Estimation of a two-component mixture model with applications to multiple testing

Rohit Kumar Patra and Bodhisattva Sen

Columbia University, New York, USA

[Received May 2012. Final revision August 2015]

Summary. We consider a two-component mixture model with one known component. We develop methods for estimating the mixing proportion and the unknown distribution non-parametrically, given independent and identically distributed data from the mixture model, using ideas from shape-restricted function estimation. We establish the consistency of our estimators. We find the rate of convergence and asymptotic limit of the estimator for the mixing proportion. Completely automated distribution-free honest finite sample lower confidence bounds are developed for the mixing proportion. Connection to the problem of multiple testing is discussed. The identifiability of the model and the estimation of the density of the unknown distribution are also addressed. We compare the estimators proposed, which are easily implementable, with some of the existing procedures through simulation studies and analyse two data sets: one arising from an application in astronomy and the other from a microarray experiment.

Keywords: Cramér–von Mises statistic; Cross-validation; Functional delta method; Identifiability; Local false discovery rate; Lower confidence bound; Microarray experiment; Projection operator; Shape-restricted function estimation

1. Introduction

Consider a mixture model with two components, i.e.

\[ F(x) = \alpha F_s(x) + (1 - \alpha) F_b(x), \]

where the cumulative distribution function (CDF) \( F_b \) is known, but the mixing proportion \( \alpha \in [0, 1] \) and the CDF \( F_s (\neq F_b) \) are unknown. Given a random sample from \( F \), we wish to estimate (non-parametrically) \( F_s \) and the parameter \( \alpha \).

This model appears in many contexts. In multiple-testing problems (microarray analysis, neuroimaging) the \( p \)-values, obtained from the numerous (independent) hypotheses tests, are uniformly distributed on \([0,1]\), under hypothesis \( H_0 \), whereas their distribution associated with \( H_1 \) is unknown; see, for example, Efron (2010) and Robin et al. (2007). Translated to the setting of model (1), \( F_b \) is the uniform distribution and the goal is to estimate the proportion of false null hypotheses \( \alpha \) and the distribution of the \( p \)-values under the alternative. In addition, a reliable estimator of \( \alpha \) is important when we want to assess or control multiple error rates, such as the false discovery rate of Benjamini and Hochberg (1995).

In contamination problems, the distribution \( F_b \), for which reasonable assumptions can be made, may be contaminated by an arbitrary distribution \( F_s \), yielding a sample drawn from \( F \) as in model (1); see, for example, McLachlan and Peel (2000). For example, in astronomy,
such situations arise quite often: when observing some variable(s) of interest (e.g. metallicity and radial velocity) of stars in a distant galaxy, foreground stars from the Milky Way, in the field of view, contaminate the sample; the galaxy (‘signal’) stars can be difficult to distinguish from the foreground stars as we can only observe the stereographic projections and not the three-dimensional position of the stars (see Walker et al. (2009)). Known physical models for the foreground stars help us to constrain \( F_b \), and the focus is on estimating the distribution of the variable for the signal stars, i.e. \( F_s \). We discuss such an application in more detail in Section 9.2. Such problems also arise in high energy physics where often the signature of new physics is evidence of a significant looking peak at some position on top of quite a smooth background distribution; see, for example, Lyons (2008).

Most of the previous work on this problem assumes some constraint on the form of the unknown distribution \( F_s \); for example, it is commonly assumed that the distributions belong to certain parametric models, which lead to techniques based on maximum likelihood (see, for example, Cohen (1967) and Lindsay (1983)), minimum \( \chi^2 \) (see, for example, Day (1969)) the method of moments (see, for example, Lindsay and Basak (1993)) and moment-generating functions (see, for example, Quandt and Ramsey (1978)). Bordes et al. (2006) assumed that both the components belong to an unknown symmetric location–shift family. Jin (2008) and Cai and Jin (2010) used empirical characteristic functions to estimate \( F_s \) under a semiparametric normal mixture model. In multiple testing, this problem has been addressed by various researchers and different estimators and confidence bounds for \( \alpha \) have been proposed in the literature under certain assumptions on \( F_s \) and its density; see for example, Storey (2002), Genovese and Wasserman (2004), Meinshausen and Rice (2006), Meinshausen and Bühlmann (2005), Celisse and Robin (2010) and Langaaas et al. (2005). For brevity, we do not discuss these references here but come back to this application in Section 7.

In this paper we provide a methodology to estimate \( \alpha \) and \( F_s \) (non-parametrically), without assuming any constraint on the form of \( F_s \). The main contributions of our paper can be summarized as follows.

(a) We investigate the identifiability of model (1) in complete generality.

(b) When \( F \) is a continuous CDF, we develop an honest finite sample lower confidence bound for the mixing proportion \( \alpha \). We believe that this is the first attempt to construct a distribution-free lower confidence bound for \( \alpha \) that is also tuning parameter free.

(c) Two estimators of \( \alpha \) are proposed and studied. We derive the rate of convergence and asymptotic limit for one of the estimators proposed.

(d) A non-parametric estimator of \( F_s \) by using ideas from shape-restricted function estimation is proposed and its consistency is proved. Further, if \( F_s \) has a non-increasing density \( f_s \), we can also consistently estimate \( f_s \).

The paper is organized as follows. In Section 2 we address the identifiability of the model given in expression (1). In Section 3 we propose an estimator of \( \alpha \) and investigate its theoretical properties, including its consistency, rate of convergence and asymptotic limit. In Section 4 we develop a completely automated distribution-free honest finite sample lower confidence bound for \( \alpha \). As the performance of the estimator proposed in Section 3 depends on the choice of a tuning parameter, in Section 5 we study a tuning-parameter-free heuristic estimator of \( \alpha \). We discuss the estimation of \( F_s \) and its density \( f_s \) in Section 6. Connection to the multiple-testing problem is developed in Section 7. In Section 8 we compare the finite sample performance of our procedures, including a plug-in and cross-validated choice of the tuning parameter for the estimator that is proposed in Section 3, with other methods that are available in the literature through simulation studies, and we provide a clear recommendation to the practitioner. Two
real data examples, one arising in astronomy and the other from a microarray experiment, are analysed in Section 9. Appendix A gives the proofs of some of the main results in the paper. The proofs of the results that are not given in Appendix A can be found in section 15 of the on-line supplementary material.

2. The model and identifiability

2.1. When \( \alpha \) is known

Suppose that we observe an independent and identically distributed sample \( X_1, X_2, \ldots, X_n \) from \( F \) as in model (1). If \( \alpha \in (0, 1) \) were known, a naive estimator of \( F \) would be

\[
\hat{F}_{\alpha} = \frac{F_n - (1 - \alpha)F_b}{\alpha},
\]

where \( F_n \) is the empirical CDF of the observed sample, i.e. \( F_n(x) = \sum_{i=1}^{n} I\{X_i \leq x\} / n \). Although this estimator is consistent, it does not satisfy the basic requirements of a CDF: \( \hat{F}_{\alpha} \) need not be non-decreasing or lie between 0 and 1. This naive estimator can be improved by imposing the known shape constraint of monotonicity. This can be accomplished by minimizing

\[
\int \{W(x) - \hat{F}_{\alpha}^{s,n}(x)\}^2 dF_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} \{W(X_i) - \hat{F}_{\alpha}^{s,n}(X_i)\}^2
\]

over all CDFs \( W \). Let \( \hat{F}_{\alpha}^{s,n} \) be a CDF that minimizes expression (3). The above optimization problem is the same as minimizing \( \|\theta - V\|^2 \) over \( \theta = (\theta_1, \ldots, \theta_n) \in \Theta_{inc} \) where

\[
\Theta_{inc} = \{\theta \in \mathbb{R}^n : 0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n \leq 1\},
\]

\( V = (V_1, V_2, \ldots, V_n), V_i := \hat{F}_{\alpha}^{s,n}(X_{(i)}), i = 1, 2, \ldots, n, X_{(i)} \) being the \( i \)th order statistic of the sample, and \( \| \cdot \| \) denotes the usual Euclidean norm in \( \mathbb{R}^n \). The estimator \( \hat{\theta} \) is uniquely defined by the projection theorem (see, for example, proposition 2.2.1 on page 88 of Bertsekas (2003)): it is the Euclidean projection of \( V \) on the closed convex set \( \Theta_{inc} \subset \mathbb{R}^n \). \( \hat{\theta} \) is related to \( \hat{F}_{\alpha}^{s,n}(X_{(i)}) = \hat{\theta}_i \) and can be easily computed by using the pool adjacent violators algorithm; see section 1.2 of Robertson et al. (1988). Thus, \( \hat{F}_{\alpha}^{s,n} \) is uniquely defined at the data points \( X_i \), for all \( i = 1, \ldots, n \), and can be defined on the entire real line by extending it to a piecewise constant right continuous function with possible jumps only at the data points. The following result, which is derived easily from chapter 1 of Robertson et al. (1988), characterizes \( \hat{F}_{\alpha}^{s,n} \).

**Lemma 1.** Let \( F_{\alpha}^{s,n} \) be the isotonic regression (see for example, page 4 of Robertson et al. (1988)) of the set of points \( \{\hat{F}_{\alpha}^{s,n}(X_{(i)})\}_{i=1}^{n} \). Then \( F_{\alpha}^{s,n} \) is characterized as the right-hand slope of the greatest convex minorant of the set of points \( \{i/n, \Sigma_{j=0}^{i} \hat{F}_{\alpha}^{s,n}(X_{(j)})\}_{i=0}^{n} \). The restriction of \( F_{\alpha}^{s,n} \) to \([0, 1] \), i.e. \( \hat{F}_{\alpha}^{s,n} = \min \{\max \{F_{\alpha}^{s,n}, 0\}, 1\} \), minimizes expression (3) over all CDFs.

Isotonic regression and the pool adjacent violators algorithm have been very well studied in the statistical literature with many textbook length treatments; see, for example, Robertson et al. (1988) and Barlow et al. (1972). If skilfully implemented, the pool adjacent violators algorithm has a computational complexity of \( O(n) \) (see Grotzinger and Witzgall (1984)).

2.2. Identifiability of \( F_\alpha \)

When \( \alpha \) is unknown, the problem is considerably more difficult; in fact, it is non-identifiable. If model (1) holds for some \( F_b \) and \( \alpha \) then the mixture model can be rewritten as...
for $0 \leq \gamma \leq 1 - \alpha$, and the term $(\alpha F_s + \gamma F_b)/(\alpha + \gamma)$ can be thought of as the non-parametric component. A trivial solution occurs when we take $\alpha + \gamma = 1$, in which case expression (3) is minimized when $W = \mathbb{P}_n$. Hence, $\alpha$ is not uniquely defined. To handle the identifiability issue, we redefine the mixing proportion as

$$\alpha_0 := \inf\{\gamma \in (0, 1] : \{F - (1 - \gamma) F_b\}/\gamma \text{ is a CDF}\}.$$  

Intuitively, this definition makes sure that the ‘signal’ distribution $F_s$ does not include any contribution from the known background $F_b$.

In this paper we consider the estimation of $\alpha_0$ as defined in expression (4). Identifiability of mixture models has been discussed in many references, but generally with parametric assumptions on the model. Genovese and Wasserman (2004) discussed identifiability when $F_b$ is the uniform distribution and $F$ has a density. Hunter et al. (2007) and Bordes et al. (2006) discussed identifiability for location–shift mixtures of symmetric distributions. Most researchers try to find conditions for the identifiability of their model, whereas we go a step further and quantify the non-identifiability by calculating $\alpha_0$ and investigating the difference between $\alpha$ and $\alpha_0$. In fact, most of our results are valid even when model (1) is non-identifiable.

