Estimation of a Two-component Mixture Model with Applications to Multiple Testing

Rohit Kumar Patra and Bodhisattva Sen

Columbia University, New York

Abstract

We consider a two-component mixture model with one known component. We develop methods for estimating the mixing proportion and the unknown distribution nonparametrically, given i.i.d. 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 proposed estimators, 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.,

$$F(x) = \alpha F_s(x) + (1 - \alpha)F_b(x), \tag{1}$$

where the cumulative distribution function (CDF) F_b is known, but the mixing proportion $\alpha \in [0, 1]$ and the CDF F_s ($\neq F_b$) are unknown. Given a random sample from F, we wish to (nonparametrically) estimate 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 H_0, while their distribution associated with H_1 is unknown; see e.g., Efron (2010) and Robin et al. (2007). Translated to the setting of (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 (1); see e.g., McLachlan and Peel (2000). For example, in astronomy, such situations arise quite often: when observing some variable(s) of interest (e.g., metallicity, 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 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 a rather smooth background distribution; see e.g., Lyons (2008).

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

In this paper we provide a methodology to estimate α and F_s (nonparametrically), without assuming any constraint on the form of F_s. The main contributions of our paper can be summarised in the following.

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

- Two different estimators of α are proposed and studied. We derive the rate of convergence and asymptotic limit for one of the proposed estimators.

- A nonparametric estimator of F_s 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 organised as follows. In Section 2 we address the identifiability of the model given in (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 proposed in Section 3, with other methods available in the literature through simulation studies, and 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 the some of the main results in the paper. The proofs of the results not given in Appendix A can be found in Appendix E.

§2 The model and identifiability

§2.1 When α is known

Suppose that we observe an i.i.d. sample $X_1, X_2, \ldots, X_n$ from F as in (1). If $\alpha \in (0, 1]$ were known, a naive estimator of F_s would be

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

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

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

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

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

$\mathbf{V} = (V_1, V_2, \ldots, V_n)$, $V_i := \hat{F}_{s,n}^{\alpha}(X_{(i)})$, $i = 1, 2, \ldots, n$, $X_{(i)}$ being the i-th order statistic of the sample, and $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^n$. The estimator $\hat{\boldsymbol{\theta}}$ is uniquely defined by the projection theorem (see e.g., Proposition 2.2.1 on page 88 of Bertsekas (2003)); it is the Euclidean projection of $\mathbf{V}$ on the closed convex set $\Theta_{inc} \subset \mathbb{R}^n$. $\hat{\boldsymbol{\theta}}$ is related to $\check{F}_{s,n}^{\alpha}$ via $\check{F}_{s,n}^{\alpha}(X_{(i)}) = \hat{\theta}_i$, and can be easily computed using the pool-adjacent-violators algorithm (PAVA); see Section 1.2 of Robertson et al. (1988). Thus, $\check{F}_{s,n}^{\alpha}$ 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 piece-wise constant right continuous function with possible jumps only at the data points. The following result, derived easily from Chapter 1 of Robertson et al. (1988), characterises $\check{F}_{s,n}^{\alpha}$.

Lemma 2.1. *Let $\tilde{F}_{s,n}^{\alpha}$ be the isotonic regression (see e.g., page 4 of Robertson et al. (1988)) of the set of points $\{\hat{F}_{s,n}^{\alpha}(X_{(i)})\}_{i=1}^{n}$. Then $\tilde{F}_{s,n}^{\alpha}$ is characterised as the right-hand slope of the greatest convex minorant of the set of points $\{i/n, \sum_{j=0}^{i} \hat{F}_{s,n}^{\alpha}(X_{(j)})\}_{i=0}^{n}$. The restriction of $\tilde{F}_{s,n}^{\alpha}$ to $[0,1]$, i.e., $\check{F}_{s,n}^{\alpha} = \min\{\max\{\tilde{F}_{s,n}^{\alpha}, 0\}, 1\}$, minimises (3) over all CDFs.*

Isotonic regression and the PAVA are very well studied in the statistical literature with many text-book length treatments; see e.g., Robertson et al. (1988) and Barlow et al. (1972). If skillfully implemented, PAVA 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 harder; in fact, it is non-identifiable. If (1) holds for some F_b and α then the mixture model can be re-written as

$$F = (\alpha + \gamma)\left(\frac{\alpha}{\alpha + \gamma} F_s + \frac{\gamma}{\alpha + \gamma} F_b\right) + (1 - \alpha - \gamma) F_b,$$

for $0 \leq \gamma \leq 1 - \alpha$, and the term $(\alpha F_s + \gamma F_b)/(\alpha + \gamma)$ can be thought of as the nonparametric component. A trivial solution occurs when we take $\alpha + \gamma = 1$, in which case (3) is minimised when $W = \mathbb{F}_n$. Hence, α is not uniquely defined. To handle the identifiability issue, we redefine the mixing proportion as

$$\alpha_0 := \inf\left\{\gamma \in (0, 1] : [F - (1 - \gamma) F_b]/\gamma \text{ is a CDF}\right\}. \tag{4}$$

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 (4). Identifiability of mixture models has been discussed in many papers, but generally with parametric assumptions on the model. Genovese and Wasserman (2004) discuss identifiability when F_b is the uniform distribution and F has a density. Hunter et al. (2007) and Bordes et al. (2006) discuss identifiability for location shift mixtures of symmetric distributions. Most authors try to find conditions for the identifiability of their model, while 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 (1) is non-identifiable.

Suppose that we start with a fixed F_s, F_b and α satisfying (1). As seen from the above discussion we can only hope to estimate α_0, which, from its definition in (4), is smaller than α, 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, proved in the Appendix A.1, gives the connection between α and α_0.

Lemma 2.2. *Let F be as in (1) and α_0 as defined in (4). Then*

$$\alpha_0 = \alpha - \sup \{0 \leq \epsilon \leq 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\}, \tag{5}$$

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 $\alpha F_s - \epsilon F_b$ is a sub-CDF. Furthermore, $\alpha_0 = 0$ if and only if $F = F_b$.

In the following 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 Appendix E. 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 $\eta_3 = 0$, since singular functions almost never occur in practice; see e.g., Parzen (1960). Hence, we may assume

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

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 $J_G : d(G) \to [0, 1]$ to be a function 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 2.3. *Let F_s and F_b be discrete CDFs. If $d(F_b) \not\subset d(F_s)$, then $\alpha_0 = \alpha$, i.e., (1) is identifiable. If $d(F_b) \subset d(F_s)$, then $\alpha_0 = \alpha \left\{1 - \inf_{x \in d(F_b)} J_{F_s}(x)/J_{F_b}(x)\right\}$. 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 2.4. *Suppose that F_s and F_b are absolutely continuous, i.e., they have densities f_s and f_b, respectively. Then*

$$\alpha_0 = \alpha \left\{ 1 - \operatorname{ess\,inf} \frac{f_s}{f_b} \right\},$$

where, for any function g, $\operatorname{ess\,inf} g = \sup\{a \in \mathbb{R} : \mathfrak{m}(\{x : g(x) < a\}) = 0\}$, $\mathfrak{m}$ 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 w.r.t. $\mathfrak{m}$.

The above lemma 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 α.

In Appendix B we apply the above two lemmas 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, proved in greater generality in Appendix B, we give conditions under which a general CDF F, that can be represented as in (6), is identifiable.

Lemma 2.5. *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 (1) is identifiable, if either $F^{(a)}$ or $F^{(d)}$ 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 (5). For the rest of the paper, unless otherwise noted, we assume

$$X_1, X_2, \ldots, X_n \text{ is an i.i.d. sample from } F \text{ as in (1)}.$$

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

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

where c_n is a sequence of constants and d_n stands for the $L_2(\mathbb{F}_n)$ distance, i.e., if $g, h : \mathbb{R} \to \mathbb{R}$ 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

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

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

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

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 α_0. We will 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 $\Psi^{-1}(y) := \inf\{t \in \mathbb{R} : y \leq \Psi(t)\}$, and $Y_i := \Psi^{-1}(X_i)$. It is easy to see that $Y_1, Y_2, \ldots, Y_n$ is an i.i.d. 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 α. We can define α_0^Y as in (4) but with $\{G, F_b \circ \Psi\}$ instead of $\{F, F_b\}$. The following result proved in Appendix E.3 shows that α_0 and its estimators proposed in this paper are invariant under such monotonic transformations.

