ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)

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.

$$F(x) = \alpha\, F_{\mathrm{s}}(x) + (1-\alpha)\, F_{\mathrm{b}}(x), \tag{1}$$

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

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 α and the distribution of the p-values under the alternative. In addition, a reliable estimator of α 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,

Address for correspondence: Bodhisattva Sen, Department of Statistics, Columbia University, Room 1032, 1255 Amsterdam Avenue, New York, NY 10027, USA.
E-mail: bodhi@stat.columbia.edu

1369–7412/16/78000

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 χ^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 α 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 Langaas *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 α 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 α. We believe that this is the first attempt to construct a distribution-free lower confidence bound for α that is also tuning parameter free.
(c) Two estimators of α 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 α 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 α. 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 α. 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 α 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_{s} would be

$$\hat{F}^{\alpha}_{\mathrm{s},n} = \frac{\mathbb{F}_n - (1-\alpha)F_{\mathrm{b}}}{\alpha}, \tag{2}$$

where $\mathbb{F}_n$ is the empirical CDF of the observed sample, i.e. $\mathbb{F}_n(x) = \Sigma_{i=1}^n \mathbf{1}\{X_i \leqslant x\}/n$. Although this estimator is consistent, it does not satisfy the basic requirements of a CDF: $\hat{F}^{\alpha}_{\mathrm{s},n}$ 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

$$\int \{W(x) - \hat{F}^{\alpha}_{\mathrm{s},n}(x)\}^2 \,\mathrm{d}\mathbb{F}_n(x) \equiv \frac{1}{n}\sum_{i=1}^{n} \{W(X_i) - \hat{F}^{\alpha}_{\mathrm{s},n}(X_i)\}^2 \tag{3}$$

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

$$\Theta_{\mathrm{inc}} = \{\boldsymbol{\theta} \in \mathbb{R}^n : 0 \leqslant \theta_1 \leqslant \theta_2 \leqslant \ldots \leqslant \theta_n \leqslant 1\},$$

$\mathbf{V} = (V_1, V_2, \ldots, V_n)$, $V_i := \hat{F}^{\alpha}_{\mathrm{s},n}(X_{(i)})$, $i = 1, 2, \ldots, n$, $X_{(i)}$ being the ith order statistic of the sample, and $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^n$. The estimator $\hat{\boldsymbol{\theta}}$ 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 $\mathbf{V}$ on the closed convex set $\Theta_{\mathrm{inc}} \subset \mathbb{R}^n$. $\hat{\boldsymbol{\theta}}$ is related to $\check{F}^{\alpha}_{\mathrm{s},n}$ via $\check{F}^{\alpha}_{\mathrm{s},n}(X_{(i)}) = \hat{\theta}_i$ and can be easily computed by using the pool adjacent violators algorithm; see section 1.2 of Robertson *et al.* (1988). Thus, $\check{F}^{\alpha}_{\mathrm{s},n}$ 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 $\check{F}^{\alpha}_{\mathrm{s},n}$.

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

When α is unknown, the problem is considerably more difficult; in fact, it is non-identifiable. If model (1) holds for some F_{b} and α then the mixture model can be rewritten as

$$F = (\alpha+\gamma)\left(\frac{\alpha}{\alpha+\gamma}F_{\mathrm{s}} + \frac{\gamma}{\alpha+\gamma}F_{\mathrm{b}}\right) + (1-\alpha-\gamma)F_{\mathrm{b}},$$

for $0 \leqslant \gamma \leqslant 1-\alpha$, and the term $(\alpha F_{\mathrm{s}} + \gamma F_{\mathrm{b}})/(\alpha+\gamma)$ 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 $W = \mathbb{F}_n$. Hence, α is not uniquely defined. To handle the identifiability issue, we redefine the mixing proportion as

$$\alpha_0 := \inf\{\gamma \in (0,1] : \{F - (1-\gamma)F_{\mathrm{b}}\}/\gamma \text{ is a CDF}\}. \tag{4}$$

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 α_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 α_0 and investigating the difference between α and α_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 α satisfying model (1). As seen from the above discussion we can only hope to estimate α_0, which, from its definition in expression (4), is smaller than α, i.e. $\alpha_0 \leqslant \alpha$. 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 α and α_0.

Lemma 2. Let F be as in model (1) and α_0 as defined in expression (4). Then

$$\alpha_0 = \alpha - \sup\{0 \leqslant \epsilon \leqslant 1 : \alpha F_{\mathrm{s}} - \epsilon F_{\mathrm{b}} \text{ is a sub-CDF}\}, \tag{5}$$

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 $\alpha F_{\mathrm{s}} - \epsilon F_{\mathrm{b}}$ is a sub-CDF. Furthermore, $\alpha_0 = 0$ if and only if $F = F_{\mathrm{b}}$.

In what follows we separately identify α_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 $G^{(\mathrm{d})}$, an absolutely continuous CDF $G^{(\mathrm{a})}$ and a continuous but singular CDF $G^{(\mathrm{s})}$, i.e. $G = \eta_1 G^{(\mathrm{a})} + \eta_2 G^{(\mathrm{d})} + \eta_3 G^{(\mathrm{s})}$, where $\eta_i \geqslant 0$, 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

$$G = \eta G^{(\mathrm{a})} + (1-\eta) G^{(\mathrm{d})}, \tag{6}$$

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 $J_G : d(G) \rightarrow [0,1]$ 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 $d(F_{\mathrm{b}}) \not\subset d(F_{\mathrm{s}})$, then $\alpha_0 = \alpha$, i.e. model (1) is identifiable. If $d(F_\mathrm{b}) \subset d(F_\mathrm{s})$, then $\alpha_0 = \alpha\{1 - \inf_{x \in d(F_\mathrm{b})} J_{F_\mathrm{s}}(x)/J_{F_\mathrm{b}}(x)\}$. Thus, $\alpha_0 = \alpha$ if and only if $\inf_{x \in d(F_\mathrm{b})} J_{F_\mathrm{s}}(x)/J_{F_\mathrm{b}}(x) = 0$.

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

$$\alpha_0 = \alpha\left(1 - \operatorname{ess\,inf} \frac{f_\mathrm{s}}{f_\mathrm{b}}\right),$$

where, for any function g, $\operatorname{ess\,inf} g = \sup\{a \in \mathbb{R} : m\{\{x : g(x) < a\}\} = 0\}$, m being the Lebesgue measure. As a consequence, $\alpha_0 < \alpha$ if and only if there exists $c > 0$ such that $f_\mathrm{s} \geqslant c f_\mathrm{b}$, almost everywhere m.

Lemma 4 states that if there does not exist any $c > 0$ for which $f_\mathrm{s}(x) \geqslant c\, f_\mathrm{b}(x)$, 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 α.

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

3. Estimation

3.1. Estimation of the mixing proportion α_0

In this section we consider the estimation of α_0 as defined in equation (5). For the rest of the paper, unless otherwise noted, we assume that

$X_1, X_2, \ldots, X_n$ is an independent and identically distributed sample from F as in model (1).

Recall the definitions of $\hat{F}^\gamma_{\mathrm{s},n}$ and $\check{F}^\gamma_{\mathrm{s},n}$, for $\gamma \in (0,1]$; see expressions (2) and (3). When $\gamma = 1$, we have $\hat{F}^\gamma_{\mathrm{s},n} = \mathbb{F}_n = \check{F}^\gamma_{\mathrm{s},n}$ as $\hat{F}^\gamma_{\mathrm{s},n}$ (for $\gamma = 1$) is a CDF, whereas, when γ is much smaller than α_0, the regularization of $\hat{F}^\gamma_{\mathrm{s},n}$ modifies it, and thus $\hat{F}^\gamma_{\mathrm{s},n}$ and $\check{F}^\gamma_{\mathrm{s},n}$ are quite different. We would like to compare the naive and isotonized estimators $\hat{F}^\gamma_{\mathrm{s},n}$ and $\check{F}^\gamma_{\mathrm{s},n}$ respectively, and to choose the smallest γ for which their distance is still small. This leads to the following estimator of α_0:

$$\hat{\alpha}_0^{c_n} = \inf\left\{\gamma \in (0,1] : \gamma\, d_n(\hat{F}^\gamma_{\mathrm{s},n}, \check{F}^\gamma_{\mathrm{s},n}) \leqslant \frac{c_n}{\sqrt{n}}\right\}, \tag{7}$$

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

$$d_n\{\mathbb{F}_n, \gamma \check{F}^\gamma_{\mathrm{s},n} + (1-\gamma) F_\mathrm{b}\} = \gamma\, d_n(\hat{F}^\gamma_{\mathrm{s},n}, \check{F}^\gamma_{\mathrm{s},n}). \tag{8}$$

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

$$\lim_{\gamma \to 0+} \gamma\, d_n(\hat{F}^\gamma_{\mathrm{s},n}, \check{F}^\gamma_{\mathrm{s},n}) = d_n(\mathbb{F}_n, F_\mathrm{b}). \tag{9}$$

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 α_0. We shall also show that particular (distribution-free) choices of c_n will lead to honest lower confidence bounds for α_0.

Next, we prove a result which implies that, in the multiple-testing problem, estimators of α_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 $\Psi^{-1}(y):=\inf\{t\in\mathbb{R}: y\leqslant\Psi(t)\}$, 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 $G:=\alpha F_\mathrm{s}\circ\Psi+(1-\alpha)F_\mathrm{b}\circ\Psi$. 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 α. We can define α_0^Y as in equation (4) but with $\{G,F_\mathrm{b}\circ\Psi\}$ instead of $\{F,F_\mathrm{b}\}$. The following result, which is proved in the on-line supplementary material, shows that the α_0 and its estimators, proposed in this paper, are invariant under such monotonic transformations.