Suppose that we start with a fixed $F_s$, $F_b$ and $\alpha$ satisfying model (1). As seen from the above discussion we can only hope to estimate $\alpha_0$, which, from its definition in expression (4), is smaller than $\alpha$, i.e. $\alpha_0 \leq \alpha$. A natural question that arises now is: under what condition(s) can we guarantee that the problem is identifiable, i.e. $\alpha_0 = \alpha$? The following lemma, which is proved in Appendix A, gives the connection between $\alpha$ and $\alpha_0$.

**Lemma 2.** Let $F$ be as in model (1) and $\alpha_0$ as defined in expression (4). Then

$$\alpha_0 = \alpha - \sup\{0 \leq \epsilon \leq 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\},$$

where sub-CDF is a non-decreasing right continuous function taking values between 0 and 1. In particular, $\alpha_0 < \alpha$ if and only if there exists $\epsilon \in (0, 1)$ such that $\alpha F_s - \epsilon F_b$ is a sub-CDF. Furthermore, $\alpha_0 = 0$ if and only if $F = F_b$.

In what follows we separately identify $\alpha_0$ for any distribution, be it continuous or discrete or a mixture of the two, with a series of lemmas proved in the on-line supplementary material. By an application of the Lebesgue decomposition theorem in conjunction with the Jordan decomposition theorem (see page 142, chapter V, section 3a* of Feller (1971)), we have that any CDF $G$ can be uniquely represented as a weighted sum of a piecewise constant CDF $G^{(d)}$, an absolutely continuous CDF $G^{(a)}$ and a continuous but singular CDF $G^{(s)}$, i.e. $G = \eta_1 G^{(a)} + \eta_2 G^{(d)} + \eta_3 G^{(s)}$, where $\eta_i \geq 0$, for $i = 1, 2, 3$, and $\eta_1 + \eta_2 + \eta_3 = 1$. However, from a practical point of view, we can assume that $\eta_3 = 0$, since singular functions almost never occur in practice; see, for example, Parzen (1960). Hence, we may assume that

$$G = \eta G^{(a)} + (1 - \eta) G^{(d)},$$

where $1 - \eta$ is the sum total of all the point masses of $G$. Let $d(G)$ denote the set of all jump discontinuities of $G$, i.e. $d(G) = \{x \in \mathbb{R} : G(x) - G(x-) > 0\}$. Let us define $J_G : d(G) \to [0, 1]$ to be a function that is defined only on the jump points of $G$ such that $J_G(x) = G(x) - G(x-)$ for all $x \in d(G)$. The following result addresses the identifiability issue when both $F_s$ and $F_b$ are discrete CDFs.

**Lemma 3.** Let $F_s$ and $F_b$ be discrete CDFs. If $d(F_b) \not\subset d(F_s)$, then $\alpha_0 = \alpha$, i.e. model (1) is
identifiable. If \( d(F_b) \subset d(F_s) \), then \( \alpha_0 = \alpha \{ 1 - \inf_{x \in d(F_b)} J_{F_s}(x)/J_{F_b}(x) \} \). Thus, \( \alpha_0 = \alpha \) if and only if \( \inf_{x \in d(F_b)} J_{F_s}(x)/J_{F_b}(x) = 0 \).

Next, let us assume that both \( F_s \) and \( F_b \) are absolutely continuous CDFs.

**Lemma 4.** Suppose that \( F_s \) and \( F_b \) are absolutely continuous, i.e. they have densities \( f_s \) and \( f_b \) respectively. Then

\[
\alpha_0 = \alpha \left( 1 - \text{ess inf } \frac{f_b}{f_b} \right),
\]

where, for any function \( g \), \( \text{ess inf } g = \sup \{ a \in \mathbb{R} : m \{ \{ x : g(x) < a \} \} = 0 \} \), \( m \) being the Lebesgue measure. As a consequence, \( \alpha_0 < \alpha \) if and only if there exists \( c > 0 \) such that \( f_s \geq c f_b \), almost everywhere \( F \).

Lemma 4 states that if there does not exist any \( c > 0 \) for which \( f_s(x) \geq c f_b(x) \), for almost every \( x \), then \( \alpha_0 = \alpha \) and we can estimate the mixing proportion correctly. Note that, in particular, if the support of \( F_s \) is strictly contained in that of \( F_b \), then the problem is identifiable and we can estimate \( \alpha \).

In section 12 of the on-line supplementary material we apply lemmas 3 and 4 to two discrete (Poisson and binomial) distributions and two absolutely continuous (exponential and normal) distributions to obtain the exact relationship between \( \alpha \) and \( \alpha_0 \). In the following lemma, which is proved in greater generality in section 12 of the on-line supplementary material, we give conditions under which a general CDF \( F \), that can be represented as in equation (6), is identifiable.

**Lemma 5.** Suppose that \( F = \kappa F^{(a)} + (1 - \kappa) F^{(d)} \), where \( F^{(a)} \) is an absolutely continuous CDF and \( F^{(d)} \) is a piecewise constant CDF, for some \( \kappa \in (0, 1) \). Then model (1) is identifiable, if either \( F^{(a)} \) or \( F^{(d)} \) are identifiable.

### 3. Estimation

#### 3.1. Estimation of the mixing proportion \( \alpha_0 \)

In this section we consider the estimation of \( \alpha_0 \) as defined in equation (5). For the rest of the paper, unless otherwise noted, we assume that

\[ X_1, X_2, \ldots, X_n \text{ is an independent and identically distributed sample from } F \text{ as in model (1)}. \]

Recall the definitions of \( \hat{F}_{s,n}^\gamma \) and \( \hat{F}_{s,n}^\gamma \), for \( \gamma \in (0, 1] \); see expressions (2) and (3). When \( \gamma = 1 \), we have \( \hat{F}_{s,n}^\gamma = \mathbb{F}_n = \hat{F}_{s,n}^\gamma \) as \( \hat{F}_{s,n}^{\gamma} \) (for \( \gamma = 1 \)) is a CDF, whereas, when \( \gamma \) is much smaller than \( \alpha_0 \), the regularization of \( \hat{F}_{s,n}^{\gamma} \) modifies it, and thus \( \hat{F}_{s,n}^{\gamma} \) and \( \hat{F}_{s,n}^{\gamma} \) are quite different. We would like to compare the naive and isotonized estimators \( \hat{F}_{s,n}^{\gamma} \) and \( \hat{F}_{s,n}^{\gamma} \) respectively, and to choose the smallest \( \gamma \) for which their distance is still small. This leads to the following estimator of \( \alpha_0 \):

\[
\hat{\alpha}_0^{c_n} = \inf \left\{ \gamma \in (0, 1] : \gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) \leq \frac{c_n}{\sqrt{n}} \right\}, \tag{7}
\]

where \( c_n \) is a sequence of constants and \( d_n \) stands for the \( L_2(\mathbb{F}_n) \) distance, i.e., if \( g, h : \mathbb{R} \rightarrow \mathbb{R} \) are two functions, then \( d_n^2(g, h) = \int \{|g(x) - h(x)|^2\} \,d\mathbb{F}_n(x) \). It is easy to see that

\[
d_n(\mathbb{F}_n, \gamma \hat{F}_{s,n}^\gamma + (1 - \gamma) F_b) = \gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma). \tag{8}
\]

For simplicity of notation, using equation (8), we define \( \gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) \) for \( \gamma = 0 \) as

\[
\lim_{\gamma \to 0^+} \gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) = d_n(\mathbb{F}_n, F_b). \tag{9}
\]
This convention is followed in the rest of the paper.

The choice of \( c_n \) is important, and in the following sections we address this issue in detail. We derive conditions on \( c_n \) that lead to consistent estimators of \( \alpha_0 \). We shall also show that particular (distribution-free) choices of \( c_n \) will lead to lower honest confidence bounds for \( \alpha_0 \).

Next, we prove a result which implies that, in the multiple-testing problem, estimators of \( \alpha_0 \) do not depend on whether we use \( p \)-values or \( z \)-values to perform our analysis. Let \( \Psi : \mathbb{R} \to \mathbb{R} \) be a known continuous non-decreasing function. We define \( \gamma \) criterion function

We start with two elementary results, which are proved in Appendix A, on the behaviour of our performance for estimating \( \alpha_0 \); see Section 8 and section 13 of the on-line supplementary material.

The choice of \( c_n \) is important, and in the following sections we address this issue in detail. We derive conditions on \( c_n \) that lead to consistent estimators of \( \alpha_0 \). We shall also show that particular (distribution-free) choices of \( c_n \) will lead to lower honest confidence bounds for \( \alpha_0 \).

Next, we prove a result which implies that, in the multiple-testing problem, estimators of \( \alpha_0 \) do not depend on whether we use \( p \)-values or \( z \)-values to perform our analysis. Let \( \Psi : \mathbb{R} \to \mathbb{R} \) be a known continuous non-decreasing function. We define \( \Psi^{-1}(y) := \inf \{ t \in \mathbb{R} : y \leq \Psi(t) \} \), and \( Y_i := \Psi^{-1}(X_i) \). It is easy to see that \( Y_1, Y_2, \ldots, Y_n \) is an independent and identically distributed sample from \( G := \alpha F_S \circ \Psi + (1 - \alpha) F_b \circ \Psi \). Suppose now that we work with \( Y_1, Y_2, \ldots, Y_n \), instead of \( X_1, X_2, \ldots, X_n \), and want to estimate \( \alpha \). We can define \( \alpha_0^\gamma \) as in equation (4) but with \( \{ G, F_b \circ \Psi \} \) instead of \( \{ F, F_b \} \). The following result, which is proved in the on-line supplementary material, shows that the \( \alpha_0 \) and its estimators, proposed in this paper, are invariant under such monotonic transformations.

**Theorem 1.** Let \( G_n \) be the empirical CDF of \( Y_1, Y_2, \ldots, Y_n \). Also, let \( \hat{G}_{s,n} \) and \( \hat{G}_{s,n}^\gamma \) be as defined in expressions (2) and (3) respectively, but with \( \{ G_n, F_b \circ \Psi \} \) instead of \( \{ F, F_b \} \). Then \( \alpha_0 = \alpha_0^\gamma \) and \( \gamma d_n(\hat{G}_{s,n}^\gamma, \hat{G}_{s,n}) = \gamma d_n(\hat{G}_{s,n}^\gamma, \hat{G}_{s,n}) \) for all \( \gamma \in (0, 1] \).

### 3.2. Consistency of \( \hat{\alpha}_0^{c_n} \)

We start with two elementary results, which are proved in Appendix A, on the behaviour of our criterion function \( \gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) \).

**Lemma 6.** For \( 1 \geq \gamma \geq \alpha_0 \), \( \gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) \leq d_n(\hat{F}, \hat{F}) \). Thus,

\[
\gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) \to \begin{cases} 0, & \gamma - \alpha_0 \geq 0, \\ > 0, & \gamma - \alpha_0 < 0, \end{cases}
\]

almost surely.

**Lemma 7.** The set \( A_n := \{ \gamma \in [0, 1] : \sqrt{n} \gamma d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma) \leq c_n \} \) is convex. Thus, \( A_n = [\hat{\alpha}_0^{c_n}, 1] \).

The following result, which is proved in the on-line supplementary material, shows that, for a broad range of choices of \( c_n \), our estimation procedure is consistent.