Theorem 3.1. *Let $\mathbb{G}_n$ be the empirical CDF of $Y_1, Y_2, \ldots, Y_n$. Also, let $\hat{G}_{s,n}$ and $\check{G}^{\gamma}_{s,n}$ be as defined in (2) and (3), respectively, but with $\{\mathbb{G}_n, F_b \circ \Psi\}$ instead of $\{\mathbb{F}_n, F_b\}$. Then $\alpha_0 = \alpha_0^Y$ and $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) = \gamma d_n(\hat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n})$ for all $\gamma \in (0, 1]$.*

§3.2 Consistency of $\hat{\alpha}_0^{c_n}$

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

Lemma 3.2. *For $1 \geq \gamma \geq \alpha_0$, $\gamma d_n(\check{F}^{\gamma}_{s,n}, \hat{F}^{\gamma}_{s,n}) \leq d_n(F, \mathbb{F}_n)$. Thus,*

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

Lemma 3.3. *The set $A_n := \{\gamma \in [0, 1] : \sqrt{n}\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \leq c_n\}$ is convex. Thus, $A_n = [\hat{\alpha}_0^{c_n}, 1]$.*

The following result, proved in Appendix E.4, shows that for a broad range of choices of c_n, our estimation procedure is consistent.

Theorem 3.4. *If $c_n = o(\sqrt{n})$ and $c_n \to \infty$, then $\hat{\alpha}_0^{c_n} \stackrel{P}{\to} \alpha_0$.*

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 Appendix C. 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 Appendix C 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 3.5. *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 3.6. *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 the above result is involved and we give the details in Appendix E.5.

Remark 3.7. *Genovese and Wasserman (2004) show that the estimators of α_0 proposed by Hengartner and Stark (1995) and Swanepoel (1999) have convergence rates of $(n/\log n)^{1/3}$ and $n^{2/5}/(\log n)^{\delta}$, for $\delta > 0$, respectively. Morover, both results require smoothness assumptions on F – Hengartner and Stark (1995) require F to be concave with a density that is Lipschitz of order 1, while Swanepoel (1999) requires even stronger smoothness conditions on the density. Nguyen and Matias (2013) prove that when the density of $F_s^{\alpha_0}$ vanishes at a set of points of measure zero 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 infinity very slowly. If we take $c_n = \log\log n$, we get an estimator that has a rate of convergence $\sqrt{n}/\log\log n$. In fact, as the next result (proved in Appendix E.6) 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 3.8. *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*

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

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 $\hat{\alpha}_0^{c_n}$ discussed in Section 3 depends on unknown parameters (e.g., α_0, F) in a complicated fashion and is of little practical use. Our goal in this sub-section is to construct a finite sample (honest) lower confidence bound $\hat{\alpha}_L$ with the property

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

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 one 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 papers 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) and Walther (2002) discuss the homogeneity problem under the assumption that the densities are log-concave. Donoho and Jin (2004) and Cai and Jin (2010) discuss the problem of detecting sparse heterogeneous mixtures under parametric settings using the 'higher criticism' statistic; see Appendix D.

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 4.1. *Let H_n be the CDF of $\sqrt{n}d_n(\mathbb{F}_n, F)$. Let $\hat{\alpha}_L$ be defined as in (7) with $c_n = H_n^{-1}(1-\beta)$. Then (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 the above theorem 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 uniforms. 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

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

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

Theorem 4.2. $\sup_{x\in\mathbb{R}} |H_n(x) - G_n(x)| \to 0$ *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 (e.g., 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

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

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 $F_s^{\alpha_0}$ and F.

Theorem 4.3. *Assume that $\alpha_0 > 0$. Then $\sqrt{n}\alpha_0 d_n(\hat{F}^{\alpha_0}_{s,n}, \check{F}^{\alpha_0}_{s,n}) \xrightarrow{d} U$, where U is a random variable whose distribution depends only on α_0, F, and F_b.*

The proof of the above theorem and the explicit from of U can be found in Appendix E.8. The proof of Theorem 4.2 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 Appendix D.

§5 A heuristic estimator of α_0

In simulations, we observe that the finite sample performance of (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, proved in Appendix A, that describes the shape of our criterion function, and will motivate our procedure.

Lemma 5.1. *$\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is a non-increasing convex function of γ in $(0, 1)$.*

Writing

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

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

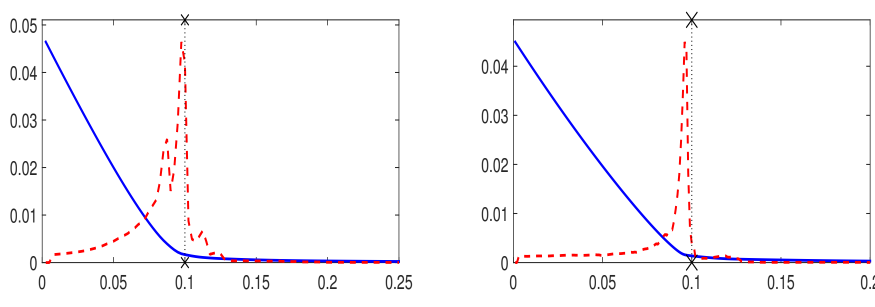

Figure 1: Plots of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red) for $\alpha_0 = 0.1$ and $n = 5000$. Left panel: setting I; right panel: setting II.

of the function $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$, where the left panel corresponds to a mixture of $N(2, 1)$ with $N(0, 1)$ (setting I) and in the right panel we have a mixture of Beta(1,10) and

Uniform$(0,1)$ (setting II). We will 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 maximal. We denote this estimator by $\tilde{\alpha}_0$. Notice 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 the above plots 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 must 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 e.g., Salvador and Chan (2004). However, in Section 8.2.4 and Appendix C, we show some simulation examples where $\tilde{\alpha}_0$ fails to consistently estimate the "elbow" of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$.

§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 (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 $\hat{\alpha}_0^{c_n}$ as discussed in the previous sections) of α_0, a natural nonparametric estimator of F_s is $\check{F}^{\check{\alpha}_n}_{s,n}$, defined as the minimiser of (3). In the following theorem (proved in Appendix E.9 we show that, indeed, $\check{F}^{\check{\alpha}_n}_{s,n}$ is uniformly consistent for estimating F_s. We also derive the rate of convergence of $\check{F}^{\check{\alpha}_n}_{s,n}$.

Theorem 6.1. *Suppose that $\check{\alpha}_n \xrightarrow{P} \alpha_0$. Then, as $n \to \infty$, $\sup_{x\in\mathbb{R}} |\check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x)| \xrightarrow{P} 0$. Furthermore, if $q_n(\check{\alpha}_n - \alpha_0) = O_P(1)$, where $q_n = o(\sqrt{n})$, then $\sup_{x\in\mathbb{R}} q_n|\check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x)| = O_P(1)$. Additionally, for $\hat{\alpha}_0^{c_n}$ as defined in (7), we have*

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

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

An immediate consequence of Theorem 6.1 is that $d_n(\check{F}^{\check{\alpha}_n}_{s,n}, \hat{F}^{\check{\alpha}_n}_{s,n}) \xrightarrow{P} 0$ as $n \to \infty$. Left panel of Fig. 2 shows our estimator $\check{F}^{\check{\alpha}_n}_{s,n}$ along with the true F_s for the same data set used in the right panel of Fig. 1.

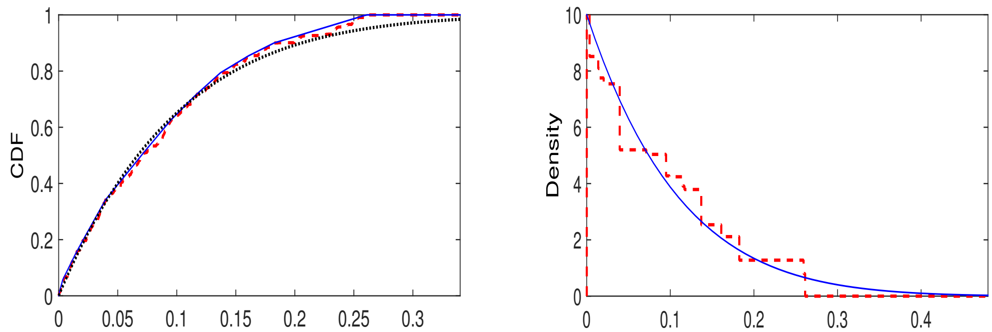

Figure 2: Left panel: Plots of $\check{F}_{s,n}^{\tilde{\alpha}_0}$ (in dashed red), $F_{s,n}^{\dagger}$ (in solid blue) and F_s (in dotted black) for setting II; right panel: plots of $f_{s,n}^{\dagger}$ (in dashed red) and f_s (in solid blue) for setting II.

§6.2 Estimating the density of F_s

Suppose now that F_s has a density f_s. Obtaining nonparametric 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 sub-section 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 authors, 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 $g : [0, \infty) \to \mathbb{R}$, 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 piece-wise constant function obtained by taking the left derivative of $F_{s,n}^{\dagger}$. In the following result, proved in Appendix A.7, we show that both $F_{s,n}^{\dagger}$ and $f_{s,n}^{\dagger}$ are consistent estimators of their population versions.