Theorem 1. Let $\mathbb{G}_n$ be the empirical CDF of $Y_1,Y_2,\ldots,Y_n$. Also, let $\hat{G}_{\mathrm{s},n}$ and $\check{G}_{\mathrm{s},n}^{\gamma}$ be as defined in expressions (2) and (3) respectively, but with $\{\mathbb{G}_n,F_\mathrm{b}\circ\Psi\}$ instead of $\{\mathbb{F}_n,F_\mathrm{b}\}$. Then $\alpha_0=\alpha_0^Y$ and $\gamma d_n(\hat{F}_{\mathrm{s},n}^{\gamma},\check{F}_{\mathrm{s},n}^{\gamma})=\gamma d_n(\hat{G}_{\mathrm{s},n}^{\gamma},\check{G}_{\mathrm{s},n}^{\gamma})$ 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 $\gamma d_n(\check{F}_{\mathrm{s},n}^{\gamma},\hat{F}_{\mathrm{s},n}^{\gamma})$.

Lemma 6. For $1\geqslant\gamma\geqslant\alpha_0$, $\gamma d_n(\check{F}_{\mathrm{s},n}^{\gamma},\hat{F}_{\mathrm{s},n}^{\gamma})\leqslant d_n(F,\mathbb{F}_n)$. Thus,

$$\gamma d_n(\hat{F}_{\mathrm{s},n}^{\gamma},\check{F}_{\mathrm{s},n}^{\gamma})\to\begin{cases}0, & \gamma-\alpha_0\geqslant 0,\\ >0, & \gamma-\alpha_0<0,\end{cases}\tag{10}$$

almost surely.

Lemma 7. The set $A_n:=\{\gamma\in[0,1]:\sqrt{n}\gamma d_n(\hat{F}_{\mathrm{s},n}^{\gamma},\check{F}_{\mathrm{s},n}^{\gamma})\leqslant c_n\}$ 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 $\hat{\alpha}_0^{c_n}\to^{\mathrm{P}}\alpha_0$.

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 α_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 α_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 $\hat{\alpha}_0^{c_n}$.

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 $r_n(\hat{\alpha}_0^{c_n} - \alpha_0) = O_P(1)$.

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 α_0 that were proposed by Hengartner and Stark (1995) and Swanepoel (1999) have rates of convergence $\{n/\log(n)\}^{1/3}$ and $n^{2/5}/\log(n)^{\delta}$, for $\delta > 0$, respectively. Morover, 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 $F_s^{\alpha_0}$ vanishes at a set of points of measure 0 and satisfies certain regularity assumptions, then any $\sqrt{n}$-consistent estimator of α_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 ∞ 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, $r_n(\hat{\alpha}_0^{c_n} - \alpha_0)$ converges to a degenerate limit. In Section 8.2, we analyse the effect of c_n on the finite sample performance of $\hat{\alpha}_0^{c_n}$ for estimating α_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

$$r_n(\hat{\alpha}_0^{c_n} - \alpha_0) \xrightarrow{\mathrm{P}} c,$$

where $c < 0$ is a constant that depends on α_0, F and F_b.

4. Lower confidence bound for α_0

The asymptotic limit of the estimator $\hat{\alpha}_0^{c_n}$ that was discussed in Section 3 depends on unknown parameters (e.g. α_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

$$P(\alpha_0 \geqslant \hat{\alpha}_L) \geqslant 1 - \beta, \tag{11}$$

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 β 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. $P(\hat{\alpha}_{\mathrm{L}} = 0) = 1 - \beta$. 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 $\sqrt{n}\, d_n(\mathbb{F}_n, F)$. Let $\hat{\alpha}_{\mathrm{L}}$ be defined as in equation (7) with $c_n = H_n^{-1}(1-\beta)$. Then inequality (11) holds. Furthermore if $\alpha_0 = 0$, then $P(\hat{\alpha}_{\mathrm{L}} = 0) = 1 - \beta$, 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. $n \geqslant 500$) the distribution H_n can be very well approximated by that of the Cramér–von Mises statistic, defined as

$$\sqrt{n}\, d(\mathbb{F}_n, F) := \sqrt{\int n\{\mathbb{F}_n(x) - F(x)\}^2 \,\mathrm{d}F(x)}.$$

Letting G_n be the CDF of $\sqrt{n}\, d(\mathbb{F}_n, F)$, 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

$$P(\alpha_0 \geqslant \hat{\alpha}_{\mathrm{L}}) = P\{\sqrt{n}\alpha_0\, d_n(\hat{F}^{\alpha_0}_{\mathrm{s},n}, \check{F}^{\alpha_0}_{\mathrm{s},n}) \geqslant H_n^{-1}(1-\beta)\}.$$

The following theorem gives the explicit asymptotic limit of $P(\alpha_0 \geqslant \hat{\alpha}_{\mathrm{L}})$ but it is not useful for practical purposes as it involves the unknown $F_{\mathrm{s}}^{\alpha_0}$ and F.

Theorem 7. Assume that $\alpha_0 > 0$. Then $\sqrt{n}\alpha_0\, d_n(\hat{F}^{\alpha_0}_{\mathrm{s},n}, \check{F}^{\alpha_0}_{\mathrm{s},n}) \to^{\mathrm{d}} U$, where U is a random variable whose distribution depends only on α_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 α_0

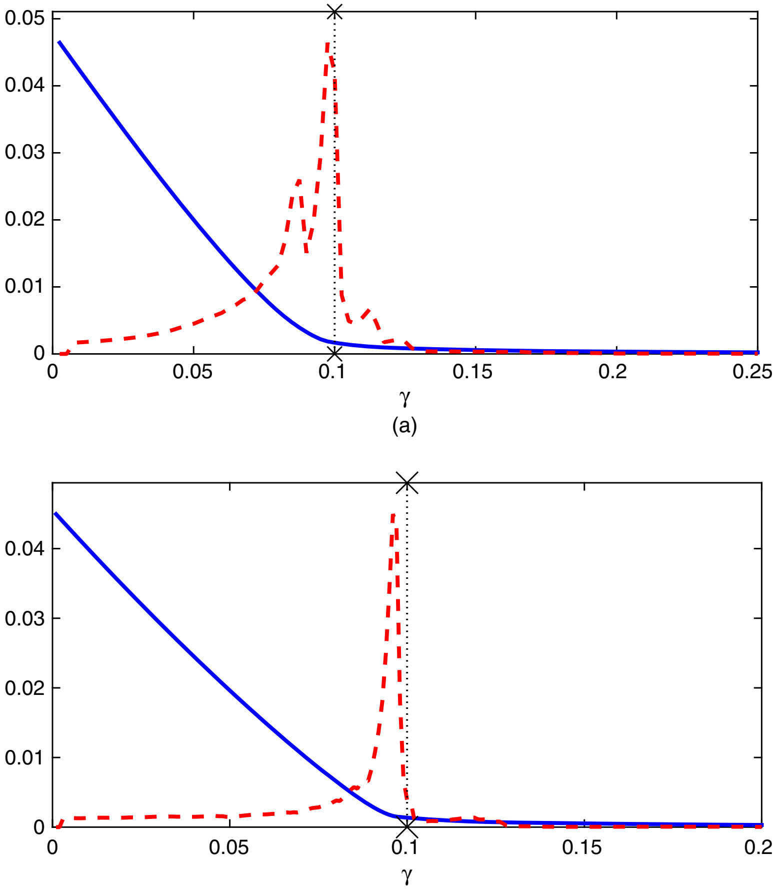
In simulations, we observe that the finite sample performance of equation (7) is affected by the choice of c_n (for an extensive simulation study on this see Section 8.2). This motivates us to propose a method to estimate α_0 that is completely automated and has good finite sample performance. We start with a lemma, which is proved in Appendix A, that describes the shape of our criterion function and will motivate our procedure.

Lemma 9. $\gamma\, d_n(\hat{F}^{\gamma}_{\mathrm{s},n}, \check{F}^{\gamma}_{\mathrm{s},n})$ is a non-increasing convex function of γ in $(0, 1)$.

Writing

$$\hat{F}^{\gamma}_{\mathrm{s},n} = \frac{\mathbb{F}_n - F}{\gamma} + \frac{\alpha_0}{\gamma} F_{\mathrm{s}}^{\alpha_0} + \left(1 - \frac{\alpha_0}{\gamma}\right) F_{\mathrm{b}},$$

we see that, for $\gamma \geqslant \alpha_0$, the second term on the right-hand side is a CDF. Thus, for $\gamma \geqslant \alpha_0$, $\hat{F}^{\gamma}_{\mathrm{s},n}$ is



(b)

Fig. 1. Plots of $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (——) overlaid with its (scaled) second derivative (- - -) for $\alpha_0 = 0.1$ and $n = 5000$: (a) setting I; (b) setting II

very close to a CDF as $\mathbb{F}_n - F = O_P(n^{-1/2})$, and hence $\check{F}^{\gamma}_{s,n}$ should also be close to $\hat{F}^{\gamma}_{s,n}$. whereas, for $\gamma < \alpha_0$, $\hat{F}^{\gamma}_{s,n}$ is not close to a CDF, and thus the distance $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is appreciably large. Therefore, at α_0, we have a 'regime' change: $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ should have a slowly decreasing segment to the right of α_0 and a steeply non-increasing segment to the left of α_0. Fig. 1 shows two typical such plots of the function $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$, 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 α_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 $\hat{\alpha}_0^{c_n}$ are derived from $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$, as a function of γ, albeit they look at two different aspects of the function.