**Theorem 2.** If \( c_n = o(\sqrt{n}) \) and \( c_n \to \infty \), then \( \hat{\alpha}_0^{c_n} \to P \alpha_0 \).

A proper choice of \( c_n \) is important and crucial for the performance of \( \hat{\alpha}_0^{c_n} \). We suggest doing cross-validation to find the optimal tuning parameter \( c_n \). In Section 8.2.1 we detail this approach and illustrate its good finite sample performance through simulation examples; see Tables 2–5, Section 8.2.4, and section 13 (in the on-line supplementary material). However, cross-validation can be computationally expensive. Another useful choice for \( c_n \) is to take \( c_n = 0.1 \log \{ \log(n) \} \). After extensive simulations, we observe that \( c_n = 0.1 \log \{ \log(n) \} \) has good finite sample performance for estimating \( \alpha_0 \); see Section 8 and section 13 of the on-line supplementary material for more details.

### 3.3. Rate of convergence and asymptotic limit

We first discuss the case \( \alpha_0 = 0 \). In this situation, under minimal assumptions, we show that, as the sample size grows, \( \hat{\alpha}_0^{c_n} \) exactly equals \( \alpha_0 \) with probability converging to 1.

**Lemma 8.** When \( \alpha_0 = 0 \), if \( c_n \to \infty \) as \( n \to \infty \), then \( P(\hat{\alpha}_0^{c_n} = 0) \to 1 \).
For the rest of this section we assume that $\alpha_0 > 0$. The following theorem gives the rate of convergence of $\hat{\alpha}_0^c$.

**Theorem 3.** Let $r_n := \sqrt{n/c_n}$. If $c_n \to \infty$ and $c_n = o(n^{1/4})$ as $n \to \infty$, then $r_n(\hat{\alpha}_0^c - \alpha_0) = O_P(1)$.

The proof of this result is involved and we give the details in section 15.5 of the on-line supplementary material.

**Remark 1.** Genovese and Wasserman (2004) showed that the estimators of $\alpha_0$ that were proposed by Hengartner and Stark (1995) and Swanepoel (1999) have rates of convergence $\{n = \log \{\log(n)\}\}^{1/3}$ and $n^{2/5}/\log(n)^{\delta}$, for $\delta > 0$, respectively. Moreover, both results require smoothness assumptions on $F$—Hengartner and Stark (1995) required $F$ to be concave with a density that is Lipschitz of order 1, whereas Swanepoel (1999) required even stronger smoothness conditions on the density. Nguyen and Matias (2014) proved that, when the density of $F_{\alpha_0}$ vanishes at a set of points of measure 0 and satisfies certain regularity assumptions, then any $\sqrt{n}$-consistent estimator of $\alpha_0$ will not have finite variance in the limit (if such an estimator exists).

We can take $r_n = \sqrt{n}/c_n$ arbitrarily close to $\sqrt{n}$ by choosing $c_n$ that increases to $\infty$ very slowly. If we take $c_n = \log\{\log(n)\}$, we obtain an estimator that has a rate of convergence $\sqrt{n}/\log\{\log(n)\}$. In fact, as the next result (which is proved in section 15.6 of the on-line supplementary material) shows, $r_n(\hat{\alpha}_0^c - \alpha_0)$ converges to a degenerate limit. In Section 8.2, we analyse the effect of $c_n$ on the finite sample performance of $\hat{\alpha}_0^c$ for estimating $\alpha_0$ through simulations and advocate a proper choice of the tuning parameter $c_n$.

**Theorem 4.** When $\alpha_0 > 0$, if $r_n \to \infty$, $c_n = o(n^{1/4})$ and $c_n \to \infty$, as $n \to \infty$, then

$$r_n(\hat{\alpha}_0^c - \alpha_0) \overset{p}{\to} c,$$

where $c < 0$ is a constant that depends on $\alpha_0$, $F$ and $F_b$.

### 4. Lower confidence bound for $\alpha_0$

The asymptotic limit of the estimator $\hat{\alpha}_0^c$ that was discussed in Section 3 depends on unknown parameters (e.g. $\alpha_0$ and $F$) in a complicated fashion and is of little practical use. Our goal in this section is to construct a finite sample (honest) lower confidence bound $\hat{\alpha}_L$ with the property

$$P(\alpha_0 \geq \hat{\alpha}_L) \geq 1 - \beta,$$

for a specified confidence level $1 - \beta$ ($0 < \beta < 1$) that is valid for any $n$ and is tuning parameter free. Such a lower bound would allow us to assert, with a specified level of confidence, that the proportion of signal is at least $\hat{\alpha}_L$.

It can also be used to test the hypothesis that there is no signal at level $\beta$ by rejecting when $\hat{\alpha}_L > 0$. The problem of no signal is known as the homogeneity problem in the statistical literature. It is easy to show that $\alpha_0 = 0$ if and only if $F = F_b$. Thus, the hypothesis of no signal or homogeneity can be addressed by testing whether $\alpha_0 = 0$ or not. There has been a considerable amount of work on the homogeneity problem, but most of the references make parametric model assumptions. Lindsay (1995) is an authoritative monograph on the homogeneity problem but the components are assumed to be from a known exponential family. Walther (2001, 2002) discussed the homogeneity problem under the assumption that the densities are log-concave. Donoho and Jin (2004) and Cai and Jin (2010) discussed the problem of detecting sparse heterogeneous mixtures under parametric settings using the ‘higher criticism’ statistic; see section 14 of the on-line supplementary material for more details.
It will be seen that our approach will lead to an exact lower confidence bound when \( \alpha_0 = 0 \), i.e. \( P(\hat{\alpha}_L = 0) = 1 - \beta \). The methods of Genovese and Wasserman (2004) and Meinshausen and Rice (2006) usually yield conservative lower bounds.

**Theorem 5.** Let \( H_n \) be the CDF of \( \sqrt{n}d_n(F_n, F) \). Let \( \hat{\alpha}_L \) be defined as in equation (7) with \( c_n = H_n^{-1}(1 - \beta) \). Then inequality (11) holds. Furthermore if \( \alpha_0 = 0 \), then \( P(\hat{\alpha}_L = 0) = 1 - \beta \), i.e. it is an exact lower bound.

The proof of theorem 5 can be found in Appendix A. Note that \( H_n \) is distribution free (i.e. it does not depend on \( F_s \) and \( F_b \)) when \( F \) is a continuous CDF and can be readily approximated by Monte Carlo simulations using a sample of uniform distributions. For moderately large \( n \) (e.g. \( n \geq 500 \)) the distribution \( H_n \) can be very well approximated by that of the Cramér–von Mises statistic, defined as

\[
\sqrt{n}d(F_n, F) := \int n\{F_n(x) - F(x)\}^2 dF(x).
\]

Letting \( G_n \) be the CDF of \( \sqrt{n}d(F_n, F) \), we have the following result.

**Theorem 6.** \( \sup_{x \in \mathbb{R}} |H_n(x) - G_n(x)| \to 0 \) as \( n \to \infty \).

Hence in practice, for moderately large \( n \), we can take \( c_n \) to be the \((1 - \beta)\)-quantile of \( G_n \) or its asymptotic limit, which are readily available (for example, see Anderson and Darling (1952)). When \( F \) is a continuous CDF, the asymptotic 95\% quantile of \( G_n \) is 0.6792 and is used in our data analysis. Note that

\[
P(\alpha_0 \geq \hat{\alpha}_L) = P\{\sqrt{n}\alpha_0 d_n(F_{\alpha_0}^{s}, \bar{F}_{\alpha_0}^{s}) \geq H_n^{-1}(1 - \beta)\}.
\]

The following theorem gives the explicit asymptotic limit of \( P(\alpha_0 \geq \hat{\alpha}_L) \) but it is not useful for practical purposes as it involves the unknown \( F_{\alpha_0}^{s} \) and \( F_b \).

**Theorem 7.** Assume that \( \alpha_0 > 0 \). Then \( \sqrt{n}\alpha_0 d_n(F_{\alpha_0}^{s}, \bar{F}_{\alpha_0}^{s}) \to U \), where \( U \) is a random variable whose distribution depends only on \( \alpha_0 \), \( F \) and \( F_b \).

The proof of theorem 7 and the explicit form of \( U \) can be found in the on-line supplementary material. The proof of theorem 6 and a detailed discussion on the performance of the lower confidence bound for detecting heterogeneity in the *moderately sparse* signal regime considered in Donoho and Jin (2004) can be found in section 14 of the on-line supplementary material.

### 5. Heuristic estimator of \( \alpha_0 \)

In simulations, we observe that the finite sample performance of equation (7) is affected by the choice of \( c_n \) (for an extensive simulation study on this see Section 8.2). This motivates us to propose a method to estimate \( \alpha_0 \) that is completely automated and has good finite sample performance. We start with a lemma, which is proved in Appendix A, that describes the shape of our criterion function and will motivate our procedure.

**Lemma 9.** \( \gamma d_n(F_{\gamma}^{s,n}, \bar{F}_{\gamma}^{s,n}) \) is a non-increasing convex function of \( \gamma \) in \((0, 1)\).

Writing

\[
\bar{F}_{\gamma}^{s,n} = \frac{\bar{F}_n - F}{\gamma} + \frac{\alpha_0}{\gamma} F_{\alpha_0}^{s} + \left(1 - \frac{\alpha_0}{\gamma}\right) F_b,
\]

we see that, for \( \gamma \geq \alpha_0 \), the second term on the right-hand side is a CDF. Thus, for \( \gamma \geq \alpha_0 \), \( \bar{F}_{\gamma}^{s,n} \) is
very close to a CDF as \( F_n - F = O_p(n^{-1/2}) \), and hence \( \hat{F}_{s,n}^{\gamma} \) should also be close to \( \hat{F}_{s,n}^{\gamma} \). Whereas, for \( \gamma < \alpha_0 \), \( \hat{F}_{s,n}^{\gamma} \) is not close to a CDF, and thus the distance \( \gamma d_n(\hat{F}_{s,n}^{\gamma}, \hat{F}_{s,n}^{\gamma}) \) is appreciably large. Therefore, at \( \alpha_0 \), we have a ‘regime’ change: \( \gamma d_n(\hat{F}_{s,n}^{\gamma}, \hat{F}_{s,n}^{\gamma}) \) should have a slowly decreasing segment to the right of \( \alpha_0 \) and a steeply non-increasing segment to the left of \( \alpha_0 \). Fig. 1 shows two typical such plots of the function \( \gamma d_n(\hat{F}_{s,n}^{\gamma}, \hat{F}_{s,n}^{\gamma}) \), where Fig. 1(a) corresponds to a mixture of \( N(2, 1) \) with \( N(0, 1) \) (setting I) and in Fig. 1(b) we have a mixture of beta(1,10) and uniform(0, 1) distributions (setting II). We shall use these two settings to illustrate our methodology in the rest of this section and also in Section 8.1.

Using the above heuristics, we can see that the ‘elbow’ of the function should provide a good estimate of \( \alpha_0 \); it is the point that has the maximum curvature, i.e. the point where the second derivative is maximum. We denote this estimator by \( \tilde{\alpha}_0 \). Note that both the estimators \( \tilde{\alpha}_0 \) and \( \hat{\alpha}_0^{\gamma} \) are derived from \( \gamma d_n(\hat{F}_{s,n}^{\gamma}, \hat{F}_{s,n}^{\gamma}) \), as a function of \( \gamma \), albeit they look at two different aspects of the function.