Theorem 6.2. *Assume that $F_s(0) = 0$ and that F_s is concave on $[0, \infty)$. If $\check{\alpha}_n \xrightarrow{P} \alpha_0$, then, as $n \to \infty$,*

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

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

Computing $F_{s,n}^{\dagger}$ and $f_{s,n}^{\dagger}$ are straightforward, an application of the PAVA gives both the estimators; see e.g., Chapter 1 of Robertson et al. (1988). In Fig. 2 the left panel shows the LCM $F_{s,n}^{\dagger}$ whereas the right panel shows its derivative $f_{s,n}^{\dagger}$ along with the true density f_s for the same data set used in the right panel of Fig. 1.

§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 (DNA) 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 $\leq \beta(1-\alpha_0)$, 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 nonparametric 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 will take f_b to be Uniform$(0,1)$. The *local false discovery rate* (LFDR) is defined as the function $l : (0,1) \to [0,\infty)$, where

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

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 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 get easily interpretable thresholding methods for reporting the "interesting" cases (e.g., $l(x) \leq 0.20$). 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:

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

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

	$n = 1000$						$n = 5000$					
	Setting I			Setting II			Setting I			Setting II		
α	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$
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

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 authors, including Genovese and Wasserman (2004) and Langaas et al. (2005).

§8 Simulation

To investigate the finite sample performance of the estimators 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 at `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 papers we found that discuss methodology for constructing lower confidence bounds for α_0. These procedures are connected and the methods in Meinshausen and Rice (2006) are extensions of those proposed in Genovese and Wasserman (2004). The lower bounds proposed in both the papers approximately satisfy (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) use a constant bounding function, $\delta(t) = 1$, with $\eta_{n,\beta} = \sqrt{\log(2/\beta)/2n}$, whereas Meinshausen and Rice (2006) suggest a class of bounding functions but observe 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 suggested by the authors. We denote the lower bound proposed in Meinshausen and Rice (2006) by $\hat{\alpha}_L^{MR}$, the bound in Genovese and Wasserman (2004) by $\hat{\alpha}_L^{GW}$, 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, 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 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 the table 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 sub-section, we illustrate and compare the performance of different 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 zero 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 realisation of X_{ij} and α be the proportion of false null hypotheses. Let $\bar{x}_i = \sum_{j=1}^{J} x_{ij}/J$ and $s_i^2 = \sum_{j=1}^{J}(x_{ij} - \bar{x}_i)^2/(J-1)$. To test H_{0i} versus H_{1i}, we calculate a two-sided p-value based on a one-sample t-test, with $p_i = 2P(T_{J-1} \geq |\bar{x}_i/\sqrt{s_i^2/J}|)$, 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 $z_i = \frac{1}{\sqrt{L+1}} \sum_{j=i}^{i+L} w_j$ 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 i.i.d. 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 $\mathrm{sgn}(m_i)$, 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 $\hat{\alpha}_0^{S,B}$ to denote the estimator proposed by Storey (2002) when bootstrapping is used to choose the required tuning parameter, and denote by $\hat{\alpha}_0^{S,\lambda}$ 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 $\hat{\alpha}_0^{L}$. We evaluate $\hat{\alpha}_0^{L}$ using the `convest` function in the R library `limma`. We also use the estimator proposed in Meinshausen

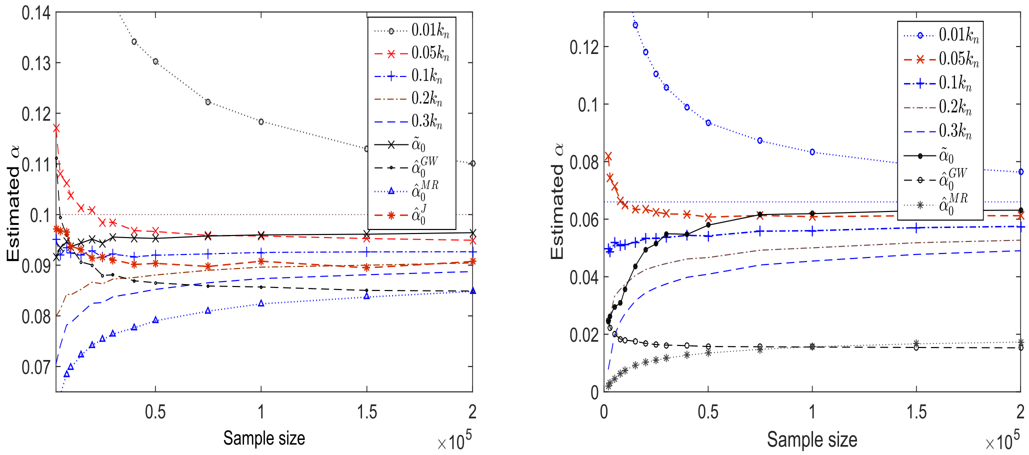

Figure 3: Plots of the means of different estimators of α_0, computed over 500 independent replications, as the sample size increases from 3000 to 2×10^5; left panel: scenario A with $\Sigma = I_{n\times n}$; right panel: scenario B with $L = 0$ and $m^* = 1$. The horizontal line (in dotted blue) indicates the value of α_0.

and Rice (2006) for two bounding functions: $\delta(t) = \sqrt{t(1-t)}$ and $\delta(t) = 1$. For its implementation, we must choose a sequence $\{\beta_n\}$ going to zero as $n \to \infty$. Meinshausen and Rice (2006) did not specify any particular choice of $\{\beta_n\}$ but required the sequence satisfy some conditions. We choose $\beta_n = 0.05/\sqrt{n}$ and denote the estimators by $\hat{\alpha}_0^{MR}$ when $\delta(t) = \sqrt{t(1-t)}$ and by $\hat{\alpha}_0^{GW}$ when $\delta(t) = 1$ (see Genovese and Wasserman (2004)). We also compare our results with $\hat{\alpha}_0^{E}$, the estimator 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) propose estimators when the model is a mixture of Gaussian distributions; we denote the estimator proposed in Section 2.2 of Jin (2008) by $\hat{\alpha}_0^{J}$ and in Section 3.1 of Cai and Jin (2010) by $\hat{\alpha}_0^{CJ}$. 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 the following we investigate a proper choice of c_n. We take $\alpha_0 = 0.1$ and evaluate the performance of $\hat{\alpha}_0^{\tau\times\log\log n}$ for different values of τ, as n increases, for scenarios A and B. The choice $c_n = \tau \times \log\log n$, for different values of τ, is suggested after extensive simulations. We also include $\tilde{\alpha}_0$, $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, and $\hat{\alpha}_0^{J}$ 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 proposed in Section 5. For both scenarios, the sample mean of the estimators of α_0 proposed in this paper converge to the true α_0, as the sample size grows. The methods developed in this paper perform favorably in comparison to $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, and $\hat{\alpha}_0^{J}$. 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.

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 $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{.1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,0.5}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
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)

8.2.1 Cross-validation

In this sub-section, we use c instead of c_n to simplify the notation. In the following we briefly describe our cross-validation procedure. For a K-fold cross validation, we randomly partition 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 defined in (7) using all data except those in $\mathcal{D}_k$ and tuning parameter c. Further, let $\check{F}_{s,n}^{\hat{\alpha}_{0,-k}^c,-k}$ be the estimator of F_s as defined in Lemma 2.1 using $\hat{\alpha}_{0,-k}^c$ and all data except those in $\mathcal{D}_k$. Define the cross-validated estimator of c as

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

where $\hat{F}^k := \hat{\alpha}_{0,-k}^c \check{F}_s^{\hat{\alpha}_{0,-k}^c,-k} + (1-\hat{\alpha}_{0,-k}^c)F_b$. 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 Appendix C.

