genovese and Wasserman (2004) and Meinshausen and Rice (2006). First, we mirror the purity condition originally defined in the spatial domain to its cousin in the frequency domain—the purity condition based on the Fourier transform, or the *FT-purity* condition for short. Second, we develop a class of new estimators and show them to be consistent for the true proportion when the FT-purity condition holds (but the original purity condition may be violated). We mention that the idea here can be largely generalized; see our forthcoming manuscript (Jin *et al.*, 2007) in detail.

To conclude this section, we mention that in order 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 one can not make *any* assumption on $h$ (e.g. the purity condition or the FT-purity condition), it is argued in Genovese and Wasserman (2004) (see also Meinshausen and Rice (2006)) that all one can hope to consistently estimate is the following quantity:
\[
\tilde{\epsilon} = \epsilon \cdot \left( 1 - \text{essinf}_x \left\{ \frac{g(x)}{\phi(x)} \right\} \right),
\]
which we call the *GW 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 proposed approach for different choices of $\omega$, signals, and parameters. Let $\omega$ be any one of the uniform/triangle/smooth density as introduced in Section 2.1, and denote the estimators proposed in Section 2.2 and Section 2.3 by
\[
\hat{\epsilon}_n(\gamma) \equiv \varphi_n(\sqrt{2\gamma \log n}; X_1, \ldots, X_n, n) \quad \text{and} \quad \hat{\epsilon}^*_n(\alpha_n) \equiv \varphi_n(t^*_n(\omega; \alpha_n); X_1, \ldots, X_n, n),
\]
(5.1)
The simulation experiment contains 6 parts, which we now describe. For clarification, we note that except for the last part of the experiment, the variance of \( x_j \) are assumed as known and equal 1.

Experiment (a). We investigate the effect of \( \gamma \) over \( \hat{\epsilon}_n(\gamma) \). Set \( n = 10^5 \), \( \epsilon_n = 0.2 \), and \( \mu_0 = 0.5, 0.75, 1, 1.25 \). For each \( \mu_0 \), we generate \( n(1 - \epsilon_n) \) samples from \( N(0, 1) \), and \( nc_n \) samples from \( N(\mu_j, 1) \); here \( |\mu_j| \) are randomly generated from \( U(\mu_0, \mu_0 + 1) \), and \( \text{sgn}(\mu_j) \) are randomly generated from \( \{-1, 1\} \) with equal probabilities, where \( U \) denotes the uniform distribution, and \( \text{sgn} \) denotes the usual sign function. As a result, the GW lower bound (see (4.4)) equals 0.079, 0.107, 0.132, 0.153 correspondingly. Next, we pick 50 different \( \gamma \) so that \( \sqrt{2\gamma} \) ranges from 0.02 to 1 with an increment of 0.02. We then apply \( \hat{\epsilon}_n(\gamma) \) to the whole sample for each \( \gamma \) and each \( \omega \). Last, we repeat the whole process independently for 100 times. For later references, we refer to samples generated in this way as \textit{samples with signals uniformly distributed (SSUD) with parameters} \( (n, \epsilon_n, \mu_0) \).

The results are summarized in Figure 2. To be brief, we report the results corresponding to the triangle density only. The results suggest the following. First, \( \hat{\epsilon}_n(\gamma) \) monotonically increases as \( \gamma \) increases. The estimator is generally conservative and underestimates the true proportion, but it becomes increasingly closer to the true proportion as \( \gamma \) approaches 0/2. This, together with more empirical studies, suggests that the best choice in this family is \( \hat{\epsilon}_n(1/2) \), and also that the difference between two estimators, \( \hat{\epsilon}_n(1/2) \) and \( \sup_{0<\gamma<1/2} \{\hat{\epsilon}_n(\gamma)\} \), is generally negligible.

Second, if we fix \( \gamma \) and let \( \mu_0 \) increase (so that the strength of the signal increases) then \( \hat{\epsilon}_n(\gamma) \) becomes increasingly accurate and becomes fairly close to the true proportion when \( \mu_0 \geq 1 \). Third, when \( \gamma \) increases, the standard deviations increase as well, which implies that the estimator gets increasingly unstable. However, the standard deviations remain in a smaller magnitude than that of the biases, so the stochastic fluctuation of the estimator is generally negligible. It is interesting to notice that the standard deviations do not respond much to the strength of the signals; they remain almost the same when the signals range from faint to strong.

Experiment (b). We compare the performances of \( \hat{\epsilon}_n(1/2) \) and \( \hat{\epsilon}_n^*(\alpha_n) \). Especially, we investigate how well the standard deviation (SD) of \( \hat{\epsilon}_n^*(\alpha_n) \) is controlled. First, for each of \( n = 10^4 \times 0.5, 1, 2, 4, 8 \), fix \( \epsilon_n = 0.2 \), generate SSUD with parameters \( (n, 0.2, 1) \). Next, we implement \( \hat{\epsilon}_n(1/2) \) and \( \hat{\epsilon}_n^*(\alpha_n) \) to the sample, with \( \alpha_n = 0.015, 0.020, 0.025 \) and \( \omega \) being the uniform/triangle/smooth density. We repeat the whole process for 100 times, and report the results in Table 1. Here, the GW lower bound does not depend on \( n \), and equals 0.132.

The results suggest the following. Firstly, the adaptive approach—\( \hat{\epsilon}_n^*(\alpha_n) \)—gives a tight control on the empirical SD; this property is not assumed by \( \hat{\epsilon}_n(1/2) \). In fact, the empirical SD of \( \hat{\epsilon}_n^*(\alpha_n) \) seldom exceed \( \alpha_n \), and if so, only by a very small amount. This suggest that, as predicted by Theorem 2.5, the empirical SD of \( \hat{\epsilon}_n^*(\alpha_n) \) is nicely bounded by \( \alpha_n \). Additionally,
Table 1: Comparison of standard deviations (SD) and root-mean squared errors (r-MSE) of \( \hat{\epsilon}_n(1/2) \) and \( \hat{\epsilon}_n^*(\alpha_n) \). See Experiment (b) for details.