In Fig. 1 we have used numerical methods to approximate the second derivative of \( \gamma d_n(\hat{F}_{s,n}^{\gamma}, \hat{F}_{s,n}^{\gamma}) \) (using the method of double differencing). We advocate plotting the function \( \gamma d_n(\hat{F}_{s,n}^{\gamma}, \hat{F}_{s,n}^{\gamma}) \) as \( \gamma \) varies between 0 and 1. In most cases, plots similar to Fig. 1 would immediately convey to
the practitioner the most appropriate choice of $\tilde{a}_0$. In some cases though, there can be multiple peaks in the second derivative, in which case some discretion on the part of the practitioner might be required. It should be noted that the idea of finding the point where the second derivative is large to detect an ‘elbow’ or ‘knee’ of a function is not uncommon; see, for example, Salvador and Chan (2004). However, in Section 8.2.4 and section 13 of the on-line supplementary material, we show some simulation examples where $\tilde{a}_0$ fails to estimate the elbow of $\gamma d_n(\hat{F}\uparrow_{s,n}, \tilde{F}\uparrow s_{s,n})$ consistently.

6. Estimation of the distribution function and its density

6.1. Estimation of $F_s$

Let us assume for the rest of this section that model (1) is identifiable, i.e. $\alpha = \alpha_0$, and $\alpha_0 > 0$. Thus $F_{s_0} = F_s$. Once we have a consistent estimator $\hat{\alpha}_n$ (which may or may not be $\hat{\alpha}^{\gamma}_{0,n}$, as discussed in the previous sections) of $\alpha_0$, a natural non-parametric estimator of $F_s$ is $\hat{F}\uparrow_{s,n}$, defined as the minimizer of expression (3). In the following theorem (which is proved in the on-line supplementary material) we show that, indeed, $\tilde{F}\uparrow_{s,n}$ is uniformly consistent for estimating $F_s$. We also derive the rate of convergence of $\hat{F}\uparrow_{s,n}$.

**Theorem 8.** Suppose that $\hat{\alpha}_n \rightarrow^P \alpha_0$. Then, as $n \rightarrow \infty$, sup$_{x \in \mathbb{R}} |\hat{F}\uparrow_{s,n}(x) - F_s(x)| \rightarrow^P 0$. Furthermore, if $q_n(\hat{\alpha}_n - \alpha_0) = O_P(1)$, where $q_n = o(\sqrt{n})$, then sup$_{x \in \mathbb{R}} q_n |\hat{F}\uparrow_{s,n}(x) - F_s(x)| = O_P(1)$.

Additionally, for $\hat{\alpha}^{\gamma}_{0,n}$ as defined in expression (7), we have

$$\sup_{x \in \mathbb{R}} |r_n(\hat{F}\uparrow^{\gamma}_{s,n} - F_s)(x) - Q(x)| \rightarrow^P 0,$$

where $r_n d(\hat{F}\uparrow^{\gamma}_{s,n}, F_s) \rightarrow^P c$ for a function $Q : \mathbb{R} \rightarrow \mathbb{R}$ and a constant $c > 0$ depending only on $\alpha_0$, $F$ and $F_0$.

An immediate consequence of theorem 8 is that $d_n(\hat{F}\uparrow_{s,n}, \tilde{F}\uparrow_{s,n}) \rightarrow^P 0$ as $n \rightarrow \infty$. Fig. 2(a) shows our estimator $\hat{F}\uparrow_{s,n}$ along with the true $F_s$ for the same data set as used in Fig. 1(b).

6.2. Estimating the density of $F_s$

Suppose now that $F_s$ has a density $f_s$. Obtaining non-parametric estimators of $f_s$ can be difficult as it requires smoothing and usually involves the choice of tuning parameter(s) (e.g. smoothing bandwidths), and especially so in our set-up.

In this subsection we describe a tuning-parameter-free approach to estimating $f_s$, under the additional assumption that $f_s$ is non-increasing. The assumption that $f_s$ is non-increasing, i.e. $F_s$ is concave on its support, is natural in many situations (see Section 7 for an application in the multiple-testing problem) and has been investigated by several researchers, including Grenander (1956), Langaas et al. (2005) and Genovese and Wasserman (2004). Without loss of generality, we assume that $f_s$ is non-increasing on $[0, \infty)$.

For a bounded function $g : [0, \infty) \rightarrow \mathbb{R}$, let us represent the least concave majorant (LCM) of $g$ by LCM[$g$]. Thus, LCM[$g$] is the smallest concave function that lies above $g$. Define $F_{s,n} := \text{LCM}[\hat{F}\uparrow^{\gamma}_{s,n}]$. Note that $F_{s,n}$ is a valid CDF. We can now estimate $f_s$ by $f_{s,n}^\uparrow$, where $f_{s,n}^\uparrow$ is the piecewise constant function that is obtained by taking the left derivative of $F_{s,n}$. In the following result, which is proved in the on-line supplementary material, we show that both $F_{s,n}$ and $f_{s,n}^\uparrow$ are consistent estimators of their population versions.

**Theorem 9.** Assume that $F_s(0) = 0$ and that $F_s$ is concave on $[0, \infty)$. If $\hat{\alpha}_n \rightarrow^P \alpha_0$, then, as $n \rightarrow \infty$,
7. Multiple-testing problem

The problem of estimating the proportion of false null hypotheses \( \alpha_0 \) is of interest in situations where a large number of hypothesis tests are performed. Recently, various such situations have arisen in applications. One major motivation is in estimating the proportion of genes that are differentially expressed in deoxyribonucleic acid microarray experiments. However, estimating the proportion of true null hypotheses is also of interest, for example, in functional magnetic resonance imaging (see Turkheimer et al. (2001)) and source detection in astrophysics (see Miller et al. (2001)).
Suppose that we wish to test \( n \) null hypotheses \( H_{01}, H_{02}, \ldots, H_{0n} \) on the basis of a data set \( X \). Let \( H_i \) denote the (unobservable) binary variable that is 0 if \( H_{0i} \) is true, and 1 otherwise, \( i = 1, \ldots, n \). We want a decision rule \( D \) that will produce a decision of ‘null’ or ‘non-null’ for each of the \( n \) cases. In their seminal work Benjamini and Hochberg (1995) argued that an important quantity to control is the false discovery rate (FDR) and proposed a procedure with the property \( \text{FDR} \leq \beta \). When \( \alpha_0 \) is significantly bigger than 0 an estimate of \( \hat{\alpha}_0 \) can be used to yield a procedure with FDR approximately equal to \( \beta \) and thus will result in an increased power. This is essentially the idea of the adapted control of FDR (see Benjamini and Hochberg (2000)). See Storey (2002), Black (2004), Langaas et al. (2005), Benjamini et al. (2006), and Donoho and Jin (2004) for a discussion on the importance of efficient estimation of \( \alpha_0 \) and some proposed estimators.

Our method can be directly used to yield an estimator of \( \alpha_0 \) that does not require the specification of any tuning parameter, as discussed in Section 5. We can also obtain a completely non-parametric estimator of \( F_S \), the distribution of the \( p \)-values arising from the alternative hypotheses. Suppose that \( F_b \) has a density \( f_b \) and \( F_S \) has a density \( f_s \). To keep the following discussion more general, we allow \( f_s \) to be any known density, although in most multiple-testing applications we shall take \( f_b \) to be uniform \( 0, \frac{1}{2} \). The local false discovery rate (LFDR) is defined as the function \( \hat{l}(x) : (0, 1) \rightarrow [0, \infty) \), where

\[
\hat{l}(x) = P(H_i = 0|X_i = x) = \frac{(1 - \alpha_0) f_b(x)}{f(x)},
\]

and \( f(x) = \alpha_0 f_s(x) + (1 - \alpha_0) f_b(x) \) is the density of the observed \( p \)-values. The estimation of the LFDR \( \hat{l} \) is important because it gives the probability that a particular null hypothesis is true given the observed \( p \)-value for the test. The LFDR method can help us to obtain easily interpretable thresholding methods for reporting the ‘interesting’ cases (e.g. \( l(x) \leq 0.20 \)). Obtaining good estimates of \( \hat{l} \) can be tricky as it involves the estimation of an unknown density, usually requiring smoothing techniques; see Section 5 of Efron (2010) for a discussion on estimation and interpretation of \( l \). From the discussion in Section 6.1, under the additional assumption that \( f_s \) is non-increasing, we have a natural tuning-parameter-free estimator \( \hat{l} \) of the LFDR:

\[
\hat{l}(x) = \frac{(1 - \hat{\alpha}_n) f_b(x)}{\hat{\alpha}_n f^*_n(x) + (1 - \hat{\alpha}_n) f_b(x)}, \quad \text{for } x \in (0, 1).
\]

The assumption that \( f_s \) is non-increasing, i.e. \( F_S \) is concave, is quite natural—when the alternative hypothesis is true the \( p \)-value is generally small—and has been investigated by several researchers, including Genovese and Wasserman (2004) and Langaas et al. (2005).

8. Simulation

To investigate the finite sample performance of the estimators that are developed in this paper, we carry out several simulation experiments. We also compare the performance of these estimators with existing methods. The R language (R Development Core Team, 2008) codes used to implement our procedures are available from [http://stat.columbia.edu/~rohit/research.html](http://stat.columbia.edu/~rohit/research.html).

8.1. Lower bounds for \( \alpha_0 \)

Although there has been some work on estimation of \( \alpha_0 \) in the multiple-testing setting, Meins-hausen and Rice (2006) and Genovese and Wasserman (2004) are the only references that we found that discuss methodology for constructing lower confidence bounds for \( \alpha_0 \). These proce-
Table 1. Coverage probabilities of nominal 95% lower confidence bounds for the three methods when 

\[ n = 1000 \] and \[ n = 5000 \]

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>Setting I</th>
<th>Setting II</th>
<th>Setting I</th>
<th>Setting II</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha}_L )</td>
<td>( \hat{\alpha}^{GW}_L )</td>
<td>( \hat{\alpha}^{MR}_L )</td>
<td>( \hat{\alpha}_L )</td>
<td>( \hat{\alpha}^{GW}_L )</td>
</tr>
<tr>
<td>0</td>
<td>0.95</td>
<td>0.98</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>0.01</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>0.05</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>0.10</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>


dures are connected and the methods in Meinshausen and Rice (2006) are extensions of those proposed in Genovese and Wasserman (2004). The lower bounds that were proposed in both the references approximately satisfy inequality (11) and have the form \( \sup_{t \in (0,1)} \left\{ \left[ \frac{1}{n} \right] \right\} \), where \( \eta_n \) is a bounding sequence for the bounding function \( \delta(t) \) at level \( \beta \); see Meinshausen and Rice (2006), Genovese and Wasserman (2004) used a constant bounding function, \( \delta(t) = 1 \), with \( \eta_n = \sqrt{\log(2/\beta)/(2n)} \), whereas Meinshausen and Rice (2006) suggested a class of bounding functions but observed that the standard deviation proportional bounding function \( \delta(t) = \sqrt{t(1-t)} \) has optimal properties among a large class of possible bounding functions. We use this bounding function and a bounding sequence that was suggested by Meinshausen and Rice (2006). We denote the lower bound proposed in Meinshausen and Rice (2006) by \( \hat{\alpha}^{MR}_L \), the bound in Genovese and Wasserman (2004) by \( \hat{\alpha}^{GW}_L \) and the lower bound discussed in Section 4 by \( \hat{\alpha}_L \). To be able to use the methods of Meinshausen and Rice (2006) and Genovese and Wasserman (2004) in setting I, which was introduced in Section 5, we transform the data such that \( F_b \) is uniform(0,1); see Section 3.1 for the details.