8.2.2 Performance under independence

In this sub-section, we take $\alpha \in \{0.01, 0.03, 0.05, 0.10\}$ and compare the performance of the different 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 2.4, it is easy to see that in scenario A, the model is identifiable (i.e., $\alpha_0 = \alpha$), while in scenario B, $\alpha_0 = \alpha \times 0.67$. For scenario A, the sample means of $\hat{\alpha}_0^{CV}$, $\tilde{\alpha}_0$, $\hat{\alpha}_0^{J}$, $\hat{\alpha}_0^{L}$, and $\hat{\alpha}_0^{0.1k_n}$ for $k_n = \log\log n$ are comparable. However, the RMSEs of $\tilde{\alpha}_0$ and $\hat{\alpha}_0^{0.1k_n}$ are lower than those of $\hat{\alpha}_0^{CV}$, $\hat{\alpha}_0^{J}$, and $\hat{\alpha}_0^{L}$. For scenario B, the sample means of $\tilde{\alpha}_0$,

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 $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,B}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
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)

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 $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,0.5}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
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)

$\hat{\alpha}_0^{CV}$, and $\hat{\alpha}_0^{0.1k_n}$ are comparable. In scenario B, the performances of $\hat{\alpha}_0^{J}$ and $\hat{\alpha}_0^{CJ}$ are not comparable to the estimators proposed in this paper, as $\hat{\alpha}_0^{J}$ and $\hat{\alpha}_0^{CJ}$ estimate α, while $\tilde{\alpha}_0$, $\hat{\alpha}_0^{CV}$, and $\hat{\alpha}_0^{c_n}$ estimate α_0. Note that $\hat{\alpha}_0^{L}$ fails to estimate α_0 because the underlying assumption inherent in their estimation procedure, that f_s be non-increasing, does not hold. In scenario A, $\hat{\alpha}_0^{S,0.5}$ has the best performance among the different values of λ, while in scenario B, $\hat{\alpha}_0^{S,\lambda}$ has poor performance for all values of $\lambda \in [0, 1]$. Furthermore, $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, $\hat{\alpha}_0^{CJ}$, $\hat{\alpha}_0^{S,B}$ and $\hat{\alpha}_0^{E}$ perform poorly in both scenarios for all values of α_0.

8.2.3 Performance under dependence

The simulation settings of this sub-section 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 of the blocks, all entries

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 $k_n = \log\log n$.

$10\alpha_0$	$\hat{\alpha}_0^{.1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,B}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
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)

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 0.067. Observe that, as in the independence setting, $\hat{\alpha}_0^{GW}$, $\hat{\alpha}_0^{MR}$, $\hat{\alpha}_0^{S,B}$, $\hat{\alpha}_0^{CJ}$, and $\hat{\alpha}_0^{E}$ perform poorly in both scenarios for all values of α_0.

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

Although the heuristic estimator $\tilde{\alpha}_0$ performs quite well in most of the simulation settings considered, there exists scenarios where $\tilde{\alpha}_0$ can fail to consistently estimate α_0. To illustrate this we consider four different CDFs F_s and fix F_b to be the uniform distribution on $(0, 1)$ (see the top left plot of Fig. 4) and compare the performance of $\hat{\alpha}_0^{CV}$, $\tilde{\alpha}_0$, $\hat{\alpha}_0^{0.1k_n}$ with the best performing competing estimators (in each setting).

We see that $\tilde{\alpha}_0$ may fail to estimate the "elbow" of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$, as a function of γ, when F_s has a multi-modal density (see the middle row of Fig. 4). Observe that $\hat{\alpha}_0^{CV}$ and $\hat{\alpha}_0^{0.1k_n}$ perform favorably compared to all competing estimators and in the two scenarios where $\tilde{\alpha}_0$ fails to consistently estimate α_0, 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 the top right plot of Fig. 4 (also see Appendix C 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 of the plot of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ can be useful in checking the

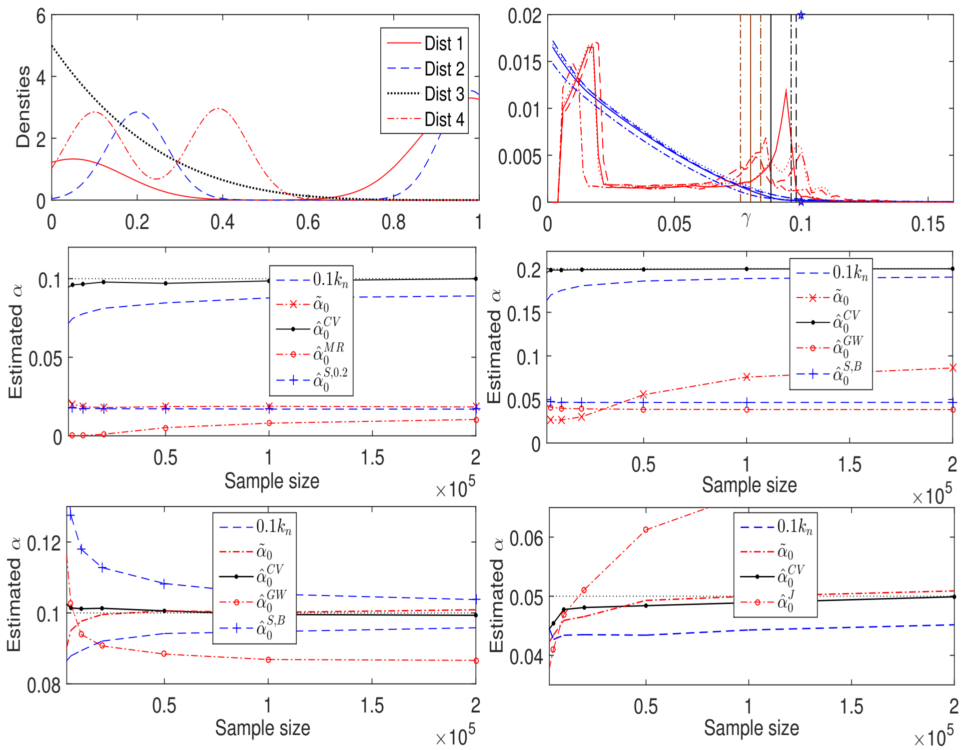

Figure 4: Top row left panel: density functions for different choices of F_s; top row right panel: plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in blue), the scaled second derivative (in red), $\hat{\alpha}_0^{CV}$ (in black), and $\hat{\alpha}_0^{0.1k_n}$ (in green) for 5 independent samples of size 5000 corresponding to "Dist 1"; the blue star denotes α_0. The bottom two rows show the means of different competing estimators of α_0, computed over 500 independent samples for Dist 1-4 (left-right, top-bottom) as sample size increases from 3000 to 2×10^5; in each figure the dotted black line denotes the true α_0.

validity of $\tilde{\alpha}_0$ as an estimator of the "elbow", and thus for α_0.

§9 Real data analysis

§9.1 Prostate 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 involved, the principal goal of the study was to discover a small number of "interesting" genes, that is, 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 is a 6033×102 matrix $\mathbb{X}$ having entries x_{ij} = expression level for gene i on patient j, $i = 1, 2, \ldots, n$, and $j =$

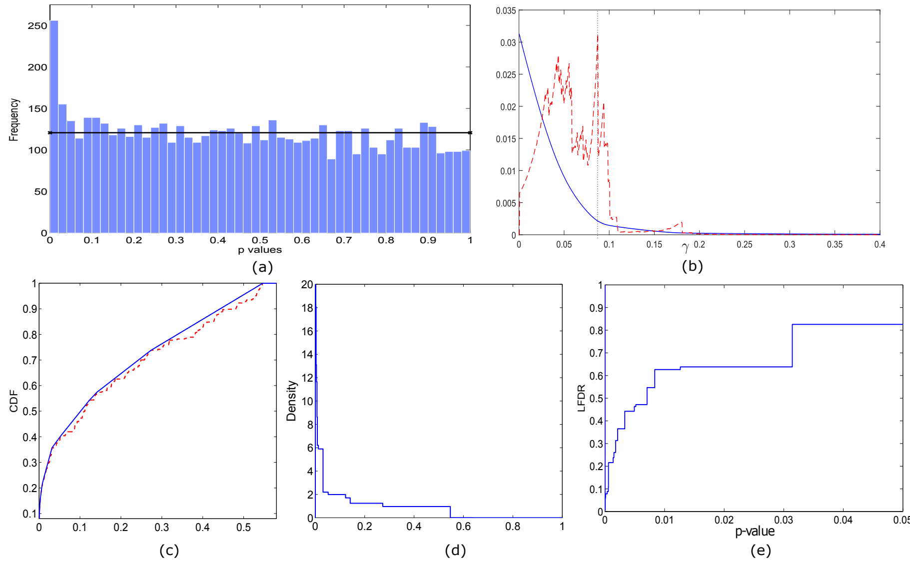

Figure 5: Plots for the prostate data: (a) Histogram of the p-values. The horizontal line (in solid black) indicates the Uniform$(0,1)$ distribution. (b) Plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red). The vertical line (in dotted black) indicates the point of maximum curvature $\tilde{\alpha}_0 = 0.088$. (c) $\check{F}^{\tilde{\alpha}_0}_{s,n}$ (in dotted red) and $F^{\dagger}_{s,n}$ (in solid blue); (d) $f^{\dagger}_{s,n}$; (e) estimated LFDR $\hat{l}$ for p-values less than 0.05.