<table>
<thead>
<tr>
<th>n</th>
<th>0.5 \times 10^4</th>
<th>10^4</th>
<th>2 \times 10^4</th>
<th>4 \times 10^4</th>
<th>8 \times 10^4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\epsilon}_n(1/2) )</td>
<td>SD 0.0878</td>
<td>0.0837</td>
<td>0.0671</td>
<td>0.0682</td>
<td>0.0699</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0906</td>
<td>0.0841</td>
<td>0.0710</td>
<td>0.0725</td>
<td>0.0713</td>
</tr>
<tr>
<td>( \alpha_n = 0.015 )</td>
<td>SD 0.0883</td>
<td>0.0105</td>
<td>0.0127</td>
<td>0.0120</td>
<td>0.0155</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0785</td>
<td>0.0283</td>
<td>0.0128</td>
<td>0.0192</td>
<td>0.0253</td>
</tr>
<tr>
<td>( \alpha_n = 0.020 )</td>
<td>SD 0.0149</td>
<td>0.0164</td>
<td>0.0172</td>
<td>0.0167</td>
<td>0.0208</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0350</td>
<td>0.0167</td>
<td>0.0122</td>
<td>0.0263</td>
<td>0.0302</td>
</tr>
<tr>
<td>( \alpha_n = 0.025 )</td>
<td>SD 0.0214</td>
<td>0.0220</td>
<td>0.0214</td>
<td>0.0214</td>
<td>0.0259</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0231</td>
<td>0.0230</td>
<td>0.0275</td>
<td>0.0312</td>
<td>0.0341</td>
</tr>
<tr>
<td>Triangle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\epsilon}_n(1/2) )</td>
<td>SD 0.0248</td>
<td>0.0206</td>
<td>0.0158</td>
<td>0.0139</td>
<td>0.0144</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0423</td>
<td>0.0391</td>
<td>0.0309</td>
<td>0.0261</td>
<td>0.0253</td>
</tr>
<tr>
<td>( \alpha_n = 0.015 )</td>
<td>SD 0.0054</td>
<td>0.0105</td>
<td>0.0118</td>
<td>0.0122</td>
<td>0.0150</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.1137</td>
<td>0.0529</td>
<td>0.0360</td>
<td>0.0271</td>
<td>0.0253</td>
</tr>
<tr>
<td>( \alpha_n = 0.020 )</td>
<td>SD 0.0145</td>
<td>0.0165</td>
<td>0.0160</td>
<td>0.0171</td>
<td>0.0200</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0566</td>
<td>0.0413</td>
<td>0.0308</td>
<td>0.0257</td>
<td>0.0267</td>
</tr>
<tr>
<td>( \alpha_n = 0.025 )</td>
<td>SD 0.0205</td>
<td>0.0220</td>
<td>0.0201</td>
<td>0.0218</td>
<td>0.0249</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0451</td>
<td>0.0390</td>
<td>0.0300</td>
<td>0.0272</td>
<td>0.0297</td>
</tr>
<tr>
<td>Smooth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\epsilon}_n(1/2) )</td>
<td>SD 0.0199</td>
<td>0.0152</td>
<td>0.0121</td>
<td>0.0092</td>
<td>0.0095</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0334</td>
<td>0.0281</td>
<td>0.0210</td>
<td>0.0149</td>
<td>0.0131</td>
</tr>
<tr>
<td>( \alpha_n = 0.015 )</td>
<td>SD 0.0055</td>
<td>0.0104</td>
<td>0.0121</td>
<td>0.0121</td>
<td>0.0151</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.1105</td>
<td>0.0401</td>
<td>0.0210</td>
<td>0.0133</td>
<td>0.0152</td>
</tr>
<tr>
<td>( \alpha_n = 0.020 )</td>
<td>SD 0.0147</td>
<td>0.0164</td>
<td>0.0163</td>
<td>0.0169</td>
<td>0.0202</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0451</td>
<td>0.0268</td>
<td>0.0182</td>
<td>0.0169</td>
<td>0.0202</td>
</tr>
<tr>
<td>( \alpha_n = 0.025 )</td>
<td>SD 0.0209</td>
<td>0.0220</td>
<td>0.0204</td>
<td>0.0217</td>
<td>0.0251</td>
</tr>
<tr>
<td></td>
<td>r-MSE 0.0324</td>
<td>0.0260</td>
<td>0.0206</td>
<td>0.0219</td>
<td>0.0253</td>
</tr>
</tbody>
</table>

the bound is tight and the empirical SD do not change much for different \( n \) and \( \omega \); except for a few cases, the empirical SD fall between 0.7\( \alpha_n \) and \( \alpha_n \). In contrast, the empirical SD of \( \hat{\epsilon}_n(1/2) \) may fluctuate for more than 7 times across different \( \omega \), and for more than 2 times across different \( n \). Secondly, in terms of the root mean squared errors (r-MSE), the performance of \( \hat{\epsilon}_n^*(\alpha_n) \) is mainly determined by \( \alpha_n \), and different choices of \( n \) and \( \omega \) do not have prominent effects. Lastly, all estimators yield a reasonably good estimate for the true proportion.

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

Experiment (c). We compare \( \hat{\epsilon}_n^*(\alpha_n) \) with Meinshausen and Rice’s estimator (Equation (5) of Meinshausen and Rice (2006)), which we denote by \( \hat{\epsilon}_n^{MR} \). The bounding function \( \beta_{n,\alpha}(\delta(t)) \) is set to \( \sqrt{2(1-t)(\log \log n)/n} \). Fix \( n = 80000 \), \( \epsilon_n = 0.2 \), \( \alpha_n = 0.015 \), and pick \( \mu_0 = 0.5, 0.75, 1, 1.25 \). Correspondingly, the GW lower bound equals 0.079, 0.107, 0.132, 0.153. For each \( \mu_0 \), we generate SSUD with parameters \( 8 \times 10^4, 0.2, \mu_0 \). We then implement \( \hat{\epsilon}_n^*(\alpha_n) \) and \( \hat{\epsilon}_n^{MR} \) to the sample, with \( \omega \) being the triangle density and the smooth density (to save space, we omit the case for the uniform density). We repeat the whole process for 100 times. The results are displayed in Figure 3, they are also summarized in terms of SD and r-MSE in Table 2. The results suggest that the performance of \( \hat{\epsilon}_n^{MR} \) is generally unsatisfactory, and \( \hat{\epsilon}_n^*(\alpha_n) \) behaves much better. In fact, \( \hat{\epsilon}_n^*(\alpha_n) \) is encouragingly accurate when \( \mu_0 \) is greater than 1.