We take \( \alpha \in \{0, 0.01, 0.03, 0.05, 0.10\} \) and compare the performance of the three lower bounds in the two different simulation settings that were discussed in Section 5. For each setting we take the sample size \( n \) to be 1000 and 5000. We present the estimated coverage probabilities, obtained by averaging over 5000 independent replications, of the lower bounds for both settings in Table 1. We can immediately see from Table 1 that the bounds are usually quite conservative. However, it is worth pointing out that, when \( \alpha_0 = 0 \), our method has exact coverage, as discussed in Section 4. Also, the fact that our procedure is simple, easy to implement and completely automated makes it very attractive.

8.2. Estimation of \( \alpha_0 \)

In this subsection, we illustrate and compare the performance of various estimators of \( \alpha_0 \) under two sampling scenarios. In scenario A, we proceed as in Langaas et al. (2005). Let \( X_j = (X_{1j}, X_{2j}, \ldots, X_{nj}) \), for \( j = 1, \ldots, J \), and assume that each \( X_j \sim N(\mu_{nj \times 1}, \Sigma_{nj \times n}) \) and that \( X_1, X_2, \ldots, X_J \) are independent. We test \( H_0: \mu_i = 0 \) versus \( H_1: \mu_i \neq 0 \) for each \( i = 1, 2, \ldots, n \). We set \( \mu_i \) to 0 for the true null hypotheses, whereas for the false null hypotheses we draw \( \mu_i \) from a symmetric bi-triangular density with parameters \( a = \log_2(1.2) = 0.263 \) and \( b = \log_2(4) = 2 \); see page 568 of Langaas et al. (2005) for the details. Let \( x_{ij} \) denote a realization of \( X_{ij} \) and \( \alpha \) be the proportion of false null hypotheses. Let \( \tilde{x}_i = \sum_{j=1}^{J} x_{ij}/J \) and \( \tilde{x}_i^2 = \sum_{j=1}^{J} (x_{ij} - \tilde{x}_i)^2/(J-1) \).
To test $H_{0i}$ versus $H_{1i}$, we calculate a two-sided $p$-value based on a one-sample $t$-test, with $p_i = 2P(T_{J-1} > |\bar{z}_i|/\sqrt{(\sum_i w_i^2)/J})$, where $T_{J-1}$ is a $t$-distributed random variable with $J-1$ degrees of freedom.

In scenario B, we generate $n + L$ independent random variables $w_1, w_2, \ldots, w_{n+L}$ from $N(0, 1)$ and set

$$z_i = \frac{1}{\sqrt{(L+1)}} \sum_{j=i}^{i+L} w_j$$

for $i = 1, 2, \ldots, n$. The dependence structure of the $z_i$s is determined by $L$. For example, $L = 0$ corresponds to the case where the $z_i$s are standard normal. Let $X_i = z_i + m_i$, for $i = 1, 2, \ldots, n$, where $m_i = 0$ under the null, and, under the alternative, $|m_i|$ is randomly generated from uniform$(m^*, m^* + 1)$ and $\text{sgn}(m_i)$, the sign of $m_i$, is randomly generated from $\{-1, 1\}$ with equal probabilities. Here $m^*$ is a suitable constant that describes the simulation setting. Let $1 - \alpha$ be the proportion of true null hypotheses. Scenario B is inspired by the numerical studies in Cai and Jin (2010) and Jin (2008).

We use $\hat{a}^{S,B}_0$ to denote the estimator that was proposed by Storey (2002) when bootstrapping is used to choose the required tuning parameter, and denote by $\hat{a}^{S,\lambda}_0$ the estimator when the value of the tuning parameter is fixed at $\lambda$. Langaas et al. (2005) proposed an estimator that is tuning parameter free but crucially uses the known shape constraint of a convex and non-increasing $f_2$; we denote it by $\hat{a}^{L}_0$. We evaluate $\hat{a}^{L}_0$ by using the convest function in the R library limma. We also use the estimator that was proposed in Meinshausen and Rice (2006) for two bounding functions: $\hat{a}^{MR}_0$ when $\delta(t) = \sqrt{\{\log(1-t)\}}$ and $\hat{a}^{GW}_0$ when $\delta(t) = 1$ (see Genovese and Wasserman (2004)). We also compare our results with $\hat{a}^{J}_0$, the estimator that was proposed in Efron (2007) using the central matching method, computed using the locfdr function in the R library locfdr. Jin (2008) and Cai and Jin (2010) proposed estimators when the model is a mixture of Gaussian distributions; we denote the estimator that was proposed in section 2.2 of Jin (2008) by $\hat{a}^{J}_0$ and in section 3.1 of Cai and Jin (2010) by $\hat{a}^{J}_0$. Some of the competing methods require $F_b$ to be of a specific form (e.g. standard normal) in which case we transform the observed data suitably.

The estimator $\hat{a}^{GW}_{0n}$ depends on the choice of $c_n$ and in what follows we investigate a proper choice of $c_n$. We take $c_0 = 0.1$ and evaluate the performance of $\hat{a}^{GW}_{0n}$ for various values of $\tau$, as $n$ increases, for scenarios A and B. The choice $c_n = \tau \log\{\log(n)\}$, for various values of $\tau$, is suggested after extensive simulations. We also include $\hat{a}_0$, $\hat{a}^{GW}_0$, $\hat{a}^{MR}_0$ and $\hat{a}^{J}_0$ in the comparison. For scenario A, we fix the sample size $n$ at 5000 and $\Sigma = I_{n \times n}$. For scenario B, we fix $n = 5 \times 10^4$, $L = 0$ and $m^* = 1$. In Fig. 3, we illustrate the effect of $c_n$ on estimation of $\alpha_0$ as $n$ varies from 3000 to $10^5$. Recall that $\hat{a}_0$ denotes the estimator that was proposed in Section 5. For both scenarios, the sample means of the estimators of $\alpha_0$ that is proposed in this paper converge to the true $\alpha_0$, as the sample size grows. The methods that are developed in this paper perform favourably in comparison with $\hat{a}^{GW}_0$, $\hat{a}^{MR}_0$ and $\hat{a}^{J}_0$. Since the choice of $c_n$ dictates the finite sample performance of $\hat{a}^{GW}_{0n}$, we propose cross-validation to find an appropriate value of the tuning parameter.

8.2.1. Cross-validation

In this subsection, we use $c$ instead of $c_n$ to simplify the notation. In what follows we briefly describe our cross-validation procedure. For a $K$-fold cross-validation, we randomly partition
the data into $K$ sets, say $D_1, \ldots, D_K$. Let $F_n^k$ be the empirical CDF of the data in $D_k$. Let $\hat{\alpha}_{0,-k}^c$ be the estimator that is defined in expression (7) using all data except those in $D_k$ and tuning parameter $c$. Further, let $F_{s_{0,-k}}^{c_{0,-k}}$ be the estimator of $F_s$ as defined in lemma 1 using $\hat{\alpha}_{0,-k}^c$ and all data except those in $D_k$. Define the cross-validated estimator of $c$ as

$$c_{cv} := \arg \min_{c \in \mathbb{R}} \sum_{k=1}^{K} \int (F_n^k - F_n^k)^2 \, dF_n^k,$$

(13)
Fig. 4. (a) Density functions for various choices of $F_S$ (— distance 1; — distance 2; — distance 3; — distance 4), (b) $\gamma d_i(F_{\alpha, n}, F_{\beta, n})$ (— distance 1; — distance 2; — distance 3; — distance 4), the scaled second derivative (— distance 1; — distance 2; — distance 3; — distance 4), $\alpha_0^{CV, \beta, n}$ and $\alpha_0^{0.1k_n}$ (— distance 1; — distance 2; — distance 3; — distance 4) for five independent samples of size 5000 corresponding to distance 1 ($\alpha_0$), and means of various competing estimators of $\alpha_0^\gamma$ computed over 500 independent samples (— distance 1; — distance 2; — distance 3; — distance 4) for (c) distance 1 ($\alpha_0$, $\hat{\alpha}_0^{CV}$, $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{S,B}$) and (f) distance 4 (— distance 1; — distance 2; — distance 3; — distance 4) as the sample size increases from 3000 to $2 \times 10^6$.

where $\hat{F}^k := \hat{\alpha}_0^{\gamma, n} F_{\alpha, n}^{\gamma, n} - k + (1 - \hat{\alpha}_0^{\gamma, n}) F_{\beta}$. In all simulations in this paper, we use $K = 10$ and denote this estimator by $\hat{\alpha}_0^{CV};$ see section 7.10 of Hastie et al. (2009) for a more detailed study of cross-validation and a justification for $K = 10$. Fig. 4 illustrates the superior performance of $\hat{\alpha}_0^{CV}$ across different simulation settings; also see Sections 8.2.2 and 8.2.4, and section 13 (in the on-line supplementary material).

8.2.2. Performance under independence

In this subsection, we take $\alpha \in \{0.01, 0.03, 0.05, 0.10\}$ and compare the performance of the various estimators under the independence setting of scenarios A and B. In Tables 2 and 3, we give the mean and root-mean-squared error RMSE of the estimators over 5000 independent replications. For scenario A, we fix the sample size $n$ at 5000 and $\Sigma = I_{n \times n}$. For scenario B, we fix $n = 5 \times 10^4$, $L = 0$ and $m^* = 1$. By an application of lemma 4, it is easy to see that, in scenario A, the model is identifiable (i.e. $\alpha_0 = \alpha$), whereas, in scenario B, $\alpha_0 = \alpha \times 0.67$. For scenario A, the sample means of $\hat{\alpha}_0^{CV}$, $\tilde{\alpha}_0$, $\hat{\alpha}_0^{L}$, and $\hat{\alpha}_0^{0.1k_n}$ for $k_n = \log\log(n)$ are comparable. However,
the RMSEs of $\tilde{\alpha}_0$ and $\tilde{\alpha}_0^{0.1kn}$ are lower than those of $\tilde{\alpha}_0^C$, $\tilde{\alpha}_0^J$ and $\tilde{\alpha}_0^L$. For scenario B, the sample means of $\tilde{\alpha}_0$, $\tilde{\alpha}_0^C$ and $\tilde{\alpha}_0^{0.1kn}$ are comparable. In scenario B, the performances of $\tilde{\alpha}_0^J$ and $\tilde{\alpha}_0^{CJ}$ are not comparable with the estimators that are proposed in this paper, as $\tilde{\alpha}_0^J$ and $\tilde{\alpha}_0^{CJ}$ estimate $\alpha$, whereas $\tilde{\alpha}_0$, $\tilde{\alpha}_0^C$ and $\tilde{\alpha}_0^{CS}$ estimate $\alpha_0$. Note that $\tilde{\alpha}_0^L$ fails to estimate $\alpha_0$ because the underlying assumption that is inherent in their estimation procedure, that $f_0$ be non-increasing, does not hold. In scenario A, $\tilde{\alpha}_0^{S,0.5}$ has the best performance among the different values of $\lambda$, whereas, in scenario B, $\tilde{\alpha}_0^{S,\lambda}$ has poor performance for all values of $\lambda \in [0, 1]$. Furthermore, $\tilde{\alpha}_0^{GW}$, $\tilde{\alpha}_0^{MR}$, $\tilde{\alpha}_0^{CJ}$, $\tilde{\alpha}_0^{S,B}$ and $\tilde{\alpha}_0^E$ perform poorly in both scenarios for all values of $\alpha_0$.