$1, 2, \ldots, m$, 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., $s_i^2 = (1/50 + 1/52)[\sum_{j=1}^{50}\{x_{ij} - \bar{x}_i(1)\}^2 + \sum_{j=51}^{102}\{x_{ij} - \bar{x}_i(2)\}^2]/100$.

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 using the method developed in Section 6.1. Note that 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 using the methodology developed in Section 6.1 and Section 7, for the prostate data example. The 95% lower confidence bound $\hat{\alpha}_L$ for this data is found to be 0.05. In Table 6, we display estimates of α_0 based on the methods considered in this paper for the prostate data and the Carina data (described below).

§9.2 Carina data – an application in astronomy

In this sub-section we analyse the radial velocity (RV) distribution of stars in Carina, a dwarf spheroidal (dSph) galaxy. The dSph galaxies are low luminosity galaxies that

Table 6: Estimates of α_0 for the two data sets.

Data set	$\hat{\alpha}_0^{0.1k_n}$	$\hat{\alpha}_0^{CV}$	$\tilde{\alpha}_0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{S,B}$	$\hat{\alpha}_0^{J}$	$\hat{\alpha}_0^{CJ}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^{E}$
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	NA

are companions of the Milky Way. The data have been obtained by Magellan and MMT 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 the RV of stars in Carina. For the contaminating stars from the Milky Way 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.

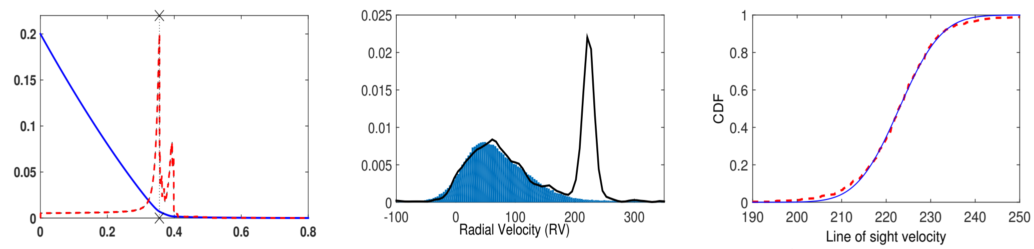

Figure 6: Plots for RV data in Carina dSph; left panel: $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red); middle panel: density of the RV distribution of the contaminating stars overlaid with the (scaled) kernel density estimator of the observed sample; right panel: $\check{F}^{\tilde{\alpha}_0}_{s,n}$ (in dashed red) overlaid with its closest Gaussian distribution (in solid blue).

The 95% lower confidence bound for α_0 is found to be 0.323. The right panel of Fig. 6 shows the estimate of F_s and the closest (in terms of minimising the $L_2(\check{F}^{\tilde{\alpha}_0}_{s,n})$ distance) fitting Gaussian distribution. Astronomers usually assume the distribution of the RVs for these dSph 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), $\hat{\alpha}_0^{CJ}$, is greater than one, while Efron's method (see Efron (2007)), implemented 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 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 proposed in this paper have superior finite sample performance than most competing methods. In contrast to most previous work on this topic the results discussed in this paper hold true even when (1) is not identifiable. Under the assumption that (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 hard 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 (e.g., we observe a second i.i.d. 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 presently unknown. The p-values in the prostate data example, considered in Section 9.1, can have slight dependence. Therefore, investigating the performance and properties of the methods 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 anonymous referees for their careful reading and constructive comments that lead to an improved version of the paper.

§A Appendix 1

§A.1 Proof of Lemma 2.2

From the definition of α_0, we have

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

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

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

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 3.2

Letting $F_s^\gamma = (F - (1-\gamma)F_b)/\gamma$, observe that

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

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

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

To prove the second part of the lemma, notice that for $\gamma \ge \alpha_0$ the result follows from above and the fact that $d_n(F, \mathbb{F}_n) \stackrel{a.s.}{\to} 0$ 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 $n \to \infty$, $\hat{F}_{s,n}^\gamma \stackrel{a.s.}{\to} F_s^\gamma$ point-wise. So, for large enough n, $\hat{F}_{s,n}^\gamma$ is not a valid CDF, whereas $\check{F}_{s,n}^\gamma$ is always a CDF. Thus, $d_n(\hat{F}_{s,n}^\gamma, \check{F}_{s,n}^\gamma)$ converges to something positive.

§A.3 Proof of Lemma 3.3

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

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

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

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

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

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

Thus $\gamma_3 \in A_n$.

§A.4 Proof of Lemma 3.5

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 4.2.

§A.5 Proof of Theorem 4.1

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

$$\begin{aligned} P(\alpha_0 \geq \hat{\alpha}_L) &= P\left(\sqrt{n}\alpha_0 \; d_n(\hat{F}^{\alpha_0}_{s,n}, \check{F}^{\alpha_0}_{s,n}) \leq c_n\right) \\ &\geq P\left(\sqrt{n}\alpha_0 \; d_n(\hat{F}^{\alpha_0}_{s,n}, F^{\alpha_0}_{s}) \leq c_n\right) = H_n(c_n) = 1-\beta, \end{aligned}$$

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

$$P(\alpha_0 \geq \hat{\alpha}_L) = P\left(\sqrt{n}\; d_n(\mathbb{F}_n, F_b) \leq c_n\right) = P\left(\sqrt{n}\; d_n(\mathbb{F}_n, F) \leq c_n\right) = 1-\beta.$$

§A.6 Proof of Lemma 5.1

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

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

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 $0 \leq \eta \leq 1$. Then, by (14) we have the desired result.

§A.7 Proof of Theorem 6.2

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

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

and therefore,

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

By Theorem 6.1, as $\epsilon_n \xrightarrow{P} 0$, we must also have (12).

The second part of the result follows immediately from the lemma is page 330 of Robertson et al. (1988), and is similar to the result in Theorem 7.2.2 of that book.

§B Identifiability of F_s

In this section we continue the discussion on the identifiability of F_s. First, we give some remarks to illustrate Lemmas 2.3 and 2.4.

Remark B.1. *We consider mixtures of Poisson and binomial distributions to illustrate Lemma 2.3. If F_s is Poisson(λ_s) and F_b is Poisson(λ_b), then*

$$\inf_{x\in d(F_b)} \frac{J_{F_s(x)}}{J_{F_b(x)}} = \inf_{k\in\mathbb{N}\cup\{0\}} \frac{\lambda_s^k \exp(-\lambda_s)}{\lambda_b^k \exp(-\lambda_b)} = \exp(\lambda_b - \lambda_s) \inf_{k\in\mathbb{N}\cup\{0\}} \left(\frac{\lambda_s}{\lambda_b}\right)^k.$$

By an application of Lemma 2.3, we have if $\lambda_s < \lambda_b$ then $\alpha_0 = \alpha$; otherwise $\alpha_0 = \alpha(1 - \exp(\lambda_b - \lambda_s))$.

In the case of a binomial mixture, i.e., $F_s = Bin(n, p_s)$ and $F_b = Bin(n, p_b)$,

$$\alpha_0 = \begin{cases} \alpha\left[1 - (\frac{1-p_s}{1-p_b})^n\right], & p_s \ge p_b, \\ \alpha\left[1 - (\frac{p_s}{p_b})^n\right], & p_s < p_b. \end{cases}$$

Remark B.2. *If F_s is $N(\mu_s, \sigma_s^2)$ and F_b ($\neq F_s$) is $N(\mu_b, \sigma_b^2)$ then it can be easily shown that the problem is identifiable if and only if $\sigma_s \le \sigma_b$. When $\sigma_s > \sigma_b$, the model is not identifiable, an application of Lemma 2.4 gives $\alpha_0 = \alpha\big[1 - (\sigma_b/\sigma_s)\exp\big(-\sigma_s\sigma_b(\mu_b - \mu_s)^2/2\big)\big]$. Thus, α_0 increases to α as $|\mu_s - \mu_b|$ tends to infinity. It should be noted that the problem is actually identifiable if we restrict ourselves to the parametric family of a two-component Gaussian mixture model.*

Remark B.3. *Now consider a mixture of exponential random variables, i.e., F_s is $E(a_s, \sigma_s)$ and F_b ($\neq F_s$) is $E(a_b, \sigma_b)$, where $E(a, \sigma)$ is the distribution that has the density $(1/\sigma)\exp(-(x-a)/\sigma)\mathbf{1}_{(a,\infty)}(x)$. In this case, the problem is identifiable if $a_s > a_b$, as this implies the support of F_s is a proper subset of the support of F_b. But when $a_s \le a_b$, the problem is identifiable if and only if $\sigma_s \le \sigma_b$.*

Remark B.4. *It is also worth pointing out that even in cases where the problem is not identifiable the difference between the true mixing proportion* α *and the estimand* α_0 *may be very small. Consider the hypothesis test* $H_0 : \theta = 0$ *versus* $H_1 : \theta \neq 0$ *for the model* $N(\theta, 1)$ *with test statistic* $\bar{X}$*. The density of the p-values under* θ *is*

$$f_\theta(p) = \frac{1}{2} e^{-m\theta^2/2} [e^{-\sqrt{m}\theta^2 \Phi^{-1}(1-p/2)} + e^{\sqrt{m}\theta^2 \Phi^{-1}(1-p/2)}],$$

where m *is the sample size. Here* $f_\theta(1) = e^{-m\theta^2/2} > 0$*, so the model is not identifiable. As* F_b *is uniform, it can be easily verified that* $\alpha_0 = \alpha - \alpha \inf_p f_\theta(p)$*. However, as the value of* f_θ *decreases exponentially with* m*, in many practical situations, where* m *is not too small, the difference between* α *and* α_0 *will be negligible.*

In the following lemma, we try to find the relationship between α and α_0 when F is a general CDF.