Experiment (d). We continue the study in Experiment (c) by letting the proportion vary from the dense regime to the moderately sparse regime. Fixing \( n = 10^6 \), \( \alpha_n = 0.002 \), for each of
The results suggest that \( \hat{\epsilon}_n \) continues to outperform the procedure by Meinshausen and Rice (2006). The results are summarized in Table 3, where we tabulated the r-MSE of \( \hat{\epsilon}_n \) being the triangle density and the smooth density. We then repeat the whole process for 100 times.

In both cases, the proposed approaches compare proportion is more 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 proposed approaches compare

<table>
<thead>
<tr>
<th>Signal Strength</th>
<th>( \mu_0 = 0.5 )</th>
<th>( \mu_0 = 0.75 )</th>
<th>( \mu_0 = 1 )</th>
<th>( \mu_0 = 1.25 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\epsilon}_n(\alpha_n) ) (( \alpha_n = 0.015 ), triangle)</td>
<td>0.3649</td>
<td>0.2194</td>
<td>0.1139</td>
<td>0.0784</td>
</tr>
<tr>
<td>( \hat{\epsilon}_n(\alpha_n) ) (( \alpha_n = 0.015 ), smooth)</td>
<td>0.2957</td>
<td>0.1375</td>
<td>0.0792</td>
<td>0.0821</td>
</tr>
<tr>
<td>Meinsahausen and Rice ( \hat{\epsilon}_{MR} )</td>
<td>0.8882</td>
<td>0.8383</td>
<td>0.7903</td>
<td>0.7435</td>
</tr>
</tbody>
</table>

Table 2: Root-mean squared errors of \( \hat{\epsilon}_n(\alpha_n) / \epsilon_n \) and \( \hat{\epsilon}_{MR} / \epsilon_n \) for different signal strengths. See Experiment (c) for details.

\( \epsilon_n = n^{-0.1}, n^{-0.2}, n^{-0.3}, n^{-0.4} \) (note \( n^{-0.1} = 0.25, n^{-0.2} = 0.063, n^{-0.3} = 0.016, n^{-0.4} = 0.004 \)), we generate SSUD with parameters \( (10^6, \epsilon_n, 1.25) \). Correspondingly, the GW lower bound equals 0.192, 0.048, 0.012, 0.003. We implement both \( \hat{\epsilon}_n(\alpha_n) \) and \( \hat{\epsilon}_{MR} \) to the whole sample, with \( \omega \) being the triangle density and the smooth density. We then repeat the whole process for 100 times. The results are summarized in Table 3, where we tabulated the r-MSE of \( \hat{\epsilon}_n(\alpha_n) / \epsilon_n \) and \( \hat{\epsilon}_{MR} / \epsilon_n \).

The results suggest that \( \hat{\epsilon}_n(\alpha_n) \) continues to perform well for the moderately sparse case, and continues to outperform the procedure by Meinshausen and Rice (2006).

<table>
<thead>
<tr>
<th>Sparsity level</th>
<th>( \epsilon_n = n^{-0.1} )</th>
<th>( \epsilon_n = n^{-0.2} )</th>
<th>( \epsilon_n = n^{-0.3} )</th>
<th>( \epsilon_n = n^{-0.4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\epsilon}_n(\alpha_n) ) (( \alpha_n = 0.0015 ), triangle)</td>
<td>0.1409</td>
<td>0.1391</td>
<td>0.1533</td>
<td>0.2854</td>
</tr>
<tr>
<td>( \hat{\epsilon}_n(\alpha_n) ) (( \alpha_n = 0.0015 ), smooth)</td>
<td>0.0611</td>
<td>0.0612</td>
<td>0.0903</td>
<td>0.2613</td>
</tr>
<tr>
<td>Meinsahausen and Rice ( \hat{\epsilon}_{MR} )</td>
<td>0.7341</td>
<td>0.7421</td>
<td>0.7697</td>
<td>0.835</td>
</tr>
</tbody>
</table>

Table 3: Root-mean squared errors of \( \hat{\epsilon}_n(\alpha_n) / \epsilon_n \) and \( \hat{\epsilon}_{MR} / \epsilon_n \) for different sparsity levels, where \( n = 10^6 \). See Experiment (d) for details.

Experiment (e). We continue the study in Experiment (c), but with a different \( \epsilon_n \) and a different way to generate nonzero \( \mu_j \)'s. Fix \( n = 80000, \epsilon_n = 0.1, \alpha_n = 0.015 \), and pick \( \sigma_0 = 1, 1.25, 1.50, 1.75 \). For each \( \sigma_0 \), we generate \( n(1 - \epsilon_n) \) samples from \( N(0, 1) \), and \( n\epsilon_n \) samples from \( N(\mu_j, 1) \), where \( \mu_j \) are sampled from \( N(0, \sigma_0^2) \). Correspondingly, the GW lower bound equals 0.029, 0.038, 0.045, 0.050. We implement both \( \hat{\epsilon}_n(\alpha_n) \) and \( \hat{\epsilon}_{MR} \) to the whole sample, with \( \omega \) being the triangle density and the smooth density. We then repeat the whole process for 100 times. The results are shown in Table 4 in terms of SD and r-MSE. In comparison, the nonzero \( \mu_j \)'s in Experiment (c) are bounded away from 0 by a distance of \( \mu_0 \), but in the current case, a certain fraction of nonzero \( \mu_j \)'s is concentrated around 0. We thus expect that the proportion is more 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 proposed approaches compare
favorably with that of Meinshausen and Rice. We mention that, the unsatisfactory behavior of $\hat{\epsilon}_n^{MR}$ is mainly due to its inconsistency in the current setting; tuning the bounding function would not be very helpful.