8.2.3. Performance under dependence

The simulation settings of this subsection are designed to investigate the effect of dependence on the performance of the estimators. For scenario A, we use the setting of Langaas et al. (2005). We take $\Sigma$ to be a block diagonal matrix with block size 100. Within blocks, the diagonal elements (i.e. variances) are set to 1 and the off-diagonal elements (within-block correlations) are set to $\rho = 0.5$. Outside the blocks, all entries are set to 0. Tables 4 and 5 show that, in both scenarios, none of the methods perform well for small values of $\alpha_0$. However, in scenario A, the performances of $\tilde{\alpha}_0^{0.1kn}$, $\tilde{\alpha}_0$ and $\tilde{\alpha}_0^J$ are comparable, for larger values of $\alpha_0$. In scenario B, $\tilde{\alpha}_0^{0.1kn}$ performs well for $\alpha_0 = 0.033$ and $\alpha_0 = 0.067$. Observe that, as in the independence setting, $\tilde{\alpha}_0^{GW}$, $\tilde{\alpha}_0^{MR}$, $\tilde{\alpha}_0^{S,B}$, $\tilde{\alpha}_0^{CJ}$ and $\tilde{\alpha}_0^E$ perform poorly in both scenarios for all values of $\alpha_0$.

### Table 2. Means × 10 and RMSEs ×100 (in parentheses) of estimators discussed in Section 8.2 for scenario A with $\Sigma = L_7 \times n$, $J = 10$, $n = 5000$ and $k_n = \log(n)$

<table>
<thead>
<tr>
<th>$10\alpha_0$</th>
<th>$\hat{\alpha}_0^{0.1kn}$</th>
<th>$\hat{\alpha}_0^C$</th>
<th>$\hat{\alpha}_0$</th>
<th>$\hat{\alpha}_0^GW$</th>
<th>$\hat{\alpha}_0^{MR}$</th>
<th>$\hat{\alpha}_0^{S,0.5}$</th>
<th>$\hat{\alpha}_0^J$</th>
<th>$\hat{\alpha}_0^{CJ}$</th>
<th>$\hat{\alpha}_0^L$</th>
<th>$\hat{\alpha}_0^E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.13</td>
<td>0.15</td>
<td>0.13</td>
<td>0.00</td>
<td>0.01</td>
<td>0.09</td>
<td>0.14</td>
<td>0.05</td>
<td>0.16</td>
<td>0.36</td>
</tr>
<tr>
<td>(1.00)</td>
<td>(1.79)</td>
<td>(0.83)</td>
<td>(1.00)</td>
<td>(0.88)</td>
<td>(1.41)</td>
<td>(1.50)</td>
<td>(5.32)</td>
<td>(1.20)</td>
<td>(3.70)</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.30</td>
<td>0.35</td>
<td>0.27</td>
<td>0.02</td>
<td>0.12</td>
<td>0.29</td>
<td>0.29</td>
<td>0.15</td>
<td>0.35</td>
<td>0.36</td>
</tr>
<tr>
<td>(1.02)</td>
<td>(1.87)</td>
<td>(1.01)</td>
<td>(2.80)</td>
<td>(1.84)</td>
<td>(1.41)</td>
<td>(1.83)</td>
<td>(5.46)</td>
<td>(1.26)</td>
<td>(3.96)</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.48</td>
<td>0.51</td>
<td>0.46</td>
<td>0.18</td>
<td>0.26</td>
<td>0.47</td>
<td>0.49</td>
<td>0.26</td>
<td>0.55</td>
<td>0.35</td>
</tr>
<tr>
<td>(1.09)</td>
<td>(1.9)</td>
<td>(1.12)</td>
<td>(3.29)</td>
<td>(2.46)</td>
<td>(1.49)</td>
<td>(1.91)</td>
<td>(5.73)</td>
<td>(1.34)</td>
<td>(3.80)</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.93</td>
<td>0.97</td>
<td>0.93</td>
<td>0.62</td>
<td>0.65</td>
<td>0.95</td>
<td>0.96</td>
<td>0.51</td>
<td>1.02</td>
<td>0.33</td>
</tr>
<tr>
<td>(1.35)</td>
<td>(1.86)</td>
<td>(1.32)</td>
<td>(3.88)</td>
<td>(3.57)</td>
<td>(1.51)</td>
<td>(1.94)</td>
<td>(7.16)</td>
<td>(1.36)</td>
<td>(3.73)</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. Means × 10 and RMSEs ×100 (in parentheses) of estimators discussed in Section 8.2 for scenario B with $L = 0$, $m = 1$, $n = 5 \times 10^4$ and $k_n = \log(\log(n))$

<table>
<thead>
<tr>
<th>$10\alpha_0$</th>
<th>$\hat{\alpha}_0^{0.1kn}$</th>
<th>$\hat{\alpha}_0^C$</th>
<th>$\hat{\alpha}_0$</th>
<th>$\hat{\alpha}_0^GW$</th>
<th>$\hat{\alpha}_0^{MR}$</th>
<th>$\hat{\alpha}_0^{S,B}$</th>
<th>$\hat{\alpha}_0^J$</th>
<th>$\hat{\alpha}_0^{CJ}$</th>
<th>$\hat{\alpha}_0^L$</th>
<th>$\hat{\alpha}_0^E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.07</td>
<td>0.03</td>
<td>0.04</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.11</td>
<td>0.19</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>(0.44)</td>
<td>(0.67)</td>
<td>(0.28)</td>
<td>(0.66)</td>
<td>(0.65)</td>
<td>(0.96)</td>
<td>(2.96)</td>
<td>(0.38)</td>
<td>(0.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.14</td>
<td>0.18</td>
<td>0.16</td>
<td>0.00</td>
<td>0.01</td>
<td>0.08</td>
<td>0.28</td>
<td>0.55</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>(0.73)</td>
<td>(0.79)</td>
<td>(0.62)</td>
<td>(1.98)</td>
<td>(1.89)</td>
<td>(2.23)</td>
<td>(1.33)</td>
<td>(4.41)</td>
<td>(1.26)</td>
<td>(1.28)</td>
<td></td>
</tr>
<tr>
<td>0.33</td>
<td>0.25</td>
<td>0.31</td>
<td>0.28</td>
<td>0.02</td>
<td>0.04</td>
<td>0.12</td>
<td>0.48</td>
<td>0.92</td>
<td>0.12</td>
<td>0.05</td>
</tr>
<tr>
<td>(0.89)</td>
<td>(0.85)</td>
<td>(0.95)</td>
<td>(3.15)</td>
<td>(2.91)</td>
<td>(3.83)</td>
<td>(1.77)</td>
<td>(6.48)</td>
<td>(2.14)</td>
<td>(1.90)</td>
<td></td>
</tr>
<tr>
<td>0.66</td>
<td>0.55</td>
<td>0.62</td>
<td>0.58</td>
<td>0.12</td>
<td>0.14</td>
<td>0.23</td>
<td>0.95</td>
<td>1.83</td>
<td>0.23</td>
<td>0.05</td>
</tr>
<tr>
<td>(1.21)</td>
<td>(1.00)</td>
<td>(1.48)</td>
<td>(5.38)</td>
<td>(5.25)</td>
<td>(7.73)</td>
<td>(3.04)</td>
<td>(11.98)</td>
<td>(4.34)</td>
<td>(3.84)</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Means $\times 10$ and RMSEs $\times 100$ (in parentheses) of estimators discussed in Section 8.2 for scenario A with $\Sigma$ as described in Section 8.2.3, $J = 10$, $n = 5000$ and $k_n = \log(\log(n))$

<table>
<thead>
<tr>
<th>$10\alpha_0$</th>
<th>$\hat{\alpha}_0^{0.1k_n}$</th>
<th>$\hat{\alpha}_0^{0.01k_n}$</th>
<th>$\hat{\alpha}_0^{CV}$</th>
<th>$\hat{\alpha}_0^{GW}$</th>
<th>$\hat{\alpha}_0^{MR}$</th>
<th>$\hat{\alpha}_0^{S0.5}$</th>
<th>$\hat{\alpha}_0^J$</th>
<th>$\hat{\alpha}_0^{\text{CJ}}$</th>
<th>$\hat{\alpha}_0^L$</th>
<th>$\hat{\alpha}_0^E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.46</td>
<td>0.42</td>
<td>0.33</td>
<td>0.07</td>
<td>0.06</td>
<td>0.28</td>
<td>0.22</td>
<td>0.07</td>
<td>0.32</td>
<td>0.37</td>
</tr>
<tr>
<td>(5.15)</td>
<td>(4.23)</td>
<td>(3.84)</td>
<td>(1.72)</td>
<td>(1.27)</td>
<td>(1.17)</td>
<td>(6.58)</td>
<td>(3.25)</td>
<td>(10.61)</td>
<td>(4.37)</td>
<td>(3.91)</td>
</tr>
<tr>
<td>0.30</td>
<td>0.52</td>
<td>0.53</td>
<td>0.41</td>
<td>0.14</td>
<td>0.17</td>
<td>0.65</td>
<td>0.34</td>
<td>0.15</td>
<td>0.49</td>
<td>0.39</td>
</tr>
<tr>
<td>(3.80)</td>
<td>(3.64)</td>
<td>(3.59)</td>
<td>(2.72)</td>
<td>(1.90)</td>
<td>(6.58)</td>
<td>(3.25)</td>
<td>(10.35)</td>
<td>(4.30)</td>
<td>(4.31)</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.66</td>
<td>0.76</td>
<td>0.54</td>
<td>0.26</td>
<td>0.31</td>
<td>0.54</td>
<td>0.49</td>
<td>0.25</td>
<td>0.66</td>
<td>0.37</td>
</tr>
<tr>
<td>(3.52)</td>
<td>(5.43)</td>
<td>(3.85)</td>
<td>(3.56)</td>
<td>(2.50)</td>
<td>(2.61)</td>
<td>(3.60)</td>
<td>(10.45)</td>
<td>(4.31)</td>
<td>(4.03)</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>1.06</td>
<td>1.13</td>
<td>0.97</td>
<td>0.68</td>
<td>0.69</td>
<td>1.15</td>
<td>0.97</td>
<td>0.53</td>
<td>1.11</td>
<td>0.36</td>
</tr>
<tr>
<td>(3.09)</td>
<td>(3.92)</td>
<td>(4.00)</td>
<td>(4.15)</td>
<td>(3.54)</td>
<td>(6.01)</td>
<td>(3.61)</td>
<td>(10.55)</td>
<td>(4.13)</td>
<td>(3.99)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Means $\times 10$ and RMSEs $\times 100$ (in parentheses) of estimators discussed in Section 8.2 for scenario B with $L = 30$, $m = 1$, $n = 5 \times 10^4$ and $k_n = \log(\log(n))$