In Fig. 1 we have used numerical methods to approximate the second derivative of $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (using the method of double differencing). We advocate plotting the function $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ as γ varies between 0 and 1. In most cases, plots similar to Fig. 1 would immediately convey to the practitioner the most appropriate choice of $\tilde{\alpha}_0$. 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 $\tilde{\alpha}_0$ fails to estimate the elbow of $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ 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 $F_s^{\alpha_0} = F_s$. Once we have a consistent estimator $\check{\alpha}_n$ (which may or may not be $\check{\alpha}_0^{c_n}$ as discussed in the previous sections) of α_0, a natural non-parametric estimator of F_s is $\check{F}_{s,n}^{\check{\alpha}_n}$, defined as the minimizer of expression (3). In the following theorem (which is proved in the on-line supplementary material) we show that, indeed, $\check{F}_{s,n}^{\check{\alpha}_n}$ is uniformly consistent for estimating F_s. We also derive the rate of convergence of $\check{F}_{s,n}^{\check{\alpha}_n}$.

Theorem 8. Suppose that $\check{\alpha}_n \rightarrow^{P} \alpha_0$. Then, as $n \rightarrow \infty$, $\sup_{x\in\mathbb{R}} |\check{F}_{s,n}^{\check{\alpha}_n}(x) - F_s(x)| \rightarrow^{P} 0$. Furthermore, if $q_n(\check{\alpha}_n - \alpha_0) = O_P(1)$, where $q_n = o(\sqrt{n})$, then $\sup_{x\in\mathbb{R}} q_n |\check{F}_{s,n}^{\check{\alpha}_n}(x) - F_s(x)| = O_P(1)$. Additionally, for $\hat{\alpha}_0^{c_n}$ as defined in expression (7), we have

$$\sup_{x\in\mathbb{R}} |r_n(\hat{F}_{s,n}^{\hat{\alpha}_0^{c_n}} - F_s)(x) - Q(x)| \xrightarrow{P} 0,$$

$$r_n\, d(\check{F}_{s,n}^{\hat{\alpha}_0^{c_n}}, F_s) \xrightarrow{P} c$$

for a function $Q: \mathbb{R} \rightarrow \mathbb{R}$ and a constant $c > 0$ depending only on α_0, F and F_b.

An immediate consequence of theorem 8 is that $d_n(\check{F}_{s,n}^{\check{\alpha}_n}, \hat{F}_{s,n}^{\check{\alpha}_n}) \rightarrow^{P} 0$ as $n \rightarrow \infty$. Fig. 2(a) shows our estimator $\check{F}_{s,n}^{\check{\alpha}_n}$ 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 $F_{s,n}^{\dagger} :=$ LCM$[\check{F}_{s,n}^{\check{\alpha}_n}]$. Note that $F_{s,n}^{\dagger}$ is a valid CDF. We can now estimate f_s by $f_{s,n}^{\dagger}$, where $f_{s,n}^{\dagger}$ is the piecewise constant function that is obtained by taking the left derivative of $F_{s,n}^{\dagger}$. In the following result, which is proved in the on-line supplementary material, we show that both $F_{s,n}^{\dagger}$ and $f_{s,n}^{\dagger}$ 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 $\check{\alpha}_n \rightarrow^{P} \alpha_0$, then, as $n \rightarrow \infty$,

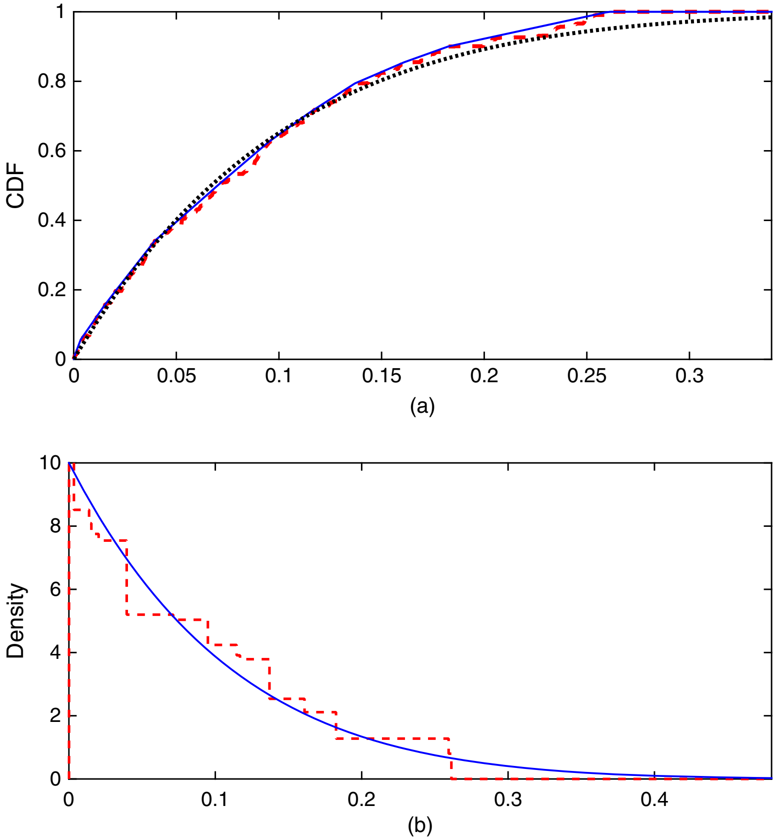


Fig. 2. (a) $\check{F}_{\mathrm{s},n}^{\tilde{\alpha}_0}$ (– – –), $F_{\mathrm{s},n}^{\dagger}$ (——) and F_{s} (· · · · · ·) for setting I and (b) $f_{\mathrm{s},n}^{\dagger}$ (– – –) and f_{s} (——) for setting II

$$\sup_{x\in\mathbb{R}} |F_{\mathrm{s},n}^{\dagger}(x) - F_{\mathrm{s}}(x)| \xrightarrow{\mathrm{P}} 0. \tag{12}$$

Further, if, for any $x > 0$, $f_{\mathrm{s}}(x)$ is continuous at x, then $f_{\mathrm{s},n}^{\dagger}(x) \rightarrow P f_{\mathrm{s}}(x)$.

Computing $F_{\mathrm{s},n}^{\dagger}$ and $f_{\mathrm{s},n}^{\dagger}$ is straightforward: an application of the pooled adjacent violators algorithm gives both the estimators; see, for example, chapter 1 of Robertson *et al.* (1988). Fig. 2(a) shows the LCM $F_{\mathrm{s},n}^{\dagger}$ whereas Fig. 2(b) shows its derivative $f_{\mathrm{s},n}^{\dagger}$ along with the true density f_{s} for the same data set as used in Fig. 1(b).

7. Multiple-testing problem

The problem of estimating the proportion of false null hypotheses α_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 $\mathbb{X}$. 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 $\mathcal{D}$ 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 FDR $\leqslant \beta(1-\alpha_0)$, where β is the user-defined level of the FDR-procedure. When α_0 is significantly bigger than 0 an estimate of α_0 can be used to yield a procedure with FDR approximately equal to β 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 α_0 and some proposed estimators.

Our method can be directly used to yield an estimator of α_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_{b} to be any known density, although in most multiple-testing applications we shall take f_{b} to be uniform$(0,1)$. The *local false discovery rate* (LFDR) is defined as the function $l: (0,1) \rightarrow [0,\infty)$, where

$$l(x) = P(H_i = 0 | X_i = x) = \frac{(1-\alpha_0)\, f_{\mathrm{b}}(x)}{f(x)},$$

and $f(x) = \alpha_0\, f_{\mathrm{s}}(x) + (1-\alpha_0)\, f_{\mathrm{b}}(x)$ is the density of the observed p-values. The estimation of the LFDR 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. $l(x) \leqslant 0.20$). Obtaining good estimates of 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:

$$\hat{l}(x) = \frac{(1-\check{\alpha}_n)\, f_{\mathrm{b}}(x)}{\check{\alpha}_n\, f^{\dagger}_{\mathrm{s},n}(x) + (1-\check{\alpha}_n)\, f_{\mathrm{b}}(x)}, \qquad \text{for } x \in (0,1).$$

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`.