<table>
<thead>
<tr>
<th>Signal Strength</th>
<th>$\sigma_0 = 1$</th>
<th>$\sigma_0 = 1.25$</th>
<th>$\sigma_0 = 1.5$</th>
<th>$\sigma_0 = 1.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\epsilon}_n^*(\alpha_n)$ ( $\alpha_n = 0.015$, triangle)</td>
<td>0.5804</td>
<td>0.5122</td>
<td>0.4532</td>
<td>0.3998</td>
</tr>
<tr>
<td>$\hat{\epsilon}_n^*(\alpha_n)$ ( $\alpha_n = 0.015$, smooth)</td>
<td>0.5389</td>
<td>0.4674</td>
<td>0.4078</td>
<td>0.3521</td>
</tr>
<tr>
<td>Meishausen and Rice $\hat{\epsilon}_n^{MR}$</td>
<td>0.9247</td>
<td>0.8943</td>
<td>0.8671</td>
<td>0.8371</td>
</tr>
</tbody>
</table>

Table 4: Root-mean squared errors of $\hat{\epsilon}_n^*(\alpha_n)/\epsilon_n$ and $\hat{\epsilon}_n^{MR}/\epsilon_n$. The nonzero $\mu_j$ are Gaussian distributed; see Experiment (e) for details.

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

![Figure 4](image)

Figure 4: The behavior of $\hat{\epsilon}_n^*(\alpha_n)$ (top two rows, with $\omega$ being the triangle density and smooth density, respectively) and $\hat{\epsilon}_n^{MR}$ (bottom row) when the normality assumption is violated. The true proportion is 0.2. The non-null effects were generated from the double exponential distribution. See Experiment (f) for details. From left to right: $\lambda = 1, 2, 3, 4$.

6 Extensions

The proposed approach 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.

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](http://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.

Example II. The sparsity level of the mean vector plays an important role in many inference problems. There are many models for the sparsity level, and the one where the sparsity level is defined as the average $\ell^p$-norm is particularly well-known (e.g. Abramovich et al. (2006)). A successful estimation for the average $\ell^p$-norm $\frac{1}{n} \sum_{j=1}^n |\mu_j|^p$ has potential applications.
Example III. A variant of the functional in Example II is \( \frac{1}{n} \sum_{j=1}^{n} \min \{ |\mu_j|^p, a^p \} \), where \( a > 0 \) is a constant which may depend on \( n \) but not \( j \). This variant is easier to handle, but is still able to 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 \( \approx \sqrt{2 \log n} \) (Shorack and Wellner, 1986), any signals with a magnitude larger than \( \sqrt{2 \log n} \) can be easily estimated individually, so it makes sense to assume that the magnitude of each \( \mu_j \) does not exceed \( \sqrt{2 \log n} \). Consequently, the functional in Example II reduces to the current one.

Motivated by these examples, we introduce a univariate function \( \pi = \pi(u; a) \) over \( \mathbb{R} \) which satisfies (i) \( \pi(u; a) = 0 \) when \( |u| > a \); (ii) \( \pi(u; a) \) is symmetric and continuous over \([-a, a] \); and (iii) \( 0 \leq \pi(u; a) \leq \pi_0 \) and \( \pi_0 > 0 \), where \( \pi_0 = \pi(0; a) \). We are interested in estimating the following functional:

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

where \( \pi_a = \pi(a; a) \). Due to the possible discontinuity of \( \pi \) at \( \pm a \), we use a randomized rule at \( \pm a \) (i.e. \( \pi(a; a) \) equals 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; similar for \( \pi(-a; a) \)). When \( \pi \) is continuous at \( \pm a \), the functional reduces to \( \lim_{n \to \infty} \Pi_n(\mu; a) = \frac{1}{n} \sum_{j=1}^{n} (\pi_0 - \pi(\mu_j; a)) \). The functional includes Example I–III as special cases. In fact, in Example I, \( \pi(u; a) = 1_{|u| \leq a} \), and \( \Pi_n(\mu; a) = \frac{1}{n} \left( \#\{ j : |\mu_j| > a \} + \#\{ j : |\mu_j| = a \} \right) \). In Example III, \( \pi(u; a) = (a^p - |u|^p)^+ \) and \( \Pi_n(\mu; a) = \frac{1}{n} \sum_{j=1}^{n} \min \{ |\mu_j|^p, a^p \} \).

The idea 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 as that of \( 1_{\{u > 0\}} \)). Similarly, we hope to construct a function \( \psi = \psi(u; t, a) \) such that for any fixed \( u \), \( \lim_{n \to \infty} \psi(u; t, a) = 0, \pi_0/2, \) and \( \pi(u; a) \) according to \( |u| > a, |u| = a, \) and \( |u| < a \). Once such a \( \psi \) is 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 \( \mu \), \( \lim_{n \to \infty} \varphi(t; \mu, n, a) = \Pi_n(\mu; a) \), and we expect that consistent estimators of \( \Pi_n(\mu; a) \) can be constructed in a similar fashion as in Section 2.

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

\[
\psi(u; t, a) = K_t(u) * \pi(u; a) = \int_{-a}^{a} K_t(u - y) \pi(y; a) \, dy \tag{6.3}
\]

It is shown in Lemma 8.4 of Jin (2007) that the function \( \psi \) can be equivalently written as

\[
\psi(u; t, a) = \frac{t}{A(\omega)} \int_{-1}^{1} \omega(\xi) \pi(t\xi; a) \cos(\xi) \, d\xi
\]

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

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

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

\[
\kappa(x; t, a) = \frac{t}{A(\omega)} \cdot \int_{-1}^{1} \omega(\xi) \pi(t\xi; a) e^{\frac{x^2}{2}} \cos(tx\xi) \, d\xi \tag{6.4}
\]

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

\[
E[\kappa(X; t, a)] = \psi(u; t, a), \quad \text{where } t > 0 \text{ and } X \sim N(u, 1) \tag{6.5}
\]
Thus if we let the empirical phase function be
\[ \varphi_n(t; X_1, \ldots, X_n, n, a) = \frac{1}{n} \sum_{j=1}^{n} [\pi_0 - \kappa(X_j; t, a)], \quad (6.6) \]
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 \( \pm a \), similar to \( \Theta_n(r, \gamma) \) (see (2.14)), we define the following set of parameters:
\[ \Theta_n^a(r) = \{ \mu \in B_{1}^{1}(r), \text{min}_{1 \leq j \leq n} \{||\mu_j| - a|| \} \geq \frac{\log \log n}{\sqrt{2 \log n}} \}, \quad (6.7) \]
where as before, \( B_{1}^{1}(r) \) is the \( \ell^1 \)-ball in \( \mathbb{R}^n \) with radius \( r \). The following theorem is proved in Section 8 of Jin (2007).