<table>
<thead>
<tr>
<th>$10\alpha_0$</th>
<th>$\hat{\alpha}_0^{0.1k_n}$</th>
<th>$\hat{\alpha}_0^{CV}$</th>
<th>$\hat{\alpha}_0^G$</th>
<th>$\hat{\alpha}_0^{MR}$</th>
<th>$\hat{\alpha}_0^{SB}$</th>
<th>$\hat{\alpha}_0^J$</th>
<th>$\hat{\alpha}_0^{\text{CJ}}$</th>
<th>$\hat{\alpha}_0^L$</th>
<th>$\hat{\alpha}_0^E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.07</td>
<td>0.29</td>
<td>0.38</td>
<td>0.17</td>
<td>0.04</td>
<td>0.05</td>
<td>0.26</td>
<td>0.20</td>
<td>0.21</td>
<td>0.13</td>
</tr>
<tr>
<td>(2.92)</td>
<td>(3.70)</td>
<td>(1.62)</td>
<td>(1.02)</td>
<td>(1.36)</td>
<td>(3.71)</td>
<td>(2.80)</td>
<td>(9.87)</td>
<td>(1.75)</td>
<td>(2.22)</td>
</tr>
<tr>
<td>0.20</td>
<td>0.30</td>
<td>0.42</td>
<td>0.18</td>
<td>0.04</td>
<td>0.04</td>
<td>0.16</td>
<td>0.33</td>
<td>0.55</td>
<td>0.13</td>
</tr>
<tr>
<td>(1.84)</td>
<td>(2.88)</td>
<td>(1.25)</td>
<td>(1.75)</td>
<td>(1.71)</td>
<td>(2.24)</td>
<td>(3.25)</td>
<td>(10.35)</td>
<td>(1.42)</td>
<td>(2.27)</td>
</tr>
<tr>
<td>0.33</td>
<td>0.38</td>
<td>0.52</td>
<td>0.20</td>
<td>0.06</td>
<td>0.06</td>
<td>0.17</td>
<td>0.50</td>
<td>0.93</td>
<td>0.16</td>
</tr>
<tr>
<td>(1.54)</td>
<td>(2.74)</td>
<td>(1.89)</td>
<td>(2.83)</td>
<td>(2.73)</td>
<td>(3.51)</td>
<td>(3.71)</td>
<td>(11.52)</td>
<td>(2.03)</td>
<td>(2.59)</td>
</tr>
<tr>
<td>0.67</td>
<td>0.63</td>
<td>0.77</td>
<td>0.31</td>
<td>0.14</td>
<td>0.15</td>
<td>0.24</td>
<td>0.95</td>
<td>1.82</td>
<td>0.25</td>
</tr>
<tr>
<td>(1.53)</td>
<td>(2.25)</td>
<td>(4.32)</td>
<td>(5.26)</td>
<td>(5.13)</td>
<td>(7.60)</td>
<td>(4.54)</td>
<td>(15.13)</td>
<td>(4.23)</td>
<td>(4.08)</td>
</tr>
</tbody>
</table>

8.2.4. Comparing the performance of $\hat{\alpha}_0^{cn}$, $\hat{\alpha}_0^{CV}$ and $\tilde{\alpha}_0$

Although the heuristic estimator $\tilde{\alpha}_0$ performs quite well in most of the simulation settings that were considered, there are scenarios where $\tilde{\alpha}_0$ can fail to estimate $\alpha_0$ consistently. To illustrate this we consider four different CDFs $F_s$ and fix $F_b$ to be the uniform distribution on $(0, 1)$ (see Fig. 4(a)) and compare the performance of $\hat{\alpha}_0^{CV}$, $\tilde{\alpha}_0$ and $\hat{\alpha}_0^{0.1k_n}$ with the best performing competing estimators (in each setting).

We see that $\tilde{\alpha}_0$ may fail to estimate the elbow of $\gamma d_n(F_{s,n}, F_{s,n}^\gamma)$, as a function of $\gamma$, when $F_s$ has a multimodal density (see Figs 4(b) and 4(c)). Observe that $\hat{\alpha}_0^{CV}$ and $\hat{\alpha}_0^{0.1k_n}$ perform favourably compared with all competing estimators and, in the two scenarios where $\tilde{\alpha}_0$ fails to estimate $\alpha_0$ consistently, all our competing estimators also fail.

The first two toy examples have been carefully constructed to demonstrate situations where the point of maximum curvature ($\tilde{\alpha}_0$) is different from the elbow of the function; see Fig. 4(b) (also see section 13 of the on-line supplementary material for further such examples).

8.2.5. Our recommendation

In this paper we study two estimators for $\alpha_0$. For $\hat{\alpha}_0^{cn}$, a proper choice of $c_n$ is important for good finite sample performance. We suggest using cross-validation to find the optimal tuning parameter $c_n$. However, cross-validation can be computationally expensive. An attractive alternative in this situation is to use $\tilde{\alpha}_0$, which is easy to implement and has very good finite sample performance in most scenarios, especially with large sample sizes. We feel that a visual analysis...
Table 6. Estimates of $\alpha_0$ for the two data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>$\hat{\alpha}_0^{0.10h}$</th>
<th>$\hat{\alpha}_0^{CV}$</th>
<th>$\hat{\alpha}_0^{GW}$</th>
<th>$\hat{\alpha}_0^{MR}$</th>
<th>$\hat{\alpha}_0^{SB}$</th>
<th>$\hat{\alpha}_0^J$</th>
<th>$\hat{\alpha}_0^{CJ}$</th>
<th>$\hat{\alpha}_0^L$</th>
<th>$\hat{\alpha}_0^E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
<td>0.04</td>
<td>0.01</td>
<td>0.19</td>
<td>0.10</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>Carina</td>
<td>0.36</td>
<td>0.35</td>
<td>0.36</td>
<td>0.31</td>
<td>0.30</td>
<td>0.45</td>
<td>0.61</td>
<td>1.00</td>
<td>0.38</td>
</tr>
</tbody>
</table>

†Not applicable.

of the plot of $\gamma_{dn}(\hat{F}_{s,n}^\gamma, \tilde{F}_{s,n}^\gamma)$ can be useful in checking the validity of $\tilde{\alpha}_0$ as an estimator of the elbow, and thus for $\alpha_0$.

9. Real data analysis

9.1. Prostate cancer data

Genetic expression levels for $n = 6033$ genes were obtained for $m = 102$ men, $m_1 = 50$ normal control subjects and $m_2 = 52$ prostate cancer patients. Without going into the biology that is involved, the principal goal of the study was to discover a small number of ‘interesting’ genes, i.e. genes whose expression levels differ between the cancer and control patients. Such genes, once identified, might be further investigated for a causal link to prostate cancer development. The prostate data are a $6033 \times 102$ matrix $X$ having entries $x_{ij}$, the expression level for gene $i$ on patient $j$, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$, with $j = 1, 2, \ldots, 50$, for the normal controls, and $j = 51, 52, \ldots, 102$, for the cancer patients. Let $\tilde{x}_i(1)$ and $\tilde{x}_i(2)$ be the averages of $x_{ij}$ for the normal controls and for the cancer patients respectively, for gene $i$. The two-sample $t$-statistic for testing significance of gene $i$ is $t_i = (\tilde{x}_i(1) - \tilde{x}_i(2)) / s_i$, where $s_i$ is an estimate of the standard error of $\tilde{x}_i(1) - \tilde{x}_i(2)$, i.e.

$$s_i^2 = \left(\frac{1}{50} + \frac{1}{52}\right) \sum_{j=1}^{50} (x_{ij} - \tilde{x}_i(1))^2 + \sum_{j=51}^{102} (x_{ij} - \tilde{x}_i(2))^2 \div 100.$$

We work with the $p$-values obtained from the 6033 two-sided $t$-tests instead of the ‘$t$-values’ as then the distribution under the alternative will have a non-increasing density which we can estimate by using the method that was developed in Section 6.1. In our analysis we ignore the dependence of the $p$-values, which is only a moderately risky assumption for the prostate data; see chapters 2 and 8 of Efron (2010) for further analysis and justification. Fig. 5 show the plots of various quantities of interest, found by using the methodology that was developed in Section 6.1 and Section 7, for the prostate data example. The 95% lower confidence bound $\hat{\alpha}_L$ for these data is found to be 0.05. In Table 6, we display estimates of $\alpha_0$ based on the methods that were considered in this paper for the prostate data and the Carina data (which are described below).

9.2. Carina data—an application in astronomy

In this subsection we analyse the distribution of radial velocities RV of stars in Carina, a dwarf spheroidal galaxy. Such galaxies are low luminosity galaxies that are companions of the Milky Way. The data have been obtained by Magellan and Multiple Mirror telescopes (see Walker et al. (2007)) and consist of radial (line-of-sight) velocity measurements of $n = 1266$ stars from Carina, contaminated with Milky Way stars in the field of view. We would like to understand the distribution of RVs of stars in Carina. For the contaminating stars from the Milky Way...
Fig. 5. Plots for the prostate data: (a) histogram of the $p$-values (---, uniform(0,1) distribution); (b) $\gamma d_n(\hat{\gamma}_n, \hat{F}_n, \gamma_n)$ (---) overlaid with its scaled derivative (---) (, point of maximum curvature $\hat{\alpha}_0 = 0.088$); (c) $F_{\hat{\alpha},\gamma}^\gamma$ (---) and $F_{\hat{\alpha},\gamma}^\gamma$ (---); (d) $f_{\hat{\alpha},\gamma}$; (e) estimated LFDR $\hat{l}$ for $p$-values less than 0.05.
in the field of view we assume a non-Gaussian velocity distribution $F_b$ that is known from the Besancon Milky Way model (Robin et al., 2003), calculated along the line of sight to Carina.

The 95% lower confidence bound for $\alpha_0$ is found to be 0.323. Fig. 6(c) shows the estimate of $F_s$ and the closest (in terms of minimizing the $L_2(\hat{F}_s, \hat{F}_s)$ distance) fitting Gaussian distribution. Astronomers usually assume the distribution of the RVs for these dwarf spheroidal galaxies to be Gaussian. Indeed we see that the estimated $F_s$ is close to a normal distribution (with mean 222.9 and standard deviation 7.51), although a formal test of this hypothesis is beyond the scope of the present paper. The estimate due to Cai and Jin (2010), $\hat{\alpha}_C$, is greater than 1, whereas Efron’s method (see Efron (2007)), implemented by using the locfdr package in R, fails to estimate $\alpha_0$.

10. Concluding remarks

In this paper we develop procedures for estimating the mixing proportion and the unknown distribution in a two-component mixture model by using ideas from shape-restricted function estimation. We discuss the identifiability of the model and introduce an identifiable parameter $\alpha_0$, under minimal assumptions on the model. We propose an honest finite sample lower con-
From the definition of \( \alpha_0 \) that is distribution free. Two point estimators of \( \alpha_0, \hat{\alpha}_0^\alpha \), and \( \tilde{\alpha}_0 \), are studied. We prove that \( \hat{\alpha}_0^\alpha \) is a consistent estimator of \( \alpha_0 \) and show that the rate of convergence of \( \hat{\alpha}_0^\alpha \) can be arbitrarily close to \( \sqrt{n} \), for proper choices of \( c_n \). These proposed estimators crucially rely on \( \gamma d_n(\hat{F}_s^n, \hat{F}_s^n) \), as a function of \( \gamma \), whose plot provides useful insights about the nature of the problem and performance of the estimators.

We observe that the estimators of \( \alpha_0 \) that are proposed in this paper have superior finite sample performance than most competing methods. In contrast with most previous work on this topic the results that are discussed in this paper hold true even when model (1) is not identifiable. Under the assumption that model (1) is identifiable, we can find an estimator of \( F \) which is uniformly consistent. Furthermore, if \( F \) is known to have a non-increasing density \( f_b \) we can find a consistent estimator of \( f_b \). All these estimators are tuning parameter free and easily implementable.