8.1. Lower bounds for α_0

Although there has been some work on estimation of α_0 in the multiple-testing setting, Meinshausen and Rice (2006) and Genovese and Wasserman (2004) are the only references that we found that discuss methodology for constructing lower confidence bounds for α_0. These proce-

Table 1. Coverage probabilities of nominal 95% lower confidence bounds for the three methods when $n = 1000$ and $n = 5000$

α	*Results for n = 1000*						*Results for n = 5000*					
	Setting I			*Setting II*			*Setting I*			*Setting II*		
	$\hat{\alpha}_{\mathrm{L}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{GW}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{MR}}$	$\hat{\alpha}_{\mathrm{L}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{GW}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{MR}}$	$\hat{\alpha}_{\mathrm{L}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{GW}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{MR}}$	$\hat{\alpha}_{\mathrm{L}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{GW}}$	$\hat{\alpha}_{\mathrm{L}}^{\mathrm{MR}}$
0	0.95	0.98	0.93	0.95	0.98	0.93	0.95	0.97	0.93	0.95	0.97	0.93
0.01	0.97	0.98	0.99	0.97	0.97	0.99	0.98	0.98	0.99	0.98	0.98	0.99
0.03	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.98	0.99
0.05	0.98	0.98	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.98	0.98	0.99
0.10	0.99	0.99	1.00	0.99	0.98	0.99	0.99	0.99	1.00	0.99	0.98	0.99

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 $\sup_{t\in(0,1)}\{\mathbb{F}_n(t) - t - \eta_{n,\beta}\,\delta(t)\}/(1-t)$, where $\eta_{n,\beta}$ is a *bounding sequence* for the *bounding function* $\delta(t)$ at level β; see Meinshausen and Rice (2006). Genovese and Wasserman (2004) used a constant bounding function, $\delta(t) = 1$, with $\eta_{n,\beta} = \surd\{\log(2/\beta)/(2n)\}$, whereas Meinshausen and Rice (2006) suggested a class of bounding functions but observed that the *standard deviation proportional* bounding function $\delta(t) = \surd\{t(1-t)\}$ 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 $\hat{\alpha}_{\mathrm{L}}^{\mathrm{MR}}$, the bound in Genovese and Wasserman (2004) by $\hat{\alpha}_{\mathrm{L}}^{\mathrm{GW}}$ and the lower bound discussed in Section 4 by $\hat{\alpha}_{\mathrm{L}}$. 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 α_0

In this subsection, we illustrate and compare the performance of various estimators of α_0 under two sampling scenarios. In scenario A, we proceed as in Langaas *et al.* (2005). Let $\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{nj})$, for $j = 1, \ldots, J$, and assume that each $\mathbf{X}_j \sim N(\mu_{n\times 1}, \Sigma_{n\times n})$ and that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J$ are independent. We test $H_{0i} : \mu_i = 0$ *versus* $H_{1i} : \mu_i \neq 0$ for each $i = 1, 2, \ldots, n$. We set μ_i to 0 for the true null hypotheses, whereas for the false null hypotheses we draw μ_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 α be the proportion of false null hypotheses. Let $\bar{x}_i = \Sigma_{j=1}^{J} x_{ij}/J$ and $s_i^2 = \Sigma_{j=1}^{J}(x_{ij} - \bar{x}_i)^2/(J-1)$.

To test H_{0i} *versus* H_{1i}, we calculate a two-sided p-value based on a one-sample t-test, with $p_i = 2\,P\{T_{J-1} \geqslant |\bar{x}_i/\surd(s_i^2/J)|\}$, 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

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

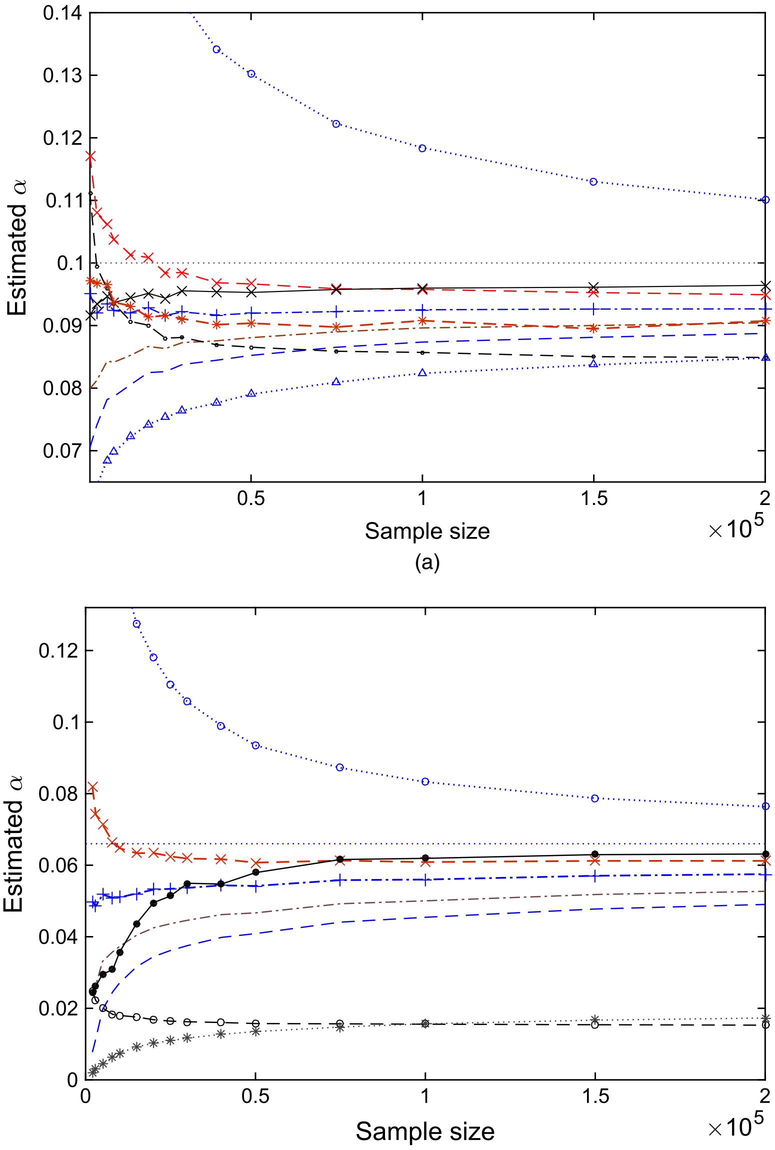
for $i = 1, 2, \ldots, n$. The dependence structure of the z_is is determined by L. For example, $L = 0$ corresponds to the case where the z_is are standard normal. Let $X_i = z_i + m_i$, for $i = 1, 2, \ldots, n$, where $m_i = 0$ under the null, and, under the alternative, $|m_i|$ is randomly generated from uniform(m^*, m^*+1) and 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 $\hat{\alpha}_0^{\mathrm{S,B}}$ to denote the estimator that was proposed by Storey (2002) when bootstrapping is used to choose the required tuning parameter, and denote by $\hat{\alpha}_0^{\mathrm{S},\lambda}$ the estimator when the value of the tuning parameter is fixed at λ. 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_{s}; we denote it by $\hat{\alpha}_0^{\mathrm{L}}$. We evaluate $\hat{\alpha}_0^{\mathrm{L}}$ 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: $\delta(t) = \surd\{t(1-t)\}$ and $\delta(t) = 1$. For its implementation, we must choose a sequence $\{\beta_n\}$ going to 0 as $n \to \infty$. Meinshausen and Rice (2006) did not specify any particular choice of $\{\beta_n\}$ but required the sequence to satisfy some conditions. We choose $\beta_n = 0.05/\surd n$ and denote the estimators by $\hat{\alpha}_0^{\mathrm{MR}}$ when $\delta(t) = \surd\{t(1-t)\}$ and by $\hat{\alpha}_0^{\mathrm{GW}}$ when $\delta(t) = 1$ (see Genovese and Wasserman (2004)). We also compare our results with $\hat{\alpha}_0^{\mathrm{E}}$, 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 $\hat{\alpha}_0^{\mathrm{J}}$ and in section 3.1 of Cai and Jin (2010) by $\hat{\alpha}_0^{\mathrm{CJ}}$. 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 $\hat{\alpha}_0^{c_n}$ depends on the choice of c_n and in what follows we investigate a proper choice of c_n. We take $\alpha_0 = 0.1$ and evaluate the performance of $\hat{\alpha}_0^{\tau \log\{\log(n)\}}$ for various values of τ, as n increases, for scenarios A and B. The choice $c_n = \tau \log\{\log(n)\}$, for various values of τ, is suggested after extensive simulations. We also include $\tilde{\alpha}_0$, $\hat{\alpha}_0^{\mathrm{GW}}$, $\hat{\alpha}_0^{\mathrm{MR}}$ and $\hat{\alpha}_0^{\mathrm{J}}$ 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 α_0 as n varies from 3000 to 10^5. Recall that $\tilde{\alpha}_0$ denotes the estimator that was proposed in Section 5. For both scenarios, the sample means of the estimators of α_0 that is proposed in this paper converge to the true α_0, as the sample size grows. The methods that are developed in this paper perform favourably in comparison with $\hat{\alpha}_0^{\mathrm{GW}}$, $\hat{\alpha}_0^{\mathrm{MR}}$ and $\hat{\alpha}_0^{\mathrm{J}}$. Since the choice of c_n dictates the finite sample performance of $\hat{\alpha}_0^{c_n}$, 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



(a)

(b)

Fig. 3. Means of various estimators of α_0 computed over 5000 independent replications as the sample size increases (· · ○ · ·, 0.01k_n; — × —, 0.05k_n; — + —, 0.1k_n; - - - -, 0.2k_n; — — —, 0.3k_n): (a) scenario A with $\Sigma = I_{n\times n}$ (— × —, $\tilde{\alpha}_0$; — ○ —, α_0^{GW}; · · △ · ·, $\hat{\alpha}_0^{\mathrm{MR}}$; — ✱ —, $\hat{\alpha}_0^{\mathrm{J}}$); (b) scenario B with $L = 0$ and $m^* = 1$ (— ● —, $\tilde{\alpha}_0$; — ○ —, α_0^{GW}; · · · ✱ · · ·, α_0^{MR})

the data into K sets, say $\mathcal{D}_1, \ldots, \mathcal{D}_K$. Let $\mathbb{F}_n^k$ be the empirical CDF of the data in $\mathcal{D}_k$. Let $\hat{\alpha}_{0,-k}^c$ be the estimator that is defined in expression (7) using all data except those in $\mathcal{D}_k$ and tuning parameter c. Further, let $\check{F}_{\mathrm{s},n}^{\hat{\alpha}_{0,-k}^c,-k}$ be the estimator of F_{s} as defined in lemma 1 using $\hat{\alpha}_{0,-k}^c$ and all data except those in $\mathcal{D}_k$. Define the cross-validated estimator of c as

$$c_{\mathrm{cv}} := \arg\min_{c\in\mathbb{R}} \sum_{k=1}^{K} \int (\mathbb{F}_n^k - \hat{F}^k)^2 \, \mathrm{d}\mathbb{F}_n^k, \qquad (13)$$