**Theorem 6.2** Fix \( a > 0, r > 0, \) and \( 0 < \gamma \leq 1/2 \), let \( \varphi_n(t; X_1, \ldots, X_n, n, a) \) be defined as in (6.6), where the density \( \omega \) is eligible with \( A(\omega) \not= 0 \), and suppose \( \pi \) is absolutely continuous over \([-a, a]\). When \( m \to \infty, \sup_{\Theta_n^a(r)} \{ ||\varphi_n(\sqrt{2} \log n; X_1, \ldots, X_n, n, a) - \Pi_n(\mu; a)|| \} \to 0 \) in probability. If additionally \( \pi \) is continuous everywhere, then \( \sup_{B_{1}^{1}(r)} \{ ||\varphi_n(\sqrt{2} \log n; X_1, \ldots, X_n, n, a) - \Pi_n(\mu; a)|| \} \to 0 \) in probability.

Again, the condition that all \( \mu_j \) are bounded away from \( \pm a \) by an amount of \( \log \log n / \sqrt{2 \log n} \) can be largely relaxed; we choose \( \Theta_n^a(r) \) only to make the presentation cleaner.

We now continue the discussion of Example I–III. Theorems 6.1–6.2 directly apply to Examples I and III. Moreover, it can be shown that Theorems 6.1–6.2 continue to hold if we take \( a = \sqrt{2 \log n} \) in Example III, so these theorems apply to Example II too. Also, we note that some explicit formulas are available for these examples. In fact, in Example I, \( \hat{\pi}(\xi, a) = 2 \sin(a\xi)/\xi, \psi(u; t, a) = \frac{1}{A(\omega)} \int_{-1}^{1} \omega(\xi) \frac{2 \sin(a\xi)\xi}{t^{2} \xi^{2}} \cos(tx\xi) d\xi, \) and \( \kappa(x; t, a) = \frac{t}{\pi \omega} \int_{-1}^{1} \omega(\xi) \frac{2(1-\cos(a\xi))\xi}{t^{2} \xi^{2}} \cos(tx\xi) d\xi. \)

In Example III, when \( p = 1, \hat{\pi}(\xi; a) = 2(1-\cos(a\xi))/\xi^{2}, \psi(u; t, a) = \frac{1}{A(\omega)} \int_{-1}^{1} \omega(\xi) \frac{2(1-\cos(a\xi))\xi}{t^{2} \xi^{2}} \cos(tx\xi) d\xi, \) and \( \kappa(x; t, a) = \frac{t}{\pi \omega} \int_{-1}^{1} \omega(\xi) \frac{2(1-\cos(a\xi))\xi}{t^{2} \xi^{2}} \cos(tx\xi) d\xi. \)

In practice, it is convenient to pick \( \omega \) as either the uniform density or the triangle density, for in both cases \( \hat{\omega} \) has an explicit formula. For example, when \( \omega \) is the uniform density, \( K(u) \equiv \hat{\omega}(u) = \sin(u)/u, A(\omega) = \pi, \) and \( \psi \) can be written as \( \frac{1}{\pi} \int_{-1}^{1} \sin(t(u-y)) \pi(y; a) dy \) (here \( \pi \approx 3.14 \) is the Archimedes’ constant). In Figure 5, let \( \omega \) be the uniform density (left column) and the triangle density (right column), we plotted the function \( \pi_0 - \psi(u; t, a) \) in Example I (top row) and Example III (bottom row). The figure 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 \( \Pi_n(\mu; a) \) tends to 0 algebraically fast (i.e. \( \Pi_n(\mu; a) \leq O(n^{-c}) \) for some constant \( c > 0 \)). The theorems above do not apply for this case as \( \Pi_n(\mu; a) \) is very small. The difficulty is that, to ensure consistency, we need to construct \( \psi \) 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 (Heisenberg Uncertainty principle (e.g. Page 32 of Mallat (1998))). However, the existence of the function \( \kappa \) critically depends on the condition that \( \hat{\psi} \) has a compact support. In fact, (6.5) can be interpreted as \( \kappa * \phi = \hat{\psi}, \) which is equivalent to \( \hat{\kappa} * \hat{\phi} = \hat{\psi} \) (recall that \( * \) denotes the usual convolution and that \( \phi \) denotes the density function of \( N(0,1) \)). Without the compact support of \( \hat{\psi}, \) the integrability of \( \hat{\kappa} \) is hard to ensure, and so does the existence of \( \kappa. \)

### 7 Discussion

In this section, we briefly mention the generalization of the proposed approach to nonGaussian data and data with dependent structures. We also make several concluding remarks.
The approach can be conveniently extended to general location shift families. In fact, consider
functions $A$ and $f$. Let $\mu$ be the characteristic function associated with the model equals $A(\theta) \cdot \left\{ \frac{1}{n} \sum_{j=1}^{n} e^{i\mu j} \right\}$, which, in a similar fashion, factors into two terms: the amplitude $A(\theta)$ and the (underlying) phase function $\left\{ \frac{1}{n} \sum_{j=1}^{n} e^{i\mu j} \right\}$. Surprisingly, the phase function does not depend on $f$ and is uniquely determined by the mean vector $\mu = \{\mu_1, \ldots, \mu_n\}$. Since the phase function is the key to the proposed approach, we expect that results presented in this paper can be extended to general location shift families.

An interesting special case is the Laplace location shift family, in which $f_0(x) = \frac{1}{2} e^{-|x|}$, and $A_0(t) = 1/(1 + t^2)$. Similarly, if we define the empirical phase function as $\phi_n(t) = \phi_n(t; X_1, \ldots, X_n, n) = \int_{-\infty}^{\infty} \omega(\xi) \cdot (1 + t^2 \xi^2) \cdot \frac{1}{n} \sum_{j=1}^{n} \cos(t X_j \xi) d\xi$, then the empirical phase function and the phase function connect to each other through $E[\phi_n(t)] = \phi(t)$. Compared to the Gaussian case, the term $e^{it^2 \xi^2/2}$ is replaced by $1 + t^2 \xi^2$. Notice that when $t$ tends to $\infty$, the latter tends to $\infty$ much slower; consequently, the empirical phase function 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 $e^{it^2 \xi^2/2}$ 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 proposed approach to be successful is that $\varphi_n(t; X_1, \ldots, X_n, n)/\epsilon_n(\mu) \approx \varphi(t; \mu, n)/\epsilon_n(\mu)$ and $\varphi(t; \mu, n)/\epsilon_n(\mu) \approx 1$. Notice that, first, the second approximation will not be affected by dependency, 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.