We conclude this section by outlining some possible future research directions. Construction of two-sided confidence intervals for \( \alpha_0 \) remains a difficult problem as the asymptotic distribution of \( \hat{\alpha}_0^\alpha \) depends on the unknown \( F \). We are currently developing estimators of \( \alpha_0 \) when we do not exactly know \( F_b \) but only have an estimator of \( F_b \) (for example, we observe a second independent and identically distributed sample from \( F_b \)). Investigating consistent alternative ways of detecting the elbow of the function \( \gamma d_n(\hat{F}_s^n, \hat{F}_s^n) \), as an estimator of \( \tilde{\alpha}_0 \), is an interesting future research direction. As we have observed in the astronomy application, formal goodness-of-fit tests for \( F_b \) are important—they can guide the practitioner to use appropriate parametric models for further analysis—but are at present unknown. The \( p \)-values in the prostate data example, which was considered in Section 9.1, can have slight dependence. Therefore, investigating the performance and properties of the methods that were introduced in this paper under appropriate dependence assumptions on \( X_1, \ldots, X_n \) is another important direction for future research.

**Acknowledgements**

We thank the Joint Editor, the Associate Editor and five referees for their careful reading and constructive comments that led to an improved version of the paper.

**Appendix A**

A.1. **Proof of lemma 2**

From the definition of \( \alpha_0 \), we have

\[
\alpha_0 = \inf \{ 0 \leq \gamma \leq \alpha : \{ F - (1 - \gamma) F_b \}/\gamma \text{ is a valid CDF} \}
= \inf \{ 0 \leq \gamma \leq \alpha : \{ \alpha F_s + (1 - \alpha) F_b - (1 - \gamma) F_b \}/\gamma \text{ is a valid CDF} \}
= \inf \{ 0 \leq \gamma \leq \alpha : \{ \alpha F_s - (\alpha - \gamma) F_b \}/\gamma \text{ is a valid CDF} \}
= \alpha - \sup \{ 0 \leq \epsilon \leq \alpha : \alpha F_s - \epsilon F_b \text{ is a sub-CDF} \}
= \alpha - \sup \{ 0 \leq \epsilon \leq 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF} \},
\]

where the final equality follows from the fact that, if \( \epsilon > \alpha \), then \( \alpha F_s - \epsilon F_b \) will not be a sub-CDF.

To show that \( \alpha_0 = 0 \) if and only if \( F = F_b \) let us define \( \delta = \alpha - \epsilon \). Note that \( \alpha_0 = 0 \), if and only if

\[
\sup \{ 0 \leq \epsilon \leq 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF} \} = \alpha
\Rightarrow \inf \{ 0 \leq \delta \leq 1 : \alpha (F_s - F_b) + \delta F_b \text{ is a sub-CDF} \} = 0.
\]

However, it is easy to see that the last equality is true if and only if \( F_s - F_b \equiv 0 \).
A.2. Proof of lemma 6

Letting $F_a^* = \{ F - (1 - \gamma) F_b \} / \gamma$, observe that

$$\gamma d_n(\hat{F}_{s, n}^{\gamma}, F_a^*) = d_n(F, \hat{F}_{s, n}^{\gamma}).$$

Also note that $F_a^*$ is a valid CDF for $\gamma \geq \alpha_0$. As $\hat{F}_{s, n}^{\gamma}$ is defined as the function that minimizes the $L_2(\mathbb{F}_a)$ distance of $\hat{F}_{s, n}^{\gamma}$ over all CDFs,

$$\gamma d_n(\hat{F}_{s, n}^{\gamma}, \hat{F}_{s, n}^{\gamma}) \leq \gamma d_n(\hat{F}_{s, n}^{\gamma}, F_a^*) = d_n(F, \hat{F}_{s, n}^{\gamma}).$$

To prove the second part of lemma 6 note that for $\gamma \geq \alpha_0$ the result follows from above and the fact that $d_n(F, \hat{F}_{s, n}^{\gamma}) \to 0$ almost surely as $n \to \infty$. For $\gamma < \alpha_0$, $F_a^*$ is not a valid CDF, by the definition of $\alpha_0$. Note that as $n \to \infty$, $\hat{F}_{s, n}^{\gamma} \to F_a^*$ almost surely, pointwise. So, for sufficiently large $n$, $\hat{F}_{s, n}^{\gamma}$ is not a valid CDF, whereas $\hat{F}_{s, n}^{\gamma}$ is always a CDF. Thus, $d_n(\hat{F}_{s, n}^{\gamma}, \hat{F}_{s, n}^{\gamma})$ converges to something positive.

A.3. Proof of lemma 7

Assume that $\gamma_1 \leq \gamma_2$ and $\gamma_1, \gamma_2 \in A_n$. If $\gamma_3 = \gamma_1 \gamma_1 + (1 - \eta) \gamma_2$, for $0 \leq \eta \leq 1$, it is easy to observe from expression (2) that

$$\eta_1 \gamma_1 \hat{F}_{s, n}^{\gamma_1} + (1 - \eta) \gamma_2 \hat{F}_{s, n}^{\gamma_2} = \gamma_3 \hat{F}_{s, n}^{\gamma_3}.$$

Note that $\{ \eta_1 \gamma_1 \hat{F}_{s, n}^{\gamma_1} + (1 - \eta) \gamma_2 \hat{F}_{s, n}^{\gamma_2} \}/\gamma_3$ is a valid CDF and thus, from the definition of $\hat{F}_{s, n}^{\gamma_3}$, we have

$$d_n(\hat{F}_{s, n}^{\gamma_3}, \hat{F}_{s, n}^{\gamma_3}) \leq d_n(\hat{F}_{s, n}^{\gamma_3}, \{ \eta_1 \gamma_1 \hat{F}_{s, n}^{\gamma_1} + (1 - \eta) \gamma_2 \hat{F}_{s, n}^{\gamma_2} \}/\gamma_3)$$

$$= d_n\left( \frac{\eta_1 \gamma_1 \hat{F}_{s, n}^{\gamma_1} + (1 - \eta) \gamma_2 \hat{F}_{s, n}^{\gamma_2}}{\eta_1 \gamma_1 + (1 - \eta) \gamma_2}, \frac{\eta_1 \gamma_1 \hat{F}_{s, n}^{\gamma_1} + (1 - \eta) \gamma_2 \hat{F}_{s, n}^{\gamma_2}}{\eta_1 \gamma_1 + (1 - \eta) \gamma_2} \right)$$

$$\leq \frac{\eta_1 \gamma_1}{\gamma_3} d_n(\hat{F}_{s, n}^{\gamma_1}, \hat{F}_{s, n}^{\gamma_1}) + \frac{(1 - \eta) \gamma_2}{\gamma_3} d_n(\hat{F}_{s, n}^{\gamma_2}, \hat{F}_{s, n}^{\gamma_2})$$

(14)

where the last step follows from the triangle inequality. But, as $\gamma_1, \gamma_2 \in A_n$, inequality (14) yields

$$d_n(\hat{F}_{s, n}^{\gamma_3}, \hat{F}_{s, n}^{\gamma_3}) \leq \frac{\eta_1 \gamma_1}{\gamma_3} \frac{c_n}{\sqrt{n} \gamma_1} + \frac{(1 - \eta) \gamma_2}{\gamma_3} \frac{c_n}{\sqrt{n} \gamma_2} = \frac{c_n}{\sqrt{n} \gamma_3}.$$

Thus $\gamma_3 \in A_n$.

A.4. Proof of lemma 8

As $\alpha_0 = 0$,

$$P(\hat{c}_{\alpha_0} \leq 0) = 1 - P(\hat{c}_{\alpha_0} > 0) = 1 - P\{ \sqrt{n} d_n(\mathbb{F}_a, F) > c_n \} \to 1,$$

(15)

since $\sqrt{n} d_n(\mathbb{F}_a, F) = O_P(1)$ by theorem 6.

A.5. Proof of theorem 5

Letting $c_n = H_n^{-1}(1 - \beta)$, we have

$$P(\alpha_0 \geq \hat{\alpha}_L) = P\{ \sqrt{n} \alpha_0 d_n(\hat{F}_{s, n}^{\alpha_0}, F) \leq c_n \} \geq P\{ \sqrt{n} \alpha_0 d_n(\hat{F}_{s, n}^{\alpha_0}, F) \leq c_n \} = H_n(c_n) = 1 - \beta,$$

where we have used the fact that $\alpha_0 d_n(\hat{F}_{s, n}^{\alpha_0}, F) = d_n(\mathbb{F}_a, F)$. Note that, when $\alpha_0 = 0$, $F = F_b$, and using expression (9) we obtain

$$P(\alpha_0 \geq \hat{\alpha}_L) = P\{ \sqrt{n} \alpha_0 d_n(\mathbb{F}_a, F) \leq c_n \} = P\{ \sqrt{n} d_n(\mathbb{F}_a, F) \leq c_n \} = 1 - \beta.$$
A.6. Proof of lemma 9
Let $0 < \gamma_1 < \gamma_2 < 1$. Then,
\[
\gamma_2 d_n(\hat{F}_{s,n}^\gamma, \tilde{F}_{s,n}^\gamma) \leq \gamma_2 d_n(\hat{F}_{s,n}^\gamma, (\gamma_1/\gamma_2)\hat{F}_{s,n}^\gamma + (1 - \gamma_1/\gamma_2) F_b) = d_n(\gamma_1 \hat{F}_{s,n}^\gamma + (\gamma_2 - \gamma_1) F_b, \gamma_1 \hat{F}_{s,n}^\gamma + (\gamma_2 - \gamma_1) F_b) \leq \gamma_1 d_n(\hat{F}_{s,n}^\gamma, \hat{F}_{s,n}^\gamma),
\]
which shows that $\gamma d_n(\hat{F}_{s,n}^\gamma, \tilde{F}_{s,n}^\gamma)$ is a non-increasing function. To show that $\gamma d_n(\hat{F}_{s,n}^\gamma, \tilde{F}_{s,n}^\gamma)$ is convex, let $0 < \gamma_1 < \gamma_2 < 1$ and $\gamma_3 = \eta \gamma_1 + (1 - \eta) \gamma_2$, for $0 \leq \eta \leq 1$. Then, by inequality (14) we have the desired result.

A.7. Proof of theorem 9
Let $\epsilon_n := \sup_{x \in \mathbb{R}} |\tildes(\hat{F}_{s,n}^\gamma (x) - F_s(x))|$. Then the function $F_s + \epsilon_n$ is concave on $[0, \infty)$ and majorizes $\hat{F}_{s,n}^\gamma$. Hence, for all $x \in [0, \infty)$, $\hat{F}_{s,n}^\alpha (x) \leq F_s (x) \leq F_s (x) + \epsilon_n$, as $F_s$ is the LCM of $\hat{F}_{s,n}^\gamma$. Thus,
\[
-\epsilon_n \leq \hat{F}_{s,n}^\alpha (x) - F_s (x) \leq \hat{F}_{s,n}^\alpha (x) - F_s (x) \leq \epsilon_n,
\]
and, therefore,
\[
\sup_{x \in \mathbb{R}} |\tildes(\hat{F}_{s,n}^\gamma (x) - F_s(x))| \leq \epsilon_n.
\]
By theorem 7, as $\epsilon_n \to 0$, we must also have result (12).

The second part of the result follows immediately from the lemma on page 330 of Robertson et al. (1988) and is similar to the result in theorem 7.2.2 there.

References
Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplement to Estimation of a two-component mixture model with applications to multiple testing’.