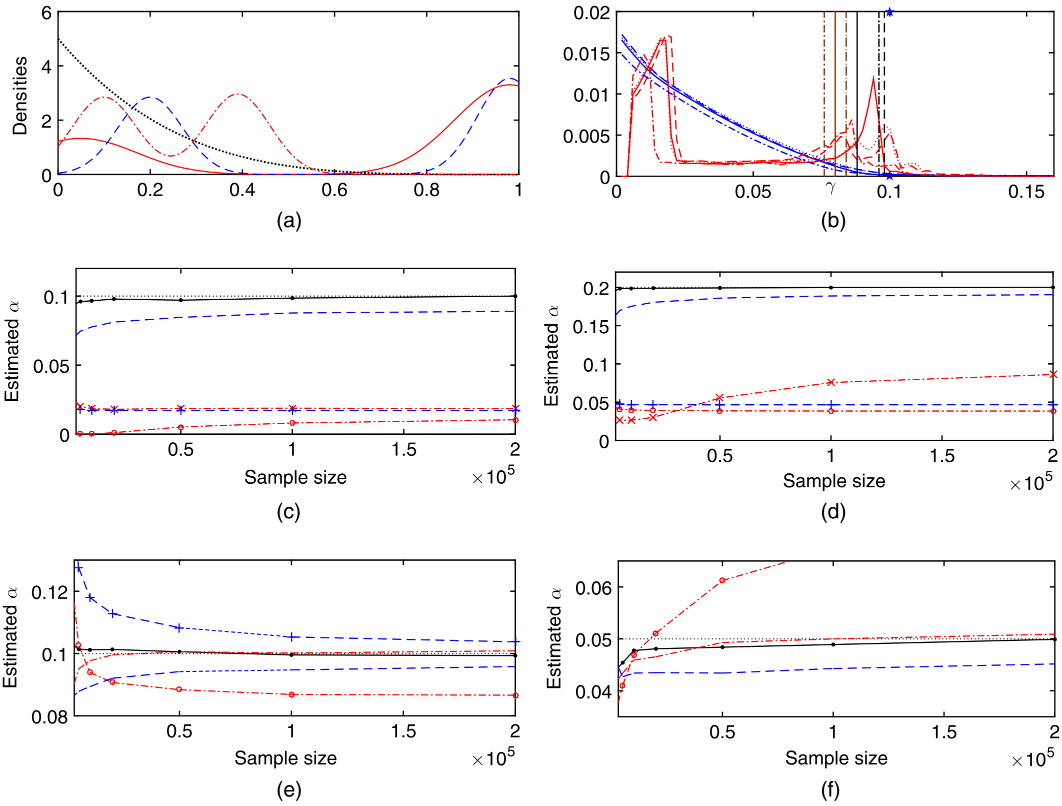


Fig. 4. (a) Density functions for various choices of F_s (——, distance 1; – – –, distance 2; ······, distance 3; – · – ·, distance 4), (b) $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (——, – – –, – · –, ······), the scaled second derivative (——, – – –, – · –, ······), $\hat{\alpha}_0^{\mathrm{CV}}$ (¦, ¦) and $\hat{\alpha}_0^{0.1k_n}$ (¦, |, ¦) for five independent samples of size 5000 corresponding to distance 1 (★, α_0), and means of various competing estimators of α_0 computed over 500 independent samples (– – –, $0.1k_n$) for (c) distance 1 (–×–, $\tilde{\alpha}_0$; –●–, $\hat{\alpha}_0^{\mathrm{CV}}$; –○–, $\hat{\alpha}_0^{\mathrm{MR}}$; –+–, $\hat{\alpha}_0^{\mathrm{S},0.2}$), (d) distance 2 (–×–, $\tilde{\alpha}_0$; –●–, $\hat{\alpha}_0^{\mathrm{CV}}$; –○–, $\hat{\alpha}_0^{\mathrm{GW}}$; –+–, $\hat{\alpha}_0^{\mathrm{S,B}}$), (e) distance 3 (- - - -, $\tilde{\alpha}_0$; –+–, $\hat{\alpha}_0^{\mathrm{CV}}$; –○–, $\hat{\alpha}_0^{\mathrm{GW}}$; –+–, $\hat{\alpha}_{0.5}^{\mathrm{S,B}}$) and (f) distance 4 (- - - -, $\tilde{\alpha}_0$; – + –, $\hat{\alpha}_0^{\mathrm{CV}}$; –○–, $\hat{\alpha}_0^{\mathrm{J}}$) as the sample size increases from 3000 to 2 $\times 10^5$

where $\hat{F}^k := \hat{\alpha}^c_{0,-k} \check{F}_s^{\hat{\alpha}^c_{0,-k},-k} + (1-\hat{\alpha}^c_{0,-k}) F_b$. In all simulations in this paper, we use $K=10$ and denote this estimator by $\hat{\alpha}_0^{\mathrm{CV}}$; 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 $\hat{\alpha}_0^{\mathrm{CV}}$ across different simulation settings; also see Sections 8.2.2 and 8.2.4, and section 13 (in the 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 $\hat{\alpha}_0^{\mathrm{CV}}$, $\tilde{\alpha}_0$, $\hat{\alpha}_0^{\mathrm{J}}$, $\hat{\alpha}_0^{\mathrm{L}}$ and $\hat{\alpha}_0^{0.1k_n}$ for $k_n = \log\{\log(n)\}$ are comparable. However,

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

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

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 $k_n = \log\{\log(n)\}$

$10\alpha_0$	$\hat{\alpha}_0^{0.1k_n}$	$\hat{\alpha}_0^{\mathrm{CV}}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{\mathrm{GW}}$	$\hat{\alpha}_0^{\mathrm{MR}}$	$\hat{\alpha}_0^{\mathrm{S,B}}$	$\hat{\alpha}_0^{\mathrm{J}}$	$\hat{\alpha}_0^{\mathrm{CJ}}$	$\hat{\alpha}_0^{\mathrm{L}}$	$\hat{\alpha}_0^{\mathrm{E}}$
0.07	0.03	0.04	0.08	0.00	0.00	0.04	0.11	0.19	0.03	0.06
	(0.44)	(0.67)	(0.28)	(0.66)	(0.66)	(0.65)	(0.96)	(2.96)	(0.38)	(0.77)
0.20	0.14	0.18	0.16	0.00	0.01	0.08	0.28	0.55	0.07	0.05
	(0.73)	(0.79)	(0.62)	(1.98)	(1.89)	(2.25)	(1.33)	(4.41)	(1.26)	(1.28)
0.33	0.25	0.31	0.28	0.02	0.04	0.12	0.48	0.92	0.12	0.05
	(0.89)	(0.85)	(0.95)	(3.15)	(2.91)	(3.83)	(1.77)	(6.48)	(2.14)	(1.90)
0.66	0.55	0.62	0.58	0.12	0.14	0.23	0.95	1.83	0.23	0.05
	(1.21)	(1.00)	(1.48)	(5.38)	(5.25)	(7.73)	(3.04)	(11.98)	(4.34)	(3.84)

the RMSEs of $\tilde{\alpha}_0$ and $\hat{\alpha}_0^{0.1k_n}$ are lower than those of $\hat{\alpha}_0^{\mathrm{CV}}$, $\hat{\alpha}_0^{\mathrm{J}}$ and $\hat{\alpha}_0^{\mathrm{L}}$. For scenario B, the sample means of $\tilde{\alpha}_0$, $\hat{\alpha}_0^{\mathrm{CV}}$ and $\hat{\alpha}_0^{0.1k_n}$ are comparable. In scenario B, the performances of $\hat{\alpha}_0^{\mathrm{J}}$ and $\hat{\alpha}_0^{\mathrm{CJ}}$ are not comparable with the estimators that are proposed in this paper, as $\hat{\alpha}_0^{\mathrm{J}}$ and $\hat{\alpha}_0^{\mathrm{CJ}}$ estimate α, whereas $\tilde{\alpha}_0$, $\hat{\alpha}_0^{\mathrm{CV}}$ and $\hat{\alpha}_0^{c_n}$ estimate α_0. Note that $\hat{\alpha}_0^{\mathrm{L}}$ fails to estimate α_0 because the underlying assumption that is inherent in their estimation procedure, that f_{s} be non-increasing, does not hold. In scenario A, $\hat{\alpha}_0^{\mathrm{S},0.5}$ has the best performance among the different values of λ, whereas, in scenario B, $\hat{\alpha}_0^{\mathrm{S},\lambda}$ has poor performance for all values of $\lambda \in [0, 1]$. Furthermore, $\hat{\alpha}_0^{\mathrm{GW}}$, $\hat{\alpha}_0^{\mathrm{MR}}$, $\hat{\alpha}_0^{\mathrm{CJ}}$, $\hat{\alpha}_0^{\mathrm{S,B}}$ and $\hat{\alpha}_0^{\mathrm{E}}$ perform poorly in both scenarios for all values of α_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 Σ 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 α_0. However, in scenario A, the performances of $\hat{\alpha}_0^{0.1k_n}$, $\tilde{\alpha}_0$ and α_0^{J} are comparable, for larger values of α_0. In scenario B, $\hat{\alpha}_0^{0.1k_n}$ performs well for $\alpha_0 = 0.033$ and $\alpha_0 = 0.067$. Observe that, as in the independence setting, $\hat{\alpha}_0^{\mathrm{GW}}$, $\hat{\alpha}_0^{\mathrm{MR}}$, $\hat{\alpha}_0^{\mathrm{S,B}}$, $\hat{\alpha}_0^{\mathrm{CJ}}$ and $\hat{\alpha}_0^{\mathrm{E}}$ perform poorly in both scenarios for all values of α_0.