7.1 Generalization

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

Let $A_0(t)$ be the characteristic function associated with $f_0$, then the underlying characteristic function associated with the model equals $A_0(t) \cdot \left\{ \frac{1}{n} \sum_{j=1}^{n} e^{i\mu_j t} \right\}$, which, in a similar fashion, factors into two terms: the amplitude $A_0(t)$ and the (underlying) phase function $\left\{ \frac{1}{n} \sum_{j=1}^{n} e^{i\mu_j t} \right\}$. 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 proposed approach, we expect that results presented in this paper can be extended to general location shift families.

Figure 5: Constructed functions $\pi_0 - \psi(u; t, a)$ (dashed curves). By choosing $\omega$ as the uniform density (top row) and the triangle density (bottom row), the plot illustrates $\psi(u; t, a)$ for Example I with $a = 1$ (top row), and Example III with $a = p = 1$ (bottom row). The solid curves are the functions $\pi_0 - \pi(u; a)$. The values of $t$ are 7.5 (left column) and 15 (right column).

7.2 Concluding remarks

We have proposed a general approach to constructing the oracle equivalence of the proportion of nonzero normal means. The oracle equivalence equals the true proportion universally for all dimensions and all normal mean vectors. The construction of the oracle equivalence reduces the
problem of estimating the proportion to that of estimating the oracle equivalence. By replacing the underlying phase function with the empirical phase function in the oracle equivalence, 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 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 CGH data on lung cancer. See Jin and Cai (2007) and Jin et al. (2007) for detail.

The proposed approach appear to provide new solutions and new opportunities in the field of large-scale multiple testing. As many procedures critically depend on the 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 (2006a, 2006b), 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 (2006a, 2006b)). 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 ready for use.

8 Proofs

8.1 Proof of Theorem 1.1

The first inequality follows directly from that for all fixed $t$, $\mu$, and $n$, $\varphi(t;\mu,n) \leq \frac{1}{\pi} \sum_{j: \mu_j \neq 0} |1 - \cos(\mu_j t)| \leq 2\epsilon_n(\mu)$. For the second inequality, write $\varphi(t;\mu,n) = \epsilon_n(\mu) \cdot \text{Ave}_{\{j: \mu_j \neq 0\}} \{1 - \cos(t\mu_j)\}$, so it is sufficient to show that, for any $k \geq 1$ and $u = (u_1, \ldots, u_k)$, when all entries of $u$ are nonzero, $\sup \{ \frac{1}{k} \sum_{k=1}^{k} (1 - \cos(u_j t)) \} \geq 1$. To show this, notice that by symmetry and scaling invariance, we can assume $u_k \geq u_{k-1} \ldots \geq u_1 = 1$ without loss of generality. Observe that for any $x > 1$, $\int_{0}^{x} \frac{1}{k} \sum_{k=1}^{k} (1 - \cos(u_j t)) dt = x - \frac{1}{k} \sum_{j=1}^{k} \frac{\sin(u_j x)}{u_j} \geq x - 1$, so $\max_{0 \leq t \leq x} \{ \frac{1}{k} \sum_{k=1}^{k} (1 - \cos(u_j t)) \} \geq (1 - \frac{1}{x})$, and the claim follows directly by letting $x \to \infty$. \hfill \Box

8.2 Proof of Theorem 2.2

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

**Lemma 8.1** With $\psi$ and $\kappa$ as defined in (2.6) and (2.8) respectively, where $\omega$ 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 $\omega$ is good, then for any $t$ and $u$, $0 \leq \psi(u; t) \leq 1$.

We now prove Theorem 2.2. Write $\varphi(t;\mu,n) = \epsilon_n(\mu) \cdot \text{Ave}_{\{j: \mu_j \neq 0\}} \{1 - \psi(\mu_j; t)\}$. For the first claim, by Lemma 8.1, $\lim_{t \to \infty} \psi(u; t) = 0$ for any $u \neq 0$, so $\lim_{s \to \infty} \left( \sup_{|t| > s} \text{Ave}_{\{j: \mu_j \neq 0\}} \{1 - \psi(\mu_j; t)\} \right) = 1$, and the first claim follows directly. For the second claim, again by Lemma 8.1, $0 \leq \psi \leq 1$, so we can strengthen the claim of $\lim_{s \to \infty} \left( \sup_{|t| > s} \text{Ave}_{\{j: \mu_j \neq 0\}} \{1 - \psi(\mu_j; t)\} \right) = 1$ into the claim of $\lim_{s \to \infty} \left( \sup_{|t| > s} \text{Ave}_{\{j: \mu_j \neq 0\}} \{1 - \psi(\mu_j; t)\} = 1 \right)$ for all $s \geq 0$; taking $s = 0$ yields the second claim. \hfill \Box

8.3 Proof of Theorem 2.3

The key for the proof is the following lemma, which is proved in Section 9 of Jin (2007).
Lemma 8.2 Consider $n$ independent random variable $X_j = \mu_j + z_j$ where $z_j \overset{iid}{\sim} N(0,1)$, suppose $\frac{1}{n} \sum_{j=1}^{n} |\mu_j| \leq r$ for some constant $r > 0$. When $n \rightarrow \infty$, for any fixed $q > 3/2$, \( P\{|0 \leq t \leq \log n| \{1 + \sum_{j=1}^{n} (\cos(t X_j) - E[\cos(t X_j)]) \} \geq \frac{\sqrt{2q \log n}}{\sqrt{n}} \} \leq 2 \cdot \log^2(n) \cdot n^{-2q/3} \cdot (1 + o(1)). \)