Lemma B.5. *Suppose that*

$$F = \kappa F^{(a)} + (1 - \kappa) F^{(d)}, \tag{16}$$

where $F^{(a)}$ *is an absolutely continuous CDF and* $F^{(d)}$ *is a piecewise constant CDF, for some* $\kappa \in (0, 1)$*. Then*

$$\alpha_0 = \alpha - \min \left\{ \frac{\alpha\kappa_s - \alpha_0^{(a)}\kappa}{\kappa_b}, \frac{\alpha(1-\kappa_s) - \alpha_0^{(d)}(1-\kappa)}{(1-\kappa_b)} \right\},$$

where $\alpha_0^{(a)}$ *and* $\alpha_0^{(d)}$ *are defined as in (4), but with* $\{F^{(a)}, F_b^{(a)}\}$ *and* $\{F^{(d)}, F_b^{(d)}\}$*, respectively (instead of* $\{F, F_b\}$*). Similarly,* κ_s *and* κ_b *are defined as in (16), but for* F_s *and* F_b*, respectively.*

Proof. From the definition of κ_s and κ_b, we have $F_s = \kappa_s F_s^{(a)} + (1 - \kappa_s) F_s^{(d)}$, and $F_b = \kappa_b F_b^{(a)} + (1 - \kappa_b) F_b^{(d)}$. Thus from (1), we get

$$F = \alpha\kappa_s F_s^{(a)} + (1-\alpha)\kappa_b F_b^{(a)} + \alpha(1-\kappa_s) F_s^{(d)} + (1-\alpha)(1-\kappa_s) F_b^{(d)}.$$

Now using the definition of κ, we see that $\kappa = \alpha\kappa_s + (1-\alpha)\kappa_b,\ 1 - \kappa = \alpha(1-\kappa_s) + (1-\alpha)(1-\kappa_b)$. If we write

$$F^{(a)} = \alpha^{(a)} F_s^{(a)} + (1 - \alpha^{(a)}) F_b^{(a)},$$

it can easily seen that $\alpha^{(a)} = \frac{\alpha\kappa_s}{\kappa}$; and similarly, $\alpha^{(d)} = \frac{\alpha(1-\kappa_s)}{1-\kappa}$. Then, we can find $\alpha_0^{(d)}$

and $\alpha_0^{(a)}$ as in Lemmas 2.3 and 2.4, respectively. Note that

$$
\begin{aligned}
& \sup\{0 \le \epsilon \le 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} \\
= \; & \sup\left\{0 \le \epsilon \le 1 : \alpha(\kappa_s F_s^{(a)} + (1-\kappa_s)F_s^{(d)}) - \epsilon(\kappa_b F_b^{(a)} + (1-\kappa_b)F_b^{(d)}) \text{ is a sub-CDF}\right\} \\
= \; & \sup\left\{0 \le \epsilon \le 1 : \text{both } \alpha\kappa_s F_s^{(a)} - \epsilon\kappa_b F_b^{(a)}, \alpha(1-\kappa_s)F_s^{(d)} - \epsilon(1-\kappa_b)F_b^{(d)} \text{ are sub-CDFs}\right\} \\
= \; & \min\left(\sup\left\{0 \le \epsilon \le 1 : \alpha\kappa_s F_s^{(a)} - \epsilon\kappa_b F_b^{(a)} \text{ is a sub-CDF}\right\},\right. \\
& \qquad\qquad \left.\sup\left\{0 \le \epsilon \le 1 : \alpha(1-\kappa_s)F_s^{(d)} - \epsilon(1-\kappa_b)F_b^{(d)} \text{ is a sub-CDF}\right\}\right) \\
= \; & \min\left(\frac{\alpha\kappa_s}{\kappa_b} \operatorname{ess\,inf} \frac{f_s^{(a)}}{f_b^{(a)}}, \frac{\alpha(1-\kappa_s)}{(1-\kappa_b)} \inf_{x \in d(F_b^{(d)})} \frac{J_{F_s^{(d)}}(x)}{J_{F_b^{(d)}}(x)}\right) \\
= \; & \min\left(\frac{(\alpha\kappa_s - \alpha_0^{(a)}\kappa)}{\kappa_b}, \frac{(\alpha(1-\kappa_s) - \alpha_0^{(d)}(1-\kappa))}{(1-\kappa_b)}\right),
\end{aligned}
$$

where J_G and $d(J_G)$ are defined before Lemma 2.3 and we use the notion that $\frac{0}{0} = 1$. Hence, by (5) the result follows. □

Lemma 2.5 is now a corollary of this result.

§C Performance comparison of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$

In Figs. 7 and 8 we present further simulation experiments to investigate the finite sample performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$ across different simulation scenarios. In each setting we also include the performance of the best performing competing estimators discussed in Section 8.2.

§D Detection of sparse heterogeneous mixtures

In this section we draw a connection between the lower confidence bound developed in Section 4 and the *Higher Criticism* method of Donoho and Jin (2004) for detection of sparse heterogeneous mixtures. The detection of heterogeneity in sparse models arises in many applications, e.g., detection of a disease outbreak (see Kulldorff et al. (2005)) or early detection of bioweapons use (see Donoho and Jin (2004)). Generally, in large scale multiple testing problems, when the non-null effect is sparse it is important to detect the existence of non-null effects (see Cai et al. (2007)).

Donoho and Jin (2004) consider n i.i.d. data from one of the two possible situations:

$$
\begin{aligned}
& H_0 : X_i \sim F_b, \quad 1 \le i \le n, \\
& H_1^{(n)} : X_i \sim F^n := \alpha_n F_{n,s} + (1-\alpha_n)F_b, \quad 1 \le i \le n,
\end{aligned}
$$