Table 4. Means × 10 and RMSEs × 100 (in parentheses) of estimators discussed in Section 8.2 for scenario A with Σ as described in Section 8.2.3, $J = 10$, $n = 5000$ and $k_n = \log\{\log(n)\}$

$10\alpha_0$	$\hat{\alpha}_0^{0.1k_n}$	$\hat{\alpha}_0^{\mathrm{CV}}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{\mathrm{GW}}$	$\hat{\alpha}_0^{\mathrm{MR}}$	$\hat{\alpha}_0^{\mathrm{S},0.5}$	$\hat{\alpha}_0^{\mathrm{J}}$	$\hat{\alpha}_0^{\mathrm{CJ}}$	$\hat{\alpha}_0^{\mathrm{L}}$	$\hat{\alpha}_0^{\mathrm{E}}$
0.10	0.46	0.42	0.33	0.07	0.06	0.28	0.22	0.07	0.32	0.37
	(5.15)	(4.23)	(3.84)	(1.72)	(1.27)	(4.11)	(3.03)	(10.61)	(4.37)	(3.91)
0.30	0.52	0.53	0.41	0.14	0.17	0.65	0.34	0.15	0.49	0.39
	(3.80)	(3.64)	(3.59)	(2.72)	(1.90)	(6.58)	(3.25)	(10.35)	(4.30)	(4.31)
0.50	0.66	0.76	0.54	0.26	0.31	0.54	0.49	0.25	0.66	0.37
	(3.52)	(5.43)	(3.85)	(3.56)	(2.50)	(2.61)	(3.60)	(10.45)	(4.31)	(4.03)
1.00	1.06	1.13	0.97	0.68	0.69	1.15	0.97	0.53	1.11	0.36
	(3.09)	(3.92)	(4.00)	(4.15)	(3.54)	(6.01)	(3.61)	(10.55)	(4.13)	(3.99)

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

$10\alpha_0$	$\hat{\alpha}_0^{0.1k_n}$	$\hat{\alpha}_0^{\mathrm{CV}}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{\mathrm{GW}}$	$\hat{\alpha}_0^{\mathrm{MR}}$	$\hat{\alpha}_0^{\mathrm{S,B}}$	$\hat{\alpha}_0^{\mathrm{J}}$	$\hat{\alpha}_0^{\mathrm{CJ}}$	$\hat{\alpha}_0^{\mathrm{L}}$	$\hat{\alpha}_0^{\mathrm{E}}$
0.07	0.29	0.38	0.17	0.04	0.05	0.26	0.20	0.21	0.13	0.22
	(2.92)	(3.70)	(1.62)	(1.02)	(1.36)	(3.71)	(2.80)	(9.87)	(1.75)	(2.22)
0.20	0.30	0.42	0.18	0.04	0.04	0.16	0.33	0.55	0.13	0.19
	(1.84)	(2.88)	(1.25)	(1.75)	(1.71)	(2.24)	(3.25)	(10.35)	(1.42)	(2.27)
0.33	0.38	0.52	0.20	0.06	0.06	0.17	0.50	0.93	0.16	0.18
	(1.54)	(2.74)	(1.89)	(2.83)	(2.73)	(3.51)	(3.71)	(11.52)	(2.03)	(2.59)
0.67	0.63	0.77	0.31	0.14	0.15	0.24	0.95	1.82	0.25	0.16
	(1.53)	(2.25)	(4.32)	(5.26)	(5.13)	(7.60)	(4.54)	(15.13)	(4.23)	(4.08)

8.2.4. *Comparing the performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{\mathrm{CV}}$ 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 α_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 $\hat{\alpha}_0^{\mathrm{CV}}$, $\tilde{\alpha}_0$ 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 $\gamma\, d_n(\hat{F}^{\gamma}_{\mathrm{s},n}, \check{F}^{\gamma}_{\mathrm{s},n})$, as a function of γ, when F_{s} has a multimodal density (see Figs 4(b) and 4(c)). Observe that $\hat{\alpha}_0^{\mathrm{CV}}$ 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 α_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 α_0. For $\hat{\alpha}_0^{c_n}$, 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 α_0 for the two data sets

Data set	$\hat{\alpha}_0^{0.1k_n}$	$\hat{\alpha}_0^{\mathrm{CV}}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{\mathrm{GW}}$	$\hat{\alpha}_0^{\mathrm{MR}}$	$\hat{\alpha}_0^{\mathrm{S,B}}$	$\hat{\alpha}_0^{\mathrm{J}}$	$\hat{\alpha}_0^{\mathrm{CJ}}$	$\hat{\alpha}_0^{\mathrm{L}}$	$\hat{\alpha}_0^{\mathrm{E}}$
Prostate	0.08	0.10	0.09	0.04	0.01	0.19	0.10	0.02	0.11	0.02
Carina	0.36	0.35	0.36	0.31	0.30	0.45	0.61	1.00	0.38	—†

†Not applicable.

of the plot of $\gamma\, d_n(\hat{F}^{\gamma}_{\mathrm{s},n}, \check{F}^{\gamma}_{\mathrm{s},n})$ can be useful in checking the validity of $\tilde{\alpha}_0$ as an estimator of the elbow, and thus for α_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 × 102 matrix $\mathbb{X}$ 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 $\bar{x}_i(1)$ and $\bar{x}_i(2)$ 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 $t_i=\{\bar{x}_i(1)-\bar{x}_i(2)\}/s_i$, where s_i is an estimate of the standard error of $\bar{x}_i(1)-\bar{x}_i(2)$, i.e.

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

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 $\hat{\alpha}_{\mathrm{L}}$ for these data is found to be 0.05. In Table 6, we display estimates of α_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