We now proceed to prove Theorem 2.3. For short, write $\varphi_n(t) = \varphi(t; X_1, \ldots, X_n)$ and $\varphi(t) = \varphi(t; \mu, n)$. Noticing $E[\cos(t X_j)] = e^{-\frac{t^2}{2}} \cos(t \mu_j)$, so by definitions, for all $t > 0$, $|\varphi_n(t) - \varphi(t)| \leq 2 \int_0^1 \omega(\xi) \epsilon \xi^2 \frac{1}{\sqrt{2\pi}} (\cos(t \xi) - E[\cos(t \xi)]) d\xi$. By Lemma 8.2, for any fixed $q > 3/2$, except for an event with probability $2 \log^2(n) n^{-2q/3}$, $\sup_{\mu \in B_n(r)} \sup_{\{0 \leq t \leq \sqrt{2q \log n}\}} |\varphi_n(t) - \varphi(t)| \leq 2 \int_0^1 \omega(\xi) \epsilon \xi^2 \frac{1}{\sqrt{2\pi}} (\cos(t \xi) - E[\cos(t \xi)]) d\xi$. Now, denote $A = \sup_{\{0 \leq t \leq 1\}} \omega(\xi)$ and write $\gamma_n = \gamma \log n$ for short. By elementary calculus, $\int_0^1 \omega(\xi) \epsilon \xi^2 d\xi \leq A \int_0^1 \epsilon \xi^2 d\xi = \frac{A}{2} \cdot \gamma_n^2 \cdot (1 + o(1))$. Combining these gives the theorem. 

\[ \square \]

8.4 Proof of Theorem 2.4

For short, write $t_n = \sqrt{2\gamma \log n}$, $\varphi(t_n) = \varphi(t; X_1, \ldots, X_n)$, $\varphi(t) = \varphi(t; \mu, n)$, and $\epsilon_n = \epsilon_n(\mu)$. Observe that for any $t$, we have the triangle inequality $|\varphi_n(t) - 1| \leq |\varphi_n(t) - \varphi(t)| + |\varphi(t) - 1|$. Now, first, using Theorem 2.3, when $n \rightarrow \infty$, 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 $\sup_{\Theta_n(\gamma, r)} \frac{\varphi_n(t_n) - \varphi(t_n)}{\epsilon_n} \leq \sup_{\Theta_n(\gamma, r)} \frac{\varphi(t_n) - \varphi(t_n)}{\epsilon_n} \leq C \sqrt{1/\log n}$. Second, by the definition of $\varphi$ and $\psi$, $|\varphi_n(t_n) - 1| = |\sup_{\Theta_n} \varphi_n(t_n) - 1| \leq \sup_{\Theta_n} \sup_{\{u \in [0, \sqrt{\log n}]\}} |\varphi(u; t_n)|$, uniformly for all $\mu \in \Theta_n(\gamma, r)$; note that by the way $\psi$ is constructed, the right hand side of the inequality tends to 0. Plugging these into the triangle inequality gives the theorem. 

\[ \square \]

8.5 Proof of Corollary 2.1

For short, write $\epsilon_n = \epsilon_n(\mu)$, $\varphi_n(t) = \varphi(t; X_1, \ldots, X_n)$, $\varphi(t) = \varphi(t; \mu, n)$, and $\Theta_n = \Theta_n(\gamma, r)$. Notice that $|\sup_{\Theta_n(\gamma, r)} \varphi_n(t) - \sup_{\Theta_n(\gamma, r)} \varphi(t)| \leq \sup_{\Theta_n(\gamma, r)} |\varphi_n(t) - \varphi(t)|$, so it is sufficient to show (a) $\sup_{\Theta_n(\gamma, r)} \frac{\sup_{\Theta_n(\gamma, r)} |\varphi_n(t_n) - \varphi(t_n)|}{\epsilon_n} \rightarrow 0$ in probability, and (b) $\lim_{n \rightarrow \infty} \sup_{\Theta_n(\gamma, r)} |\sup_{\Theta_n(\gamma, r)} |\varphi_n(t_n) - \varphi(t_n)| - 1| = 0$. First, for (a). Using Theorem 2.4, when $n \rightarrow \infty$, 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 $\sup_{\Theta_n(\gamma, r)} \frac{\sup_{\Theta_n(\gamma, r)} |\varphi_n(t_n) - \varphi(t_n)|}{\epsilon_n} \leq C \sup_{\Theta_n(\gamma, r)} \frac{1}{\sqrt{1/\log n}} \rightarrow 0$. Second, for (b). By Theorem 1.1 and symmetry, $\varphi(\sqrt{2\gamma \log n}) \leq \sup_{\Theta_n(\gamma, r)} \varphi(t) \leq \epsilon_n$, hence $\sup_{\Theta_n(\gamma, r)} |\sup_{\Theta_n(\gamma, r)} |\varphi_n(t_n) - \varphi(t_n)| - 1| \leq \sup_{\Theta_n(\gamma, r)} |\varphi_n(t_n) - 1| = 0$. Plugging these into the triangle inequality gives the theorem. 

\[ \square \]

8.6 Proof of Theorem 3.1

The proof of this Theorem is similar to that of Lemma 8.2 so we skip it.

Lemma 8.3 Consider $n$ independent random samples $X_j \sim F$ with $E_F|X|^2 < \infty$. When $n \rightarrow \infty$, there is a constant $C > 0$ such that with overwhelming probability, $\max_{\{0 \leq t \leq \log n\}} \frac{1}{n} \sum_{j=1}^{n} (\cos(t X_j) - E[\cos(t X_j)]) \leq C \cdot \frac{n \log n}{\sqrt{n}}$. 