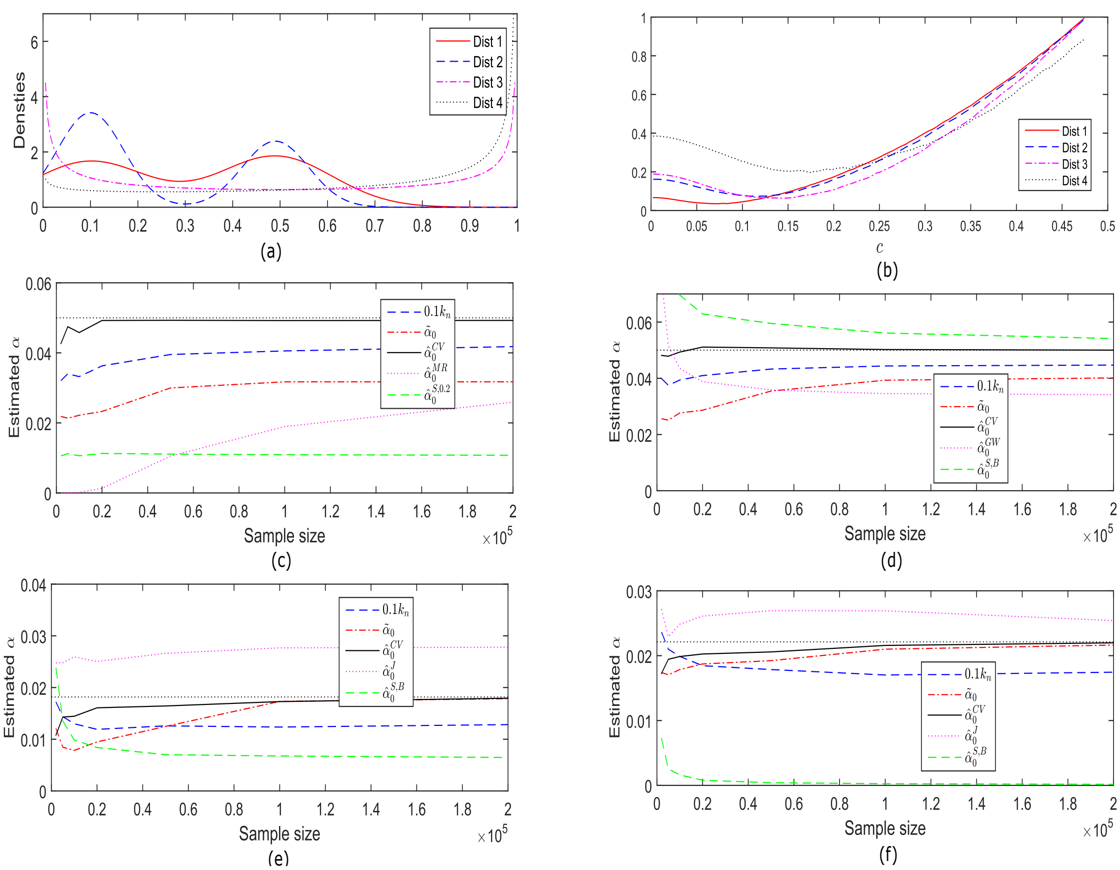

Figure 7: Plots comparing the performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$; (a) density functions for four different choices of F_s; (b) plot of the average of $\sum_{k=1}^{K} \int (\mathbb{F}_n^k - \hat{F}^k)^2 d\mathbb{F}_n^k$ (see (13) of the main paper), as a function of c, computed over 500 independent samples of size 50000 corresponding to Dist 1-4; (c)-(f) gives the means of different competing estimators of α_0, computed over 500 independent samples for Dist 1-4 respectively (in each figure the horizontal dotted black line denotes the true α_0).

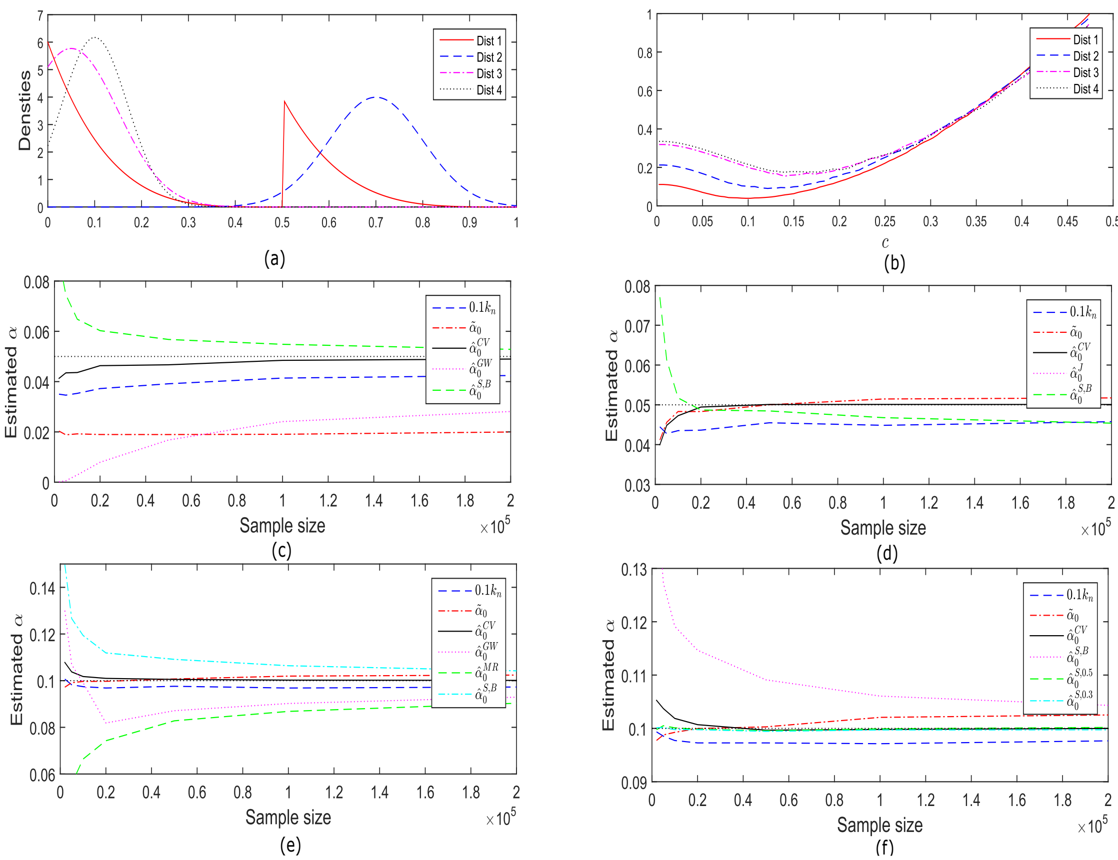

Figure 8: Plots comparing the performance of $\hat{\alpha}_0^{c_n}$, $\hat{\alpha}_0^{CV}$, and $\tilde{\alpha}_0$; (a) density functions for four different choices of F_s; (b) plot of the average of $\sum_{k=1}^{K} \int (\mathbb{F}_n^k - \hat{F}^k)^2 d\mathbb{F}_n^k$ (see (13) of the main paper), as a function of c, computed over 500 independent samples of size 50000 corresponding to Dist 1-4; (c)-(f) gives the means of different competing estimators of α_0, computed over 500 independent samples for Dist 1-4 respectively (in each figure the horizontal dotted black line denotes the true α_0).

where $\alpha_n \sim n^{-\lambda}$ and $F_{n,s}$ is such that $d(F_{n,s}, F_b)$ is bounded away from 0. In Donoho and Jin (2004) the main focus is on testing H_0, i.e., $\alpha_n = 0$. We can test this hypothesis by rejecting H_0 when $\hat{\alpha}_L > 0$. The following lemma shows that indeed this yields a valid testing procedure for $\lambda < 1/2$.

Theorem D.1. *If $\alpha_n \sim n^{-\lambda}$, for $\lambda < 1/2$, then $P_{H_0}(\text{Reject } H_0) = \beta$ and $P_{H_1^{(n)}}(\hat{\alpha}_L > 0) \to 1$ as $n \to \infty$.*

Proof. Note that $\{\hat{\alpha}_L > 0\}$ is equivalent to $\{c_n \le \sqrt{n} d_n(\mathbb{F}_n, F_b)\}$ which shows that

$$\begin{aligned} c_n &\le \sqrt{n} d_n(\mathbb{F}_n, (1-\alpha_n)F_b + \alpha_n F_{n,s}) + \sqrt{n} d_n(\alpha_n F_b, \alpha_n F_{n,s}) \\ &= \sqrt{n} d_n(\mathbb{F}_n, F^n) + \alpha_n \sqrt{n} d_n(F_{n,s}, F_b), \end{aligned}$$

where c_n is chosen as in Theorem 4.1. It is easy to see that $\sqrt{n} d_n(\mathbb{F}_n, F^n)$ is $O_P(1)$ and $\alpha_n \sqrt{n} d_n(F_{n,s}, F_b) \to \infty$, for $\lambda < 1/2$, which shows that $P_{H_1^{(n)}}(\hat{\alpha}_L > 0) \to 1$. It can be easily seen that $P_{H_0}(\hat{\alpha}_L > 0) = P_{H_0}(\text{Reject } H_0) = \beta$. $\square$

§E Proofs of theorems and lemmas in the main paper

§E.1 Proof of Lemma 2.3

When $d(F_b) \not\subset d(F_s)$, there exists a $x \in d(F_b) - d(F_s)$, i.e., there exists a x which satisfies $F_b(x) - F_b(x-) > 0$ and $F_s(x) - F_s(x-) = 0$. Then for all $\epsilon > 0$, $F_s(x-) - \epsilon F_b(x-) > F_s(x) - \epsilon F_b(x)$. This shows that $F_s - \epsilon F_b$ cannot be a sub-CDF, and hence by Lemma 2.2 the model is identifiable. Now let us assume that $d(F_b) \subset d(F_s)$.

$$\begin{aligned} \{0 \le \epsilon \le 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} &= \{0 \le \epsilon \le 1 : \alpha J_{F_s}(x) - \epsilon J_{F_b}(x) \ge 0, \ \forall x \in d(J_{F_b})\} \\ &= \left\{0 \le \epsilon \le 1 : \frac{J_{F_s}(x)}{J_{F_b}(x)} \ge \frac{\epsilon}{\alpha}, \ \forall x \in d(J_{F_b})\right\} \\ &= \left\{0 \le \epsilon \le 1 : \inf_{x \in d(F_b)} \frac{J_{F_s}(x)}{J_{F_b}(x)} \ge \frac{\epsilon}{\alpha}\right\}. \end{aligned}$$

Therefore, using (5), we get the desired result.