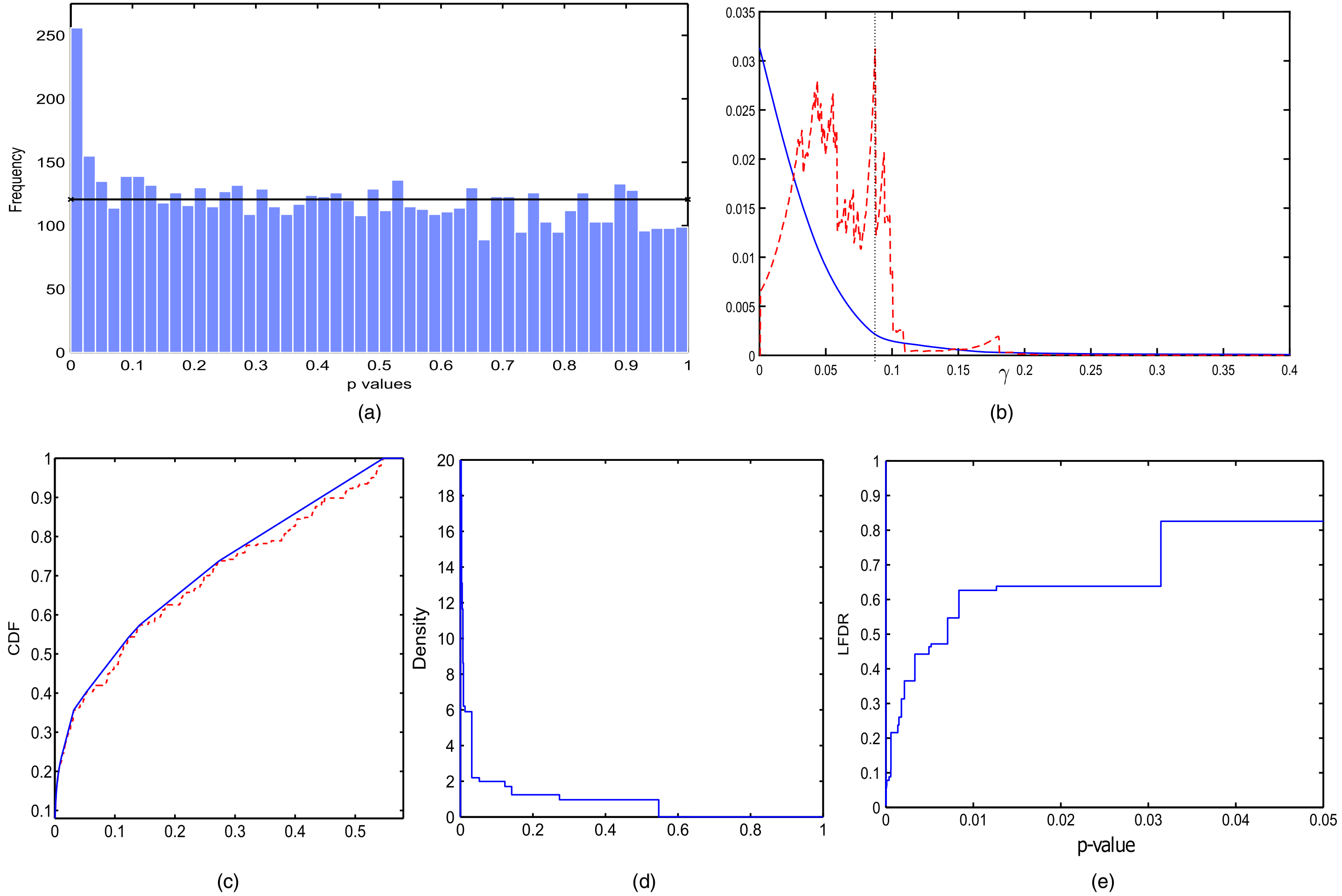
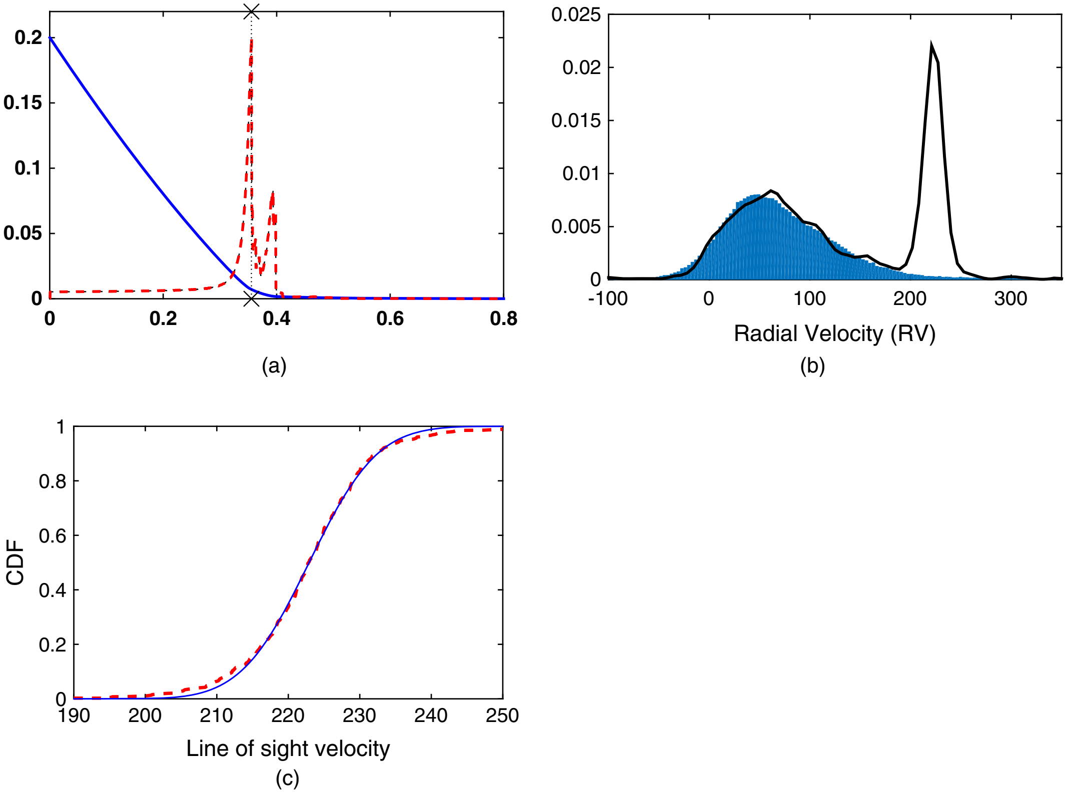


Fig. 5. Plots for the prostate data: (a) histogram of the p-values (——, uniform(0, 1) distribution); (b) $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (——) overlaid with its scaled derivative (– – –) (⋮, point of maximum curvature $\tilde{\alpha}_0 = 0.088$); (c) $\check{F}^{\tilde{\alpha}_0}_{s,n}$ (- - - -) and $F^{\dagger}_{s,n}$ (——); (d) $f^{\dagger}_{s,n}$; (e) estimated LFDR $\hat{l}$ for p-values less than 0.05



(a)

(b)

(c)

Fig. 6. Plots for the RV-data in the Carina dwarf spheroidal galaxy: (a) $\gamma\, d_n(\hat{F}^{\gamma}_{\mathrm{s},n}, \check{F}^{\gamma}_{\mathrm{s},n})$ (——) overlaid with its (scaled) second derivative (- - -); (b) density of the RV-distribution of the contaminating stars overlaid with the (scaled) kernel density estimator of the observed sample; (c) $\check{F}^{\tilde{\alpha}_0}_{\mathrm{s},n}$ (- - -) overlaid with its closest Gaussian distribution (——)

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 α_0 is found to be 0.323. Fig. 6(c) shows the estimate of F_{s} and the closest (in terms of minimizing the $L_2(\check{F}^{\tilde{\alpha}_0}_{\mathrm{s},n})$ 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), $\hat{\alpha}_0^{\mathrm{CJ}}$, is greater than 1, whereas Efron's method (see Efron (2007)), implemented by using the `locfdr` package in R, fails to estimate α_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 α_0, under minimal assumptions on the model. We propose an honest finite sample lower confidence bound of α_0 that is distribution free. Two point estimators of α_0, $\hat{\alpha}_0^{c_n}$ and $\tilde{\alpha}_0$, are studied. We prove that $\hat{\alpha}_0^{c_n}$ is a consistent estimator of α_0 and show that the rate of convergence of $\hat{\alpha}_0^{c_n}$ can be arbitrarily close to $\sqrt{n}$, for proper choices of c_n. These proposed estimators crucially rely on $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$, as a function of γ, whose plot provides useful insights about the nature of the problem and performance of the estimators.

We observe that the estimators of α_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_s which is uniformly consistent. Furthermore, if F_s is known to have a non-increasing density f_s we can find a consistent estimator of f_s. 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 α_0 remains a difficult problem as the asymptotic distribution of $\hat{\alpha}_0^{c_n}$ depends on the unknown F. We are currently developing estimators of α_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 $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$, 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_s 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 α_0, we have

$$\begin{aligned}
\alpha_0 &= \inf\{0 \leqslant \gamma \leqslant \alpha : \{F - (1-\gamma)F_b\}/\gamma \text{ is a valid CDF}\} \\
&= \inf\{0 \leqslant \gamma \leqslant \alpha : \{\alpha F_s + (1-\alpha)F_b - (1-\gamma)F_b\}/\gamma \text{ is a valid CDF}\} \\
&= \inf\{0 \leqslant \gamma \leqslant \alpha : \{\alpha F_s - (\alpha-\gamma)F_b\}/\gamma \text{ is a valid CDF}\} \\
&= \alpha - \sup\{0 \leqslant \epsilon \leqslant \alpha : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} \\
&= \alpha - \sup\{0 \leqslant \epsilon \leqslant 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\},
\end{aligned}$$

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

$$\begin{aligned}
& \sup\{0 \leqslant \epsilon \leqslant 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} = \alpha \\
\Leftrightarrow\ & \inf\{0 \leqslant \delta \leqslant 1 : \alpha(F_s - F_b) + \delta F_b \text{ is a sub-CDF}\} = 0.
\end{aligned}$$

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 $F_\mathrm{s}^\gamma = \{F - (1-\gamma)F_\mathrm{b}\}/\gamma$, observe that

$$\gamma\, d_n(\hat{F}_{\mathrm{s},n}^\gamma, F_\mathrm{s}^\gamma) = d_n(F, \mathbb{F}_n).$$

Also note that F_s^γ is a valid CDF for $\gamma \geqslant \alpha_0$. As $\check{F}_{\mathrm{s},n}^\gamma$ is defined as the function that minimizes the $L_2(\mathbb{F}_n)$ distance of $\hat{F}_{\mathrm{s},n}^\gamma$ over all CDFs,

$$\gamma\, d_n(\check{F}_{\mathrm{s},n}^\gamma, \hat{F}_{\mathrm{s},n}^\gamma) \leqslant \gamma\, d_n(\hat{F}_{\mathrm{s},n}^\gamma, F_\mathrm{s}^\gamma) = d_n(F, \mathbb{F}_n).$$

To prove the second part of lemma 6 note that for $\gamma \geqslant \alpha_0$ the result follows from above and the fact that $d_n(F, \mathbb{F}_n) \to 0$ almost surely as $n \to \infty$.

For $\gamma < \alpha_0$, F_s^γ is not a valid CDF, by the definition of α_0. Note that as $n \to \infty$, $\hat{F}_{\mathrm{s},n}^\gamma \to F_\mathrm{s}^\gamma$ almost surely, pointwise. So, for sufficiently large n, $\hat{F}_{\mathrm{s},n}^\gamma$ is not a valid CDF, whereas $\check{F}_{\mathrm{s},n}^\gamma$ is always a CDF. Thus, $d_n(\hat{F}_{\mathrm{s},n}^\gamma, \check{F}_{\mathrm{s},n}^\gamma)$ converges to something positive.