We now proceed to prove Theorem 3.1. As the proofs are similar, we only prove the first claim. Define $\varphi(t; \mu, n, F) = E_F[\varphi_n(t; X_1, \ldots, X_n)]$. Using Fubini Theorem, $\varphi(t; \mu, n, F) = \int_{-1}^{1} \omega(\xi) E[1 - e^{\frac{t^2}{2}} \cos(t \xi X_1)] d\xi = \epsilon_n \cdot \int_{-1}^{1} \omega(\xi) [\int_{-1}^{1} (1 - \cos(t \xi u)) dF(u)] d\xi$. For short, write $t_n = \sqrt{2\gamma \log n}$, $\varphi_n(t) = \varphi(t; X_1, \ldots, X_n)$ and $\varphi(t) = \varphi(t; \mu, n, F)$, by Fubini theorem, $\varphi(t_n)/\epsilon_n = \int_{-1}^{1} \omega(\xi) (1 - \cos(t_n \xi u)) d\xi = 1$ for any $u \neq 0$, using Dominant Convergence Theorem gives $\lim_{n \rightarrow \infty} \int_{-1}^{1} \omega(\xi) (1 - \cos(t_n \xi u)) d\xi = 1$.
\[ |\frac{\varphi_n(t)-\varphi(t)}{\epsilon_n} - \frac{\varphi(t)}{\epsilon_n} | \leq n^{\beta} \cdot \int_1^\infty \omega(\xi) e^{-\frac{\xi^2}{2}} \left| \frac{1}{n} \cos(t\xi X_j) - E[\cos(t\xi X_j)] \right| \leq C_n^{\alpha+\gamma-1/2}. \] Since \( \gamma + \beta \leq 1/2 \), combining this with \( \lim_{n \to \infty} |\frac{\varphi(t_n)}{\epsilon_n} - 1| = 0 \) gives the theorem. \( \square \)

8.7 Proof of Theorem 3.2

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 \( \text{Re}(\hat{g}(s)) = \hat{\phi}(s)h(s) \). Note that \( h(s) \) is a bounded function which tends to 0 as \( s \) tends to \( \infty \). On one hand, by similar arguments as that in the proof of Theorem 3.1, there is a constant \( C = C(\gamma, \omega, q) \) such that with overwhelming probability, \( |\frac{\varphi_n(t) - \varphi(t)}{\epsilon_n}| \leq C/\sqrt{\log n} \). On the other hand, direct calculation shows that \( |\frac{\varphi(t_n)}{\epsilon_n} - 1| = |\int \omega(\xi)h(t\xi)d\xi| \), where the right hand side tends to 0 as \( n \) tends to \( \infty \). Since \( |\frac{\varphi(t_n)}{\epsilon_n} - 1| \leq |\frac{\varphi(t) - \varphi(t_n)}{\epsilon_n}| + |\frac{\varphi(t_n)}{\epsilon_n} - 1| \), the claim follows directly. \( \square \)

8.8 Proof of Theorem 3.4

We employ the theory on the FDR functional developed in Donoho and Jin (2006) for the proof. The FDR functional \( T_q(\cdot) \) is defined as \( T_q(G) = \inf \{ t : \hat{G}(t) \geq \frac{1}{q} \hat{G}_0(t) \} \), where \( \hat{G} = 1 - G \) is any survival function and \( \hat{G}_0 \) 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 \( 0 < c_0 < 1/2 \), Corollary 4.2 in Donoho and Jin (2006) can be extended to the current situation and we have \( \sup_{c_0 \leq q \leq 1-c_0} |T_q(G) - T_q(G_n)| = O_p\left(\frac{1}{\sqrt{n}}\right) \).

Notice that \( G(t) \) can be written in the form of \( G(t) = (1 - c)G_0(t) + \epsilon H(t) \), where \( H(t) \) is the marginal cdf associated with the null-effects. Let \( q_a \) be the unique solution of \( \hat{H}(T_q(G)) \). Note that when \( X_j \sim H \), the probability that \( X_j \) exceeds \( T_q(G_n) \) is \( \alpha \). View \( T_q(G_n) \) as a non-stochastic oracle threshold, and treat \( X_j \) as a discovery if and only if it exceeds the threshold, then the resulted total number of true positives is distributed as Binomial\((n\epsilon, \alpha)\). As a result, the proportion of signals exceeding the threshold \( T_q(G) \) tends to \( \alpha \) in probability.

At the same time, notice that the stochastic threshold in the proposed procedure equals to \( T_q(G_n) \). So to show the theorem, it is sufficient to show that the stochastic threshold converges to the non-stochastic oracle threshold:

\[ T_q(G_n) \to T_q(G), \quad \text{in probability.} \quad (8.1) \]

We now show (8.1). For short, write \( t_0 = T_{q_0}(G) \), \( \hat{t}_n = T_q(G_n) \). Introduce a bridging quantity \( \hat{t}_n = T_q(G) \). By the definition of the FDR functional, it is not hard to show that there is a constant \( c = c(F) \in (0, 1/2) \) such that with overwhelming probability, \( \hat{q} \) falls in the interval \( [c, 1 - c] \). Recall \( \sup_{c \leq q \leq 1-c} |T_q(G) - T_q(G_n)| = O_p\left(\frac{1}{\sqrt{n}}\right) \), hence

\[ |\hat{t}_n - t_n| \to 0, \quad \text{in probability.} \quad (8.2) \]

Now, by the way that the procedure is designed, \#\{Total Discoveries\} \( q = n \cdot \hat{G}_n(\hat{t}_n) \), so \( \epsilon_n \cdot \alpha = G_n(\hat{t}_n) \cdot (1 - (1 - \epsilon)n) q_n \). Notice that, (a) \( \sup_{q} |G_n(t) - G(t)| = O_p(1/\sqrt{n}) \) by Dvoretzky-Kiefer-Wolfowitz theorem (Shorack and Wellner, 1986); (b) \( \epsilon_n/\epsilon \to 1 \) in probability; combining these with (8.2) yields:

\[ \epsilon \cdot \alpha \cdot (1 + o_p(1)) = G(\hat{t}_n) \cdot (1 - (1 - \epsilon)q_n). \quad (8.3) \]

In addition, observe that for any \( 0 < q < 1 \) and \( t \equiv T_q(G) \), \( \frac{1}{q} \hat{G}_0(t) = \hat{G}(t) = (1 - c)\hat{G}_0(t) + \epsilon \hat{H}(t) \), so \( \epsilon \cdot \hat{H}(\hat{t}_n) = G(\hat{t}_n) \cdot (1 - (1 - \epsilon)q) \); plugging this into (8.3) gives \( \alpha \cdot (1 + o_p(1)) = \hat{H}(\hat{t}_n) \). Now, comparing \( \alpha \cdot (1 + o_p(1)) = \hat{H}(\hat{t}_n) \) with the definition of \( q_n \) gives that \( |\hat{t}_n - t_n| \to 0 \) in probability, which, together with (8.2), gives (8.1). \( \square \)

9 Acknowledgments

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

References