A.3. Proof of lemma 7

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

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

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

$$\begin{aligned} d_n(\hat{F}_{\mathrm{s},n}^{\gamma_3}, \check{F}_{\mathrm{s},n}^{\gamma_3}) &\leqslant d_n[\hat{F}_{\mathrm{s},n}^{\gamma_3}, \{\eta\gamma_1 \check{F}_{\mathrm{s},n}^{\gamma_1} + (1-\eta)\gamma_2 \check{F}_{\mathrm{s},n}^{\gamma_2}\}/\gamma_3] \\ &= d_n\left(\frac{\eta\gamma_1 \hat{F}_{\mathrm{s},n}^{\gamma_1} + (1-\eta)\gamma_2 \hat{F}_{\mathrm{s},n}^{\gamma_2}}{\gamma_3}, \frac{\eta\gamma_1 \check{F}_{\mathrm{s},n}^{\gamma_1} + (1-\eta)\gamma_2 \check{F}_{\mathrm{s},n}^{\gamma_2}}{\gamma_3} \right) \\ &\leqslant \frac{\eta\gamma_1}{\gamma_3} d_n(\hat{F}_{\mathrm{s},n}^{\gamma_1}, \check{F}_{\mathrm{s},n}^{\gamma_1}) + \frac{(1-\eta)\gamma_2}{\gamma_3} d_n(\hat{F}_{\mathrm{s},n}^{\gamma_2}, \check{F}_{\mathrm{s},n}^{\gamma_2}) \end{aligned} \tag{14}$$

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

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

Thus $\gamma_3 \in A_n$.

A.4. Proof of lemma 8

As $\alpha_0 = 0$,

$$P(\hat{\alpha}_0^{c_n} = 0) = 1 - P(\hat{\alpha}_0^{c_n} > 0) = 1 - P\{\sqrt{n}\, d_n(\mathbb{F}_n, F) > c_n\} \to 1, \tag{15}$$

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

A.5. Proof of theorem 5

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

$$\begin{aligned} P(\alpha_0 \geqslant \hat{\alpha}_\mathrm{L}) &= P\{\sqrt{n}\alpha_0\, d_n(\hat{F}_{\mathrm{s},n}^{\alpha_0}, \check{F}_{\mathrm{s},n}^{\alpha_0}) \leqslant c_n\} \\ &\geqslant P\{\sqrt{n}\alpha_0\, d_n(\hat{F}_{\mathrm{s},n}^{\alpha_0}, F_\mathrm{s}^{\alpha_0}) \leqslant c_n\} = H_n(c_n) = 1 - \beta, \end{aligned}$$

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

$$P(\alpha_0 \geqslant \hat{\alpha}_\mathrm{L}) = P\{\sqrt{n}\, d_n(\mathbb{F}_n, F_\mathrm{b}) \leqslant c_n\} = P\{\sqrt{n}\, d_n(\mathbb{F}_n, F) \leqslant c_n\} = 1 - \beta.$$

A.6. Proof of lemma 9

Let $0<\gamma_1<\gamma_2<1$. Then,

$$\begin{aligned}\gamma_2\, d_n(\hat{F}^{\gamma_2}_{s,n}, \check{F}^{\gamma_2}_{s,n}) &\leqslant \gamma_2\, d_n\{\hat{F}^{\gamma_2}_{s,n}, (\gamma_1/\gamma_2)\check{F}^{\gamma_1}_{s,n} + (1-\gamma_1/\gamma_2)F_b\}\\ &= d_n\{\gamma_1 \hat{F}^{\gamma_1}_{s,n} + (\gamma_2-\gamma_1)F_b, \gamma_1 \check{F}^{\gamma_1}_{s,n} + (\gamma_2-\gamma_1)F_b\}\\ &\leqslant \gamma_1\, d_n(\hat{F}^{\gamma_1}_{s,n}, \check{F}^{\gamma_1}_{s,n}),\end{aligned}$$

which shows that $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is a non-increasing function. To show that $\gamma\, d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is convex, let $0<\gamma_1<\gamma_2<1$ and $\gamma_3=\eta\gamma_1+(1-\eta)\gamma_2$, for $0\leqslant\eta\leqslant 1$. Then, by inequality (14) we have the desired result.

A.7. Proof of theorem 9

Let $\epsilon_n := \sup_{x\in\mathbb{R}} |\check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x)|$. Then the function $F_s+\epsilon_n$ is concave on $[0,\infty)$ and majorizes $\check{F}^{\check{\alpha}_n}_{s,n}$. Hence, for all $x\in[0,\infty)$, $\check{F}^{\check{\alpha}_n}_{s,n}(x) \leqslant F^{\dagger}_{s,n}(x) \leqslant F_s(x)+\epsilon_n$, as $F^{\dagger}_{s,n}$ is the LCM of $\check{F}^{\check{\alpha}_n}_{s,n}$. Thus,

$$-\epsilon_n \leqslant \check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x) \leqslant F^{\dagger}_{s,n}(x) - F_s(x) \leqslant \epsilon_n,$$

and, therefore,

$$\sup_{x\in\mathbb{R}} |F^{\dagger}_{s,n}(x) - F_s(x)| \leqslant \epsilon_n.$$

By theorem 7, as $\epsilon_n \rightarrow^{P} 0$, 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

Anderson, T. W. and Darling, D. A. (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, **23**, 193–212.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972) *Statistical Inference under Order Restrictions: the Theory and Application of Isotonic Regression*. London: Wiley.

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

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

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

Bertsekas, D. P. (2003) *Convex Analysis and Optimization*. Belmont: Athena Scientific.

Black, M. A. (2004) A note on the adaptive control of false discovery rates. *J. R. Statist. Soc.* B, **66**, 297–304.

Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006) Semiparametric estimation of a two-component mixture model. *Ann. Statist.*, **34**, 1204–1232.

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

Celisse, A. and Robin, S. (2010) A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Planng Inf.*, **140**, 3132–3147.

Cohen, A. C. (1967) Estimation in mixtures of two normal distributions. *Technometrics*, **9**, 15–28.

Day, N. E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

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

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

Efron, B. (2010) Empirical Bayes methods for estimation, testing, and prediction. In *Large-scale Inference*. Cambridge: Cambridge University Press.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, vol. II, 2nd edn. New York: Wiley.

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

Grenander, U. (1956) On the theory of mortality measurement: I. *Skand. Akt.*, **39**, 70–96.

Grotzinger, S. J. and Witzgall, C. (1984) Projections onto order simplexes. *Appl. Math. Optimizn*, **12**, 247–270.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J. and Tibshirani, R. (2009) *The Elements of Statistical Learning*, vol. 2. New York: Springer.

Hengartner, N. W. and Stark, P. B. (1995) Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.*, **23**, 525–550.

Hunter, D. R., Wang, S. and Hettmansperger, T. P. (2007) Inference for mixtures of symmetric distributions. *Ann. Statist.*, **35**, 224–251.

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

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

Lindsay, B. G. (1983) The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, **11**, 86–94.

Lindsay, B. G. (1995) Mixture models: theory, geometry and applications. *Regl Conf. Ser. Probab. Statist.*, **5**, 1–163.

Lindsay, B. G. and Basak, P. (1993) Multivariate normal mixtures: a fast consistent method of moments. *J. Am. Statist. Ass.*, **88**, 468–476.

Lyons, L. (2008) Open statistical issues in particle physics. *Ann. Appl. Statist.*, **2**, 887–915.

McLachlan, G. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley-Interscience.

Meinshausen, N. and Bühlmann, P. (2005) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, **92**, 893–907.

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

Miller, C. J., Genovese, C., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, A. and Moore, J. (2001) Controlling the false-discovery rate in astrophysical data analysis. *Astron. J.*, **122**, 3492–3505.

Nguyen, V. H. and Matias, C. (2014) On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scand. J. Statist.*, **41**, 1167–1194.

Parzen, E. (1960) *Modern Probability Theory and Its Applications*. New York: Wiley.

Quandt, R. E. and Ramsey, J. B. (1978) Estimating mixtures of normal distributions and switching regressions (with comments). *J. Am. Statist. Ass.*, **73**, 730–752.

R Development Core Team (2008) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988) *Order Restricted Statistical Inference*. Chichester: Wiley.

Robin, S., Bar-Hen, A., Daudin, J.-J. and Pierre, L. (2007) A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Computnl Statist. Data Anal.*, **51**, 5483–5493.

Robin, A. C., Reyl, C., Derrire, S. and Picaud, S. (2003) A synthetic view on structure and evolution of the Milky Way. *Astron. Astrophys.*, **409**, 523–540.

Salvador, S. and Chan, P. (2004) Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence*: *Proc. 16th IEEE Int. Conf. Tools with Artificial Intelligence*, pp. 576–584. Piscataway: Institute of Electrical and Electronics Engineers.

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

Swanepoel, J. W. H. (1999) The limiting behavior of a modified maximal symmetric 2s-spacing with applications. *Ann. Statist.*, **27**, 24–35.

Turkheimer, F., Smith, C. and Schmidt, K. (2001) Estimation of the number of "true" null hypotheses in multivariate analysis of neuroimaging data. *Neuroimage*, **13**, 920–930.

Walker, M. G., Mateo, M., Olszewski, E. W., Gnedin, O. Y., Wang, X., Sen, B. and Woodroofe, M. (2007) Velocity dispersion profiles of seven dwarf spheroidal galaxies. *Astrophys. J.*, **667**, L53–L56.

Walker, M., Mateo, M., Olszewski, E., Sen, B. and Woodroofe, M. (2009) Clean kinematic samples in dwarf spheroidals: an algorithm for evaluating membership and estimating distribution parameters when contamination is present. *Astron. J.*, **137**, no. 2 article 3109.

Walther, G. (2001) Multiscale maximum likelihood analysis of a semiparametric model, with applications. *Ann. Statist.*, **29**, 1297–1319.

Walther, G. (2002) Detecting the presence of mixing with multiscale maximum likelihood. *J. Am. Statist. Ass.*, **97**, 508–513.

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'.