§E.2 Proof of Lemma 2.4

From (5), we have

$$\begin{aligned} \alpha_0 &= \alpha - \sup\{0 \le \epsilon \le 1 : \alpha F_s - \epsilon F_b \text{ is a sub-CDF}\} \\ &= \alpha - \sup\{0 \le \epsilon \le 1 : \alpha f_s(x) - \epsilon f_b(x) \ge 0 \text{ almost every } x\} \\ &= \alpha - \sup\left\{0 \le \epsilon \le 1 : \alpha \frac{f_s}{f_b}(x) \ge \epsilon \text{ almost every } x\right\} \\ &= \alpha \left\{1 - \operatorname{ess\,inf} \frac{f_s}{f_b}\right\}. \end{aligned}$$

§E.3 Proof of Theorem 3.1

Without loss of generality, we can assume that F_b is the uniform distribution on $(0,1)$ and, for clarity, in the following we write U instead of F_b. Let us define

$$\begin{aligned} A &:= \left\{\gamma \in (0,1] : \frac{F-(1-\gamma)U}{\gamma} \text{ is a valid CDF}\right\}, \\ A^Y &:= \left\{\gamma \in (0,1] : \frac{G-(1-\gamma)U\circ\Psi}{\gamma} \text{ is a valid CDF}\right\}. \end{aligned}$$

Since $\alpha_0 = \inf A$, and $\alpha_0^Y = \inf A^Y$ for the first part of the theorem it is enough to show that $A = A^Y$. Let us first show that $A^Y \subset A$. Suppose $\eta \in A^Y$. We first show that $(F-(1-\eta)U)/\eta$ is a non-decreasing function. For all $t_1 \le t_2$, we have that

$$\frac{G(t_1)-(1-\eta)U(\Psi(t_1))}{\eta} \le \frac{G(t_2)-(1-\eta)U(\Psi(t_2))}{\eta}.$$

Let $y_1 \le y_2$. Then,

$$\frac{G(\Psi^{-1}(y_1))-(1-\eta)U(\Psi(\Psi^{-1}(y_1)))}{\eta} \le \frac{G(\Psi^{-1}(y_2))-(1-\eta)U(\Psi(\Psi^{-1}(y_2)))}{\eta},$$

since $y_1 \le y_2 \Rightarrow \Psi^{-1}(y_1) \le \Psi^{-1}(y_2)$. However, as Ψ is continuous, $\Psi(\Psi^{-1}(y)) = y$ and $G(\Psi^{-1}(y)) = \alpha F_s(y) + (1-\alpha)U(y) = F(y)$. Hence, we have

$$\frac{F(y_1)-(1-\eta)U(y_1)}{\eta} \le \frac{F(y_2)-(1-\eta)U(y_2)}{\eta}.$$

As F and U are CDFs, it is easy to see that $\lim_{x\to-\infty}(F(x)-(1-\eta)U(x))/\eta = 0$, $\lim_{x\to\infty}(F(x)-(1-\eta)U(x))/\eta = 1$ and $(F-(1-\eta)U)/\eta$ is a right continuous function. Hence, for $\eta \in A^Y$, $(F-(1-\eta)U)/\eta$ is a CDF and thus, $\eta \in A$. We can similarly prove $A \subset A^Y$. Therefore, $A = A^Y$ and $\alpha_0 = \alpha_0^Y$.

Note that

$$\gamma d_n(\hat{F}_{s,n}^\gamma, \check{F}_{s,n}^\gamma) = \min_{W\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \{W(X_i) - \hat{F}_{s,n}^\gamma(X_i)\}^2,$$

where $\mathcal{F}$ is the class of all CDFs. For the second part of theorem it is enough to show that

$$\min_{W\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \{W(X_i) - \hat{F}_{s,n}^\gamma(X_i)\}^2 = \min_{B\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \{B(Y_i) - \hat{G}_{s,n}^\gamma(Y_i)\}^2.$$

First note that

$$\begin{aligned} \mathbb{G}_n(y) &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\Psi^{-1}(X_i) \le y\} \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le \Psi(y)\} \\ &= \mathbb{F}_n(\Psi(y)). \end{aligned}$$

Thus, from the definition of $\hat{G}^{\gamma}_{s,n}$, we have

$$\begin{aligned}\hat{G}^{\gamma}_{s,n}(Y_i) &= \frac{\mathbb{F}_n(\Psi(Y_i)) - (1-\gamma)U(\Psi(Y_i))}{\gamma} \\ &= \frac{\mathbb{F}_n(X_i) - (1-\gamma)U(X_i)}{\gamma} = \hat{F}^{\gamma}_{s,n}(X_i).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\{B(Y_i) - \hat{G}^{\gamma}_{s,n}(Y_i)\}^2 &= \frac{1}{n}\sum_{i=1}^{n}\{B(Y_i) - \hat{F}^{\gamma}_{s,n}(X_i)\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\{B(\Psi^{-1}(X_i)) - \hat{F}^{\gamma}_{s,n}(X_i)\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\{W(X_i) - \hat{F}^{\gamma}_{s,n}(X_i)\}^2,\end{aligned}$$

where $W(x) := B(\Psi^{-1}(x))$. W is a valid CDF as Ψ^{-1} is non-decreasing.

§E.4 Proof of Theorem 3.4

We need to show that $P(|\hat{\alpha}_0^{c_n} - \alpha_0| > \epsilon) \to 0$ for any $\epsilon > 0$. Let us first show that

$$P(\hat{\alpha}_0^{c_n} - \alpha_0 < -\epsilon) \to 0.$$

The statement is obviously true if $\alpha_0 \leq \epsilon$. So let us assume that $\alpha_0 > \epsilon$. Suppose $\hat{\alpha}_0^{c_n} - \alpha_0 < -\epsilon$, i.e., $\hat{\alpha}_0^{c_n} < \alpha_0 - \epsilon$. Then by the definition of $\hat{\alpha}_0^{c_n}$ and the convexity of A_n, we have $(\alpha_0 - \epsilon) \in A_n$ (as A_n is a convex set in $[0,1]$ with $1 \in A_n$ and $\hat{\alpha}_0^{c_n} \in A_n$), and thus

$$d_n(\hat{F}^{\alpha_0-\epsilon}_{s,n}, \check{F}^{\alpha_0-\epsilon}_{s,n}) \leq \frac{c_n}{\sqrt{n}(\alpha_0-\epsilon)}. \tag{17}$$

But by (10) the left-hand side of (17) goes to a non-zero constant in probability. Hence, if $c_n/\sqrt{n} \to 0$,

$$P(\hat{\alpha}_0^{c_n} - \alpha_0 < -\epsilon) \leq P\left(d_n(\hat{F}^{\alpha_0-\epsilon}_{s,n}, \check{F}^{\alpha_0-\epsilon}_{s,n}) \leq \frac{c_n}{\sqrt{n}(\alpha_0-\epsilon)}\right) \to 0.$$

This completes the proof of the first part of the claim.

Now suppose that $\hat{\alpha}_0^{c_n} - \alpha_0 > \epsilon$. Then,

$$\begin{aligned}\hat{\alpha}_0^{c_n} - \alpha_0 > \epsilon \quad &\Rightarrow \quad \sqrt{n}d_n(\hat{F}^{\alpha_0+\epsilon}_{s,n}, \check{F}^{\alpha_0+\epsilon}_{s,n}) \geq \frac{c_n}{\alpha_0+\epsilon} \\ &\Rightarrow \quad \sqrt{n}d_n(\mathbb{F}_n, F) \geq c_n.\end{aligned}$$

The first implication follows from the definition of $\hat{\alpha}_0^{c_n}$, while the second implication is true by Lemma 3.2. The right-hand side of the last inequality is (asymptotically similar to) the Cramér–von Mises statistic for which the asymptotic distribution is well-known and thus if $c_n \to \infty$ the result follows.

§E.5 Proof of Theorem 3.6

As the proof of this result is slightly involved we break it into a number of lemmas (whose proofs are provided later in this sub-section) and give the main arguments below.

We need to show that given any $\epsilon > 0$, we can find an $M > 0$ and $n_0 \in \mathbb{N}$ (depending on ϵ) for which $\sup_{n>n_0} P(r_n|\hat{\alpha}_0^{c_n} - \alpha_0| > M) \le \epsilon$.

Lemma E.1. *If $c_n \to \infty$, then for any $M > 0$, $\sup_{n>n_0} P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > M\right) < \epsilon$, for large enough $n_0 \in \mathbb{N}$.*

Finding an r_n such that $P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) < -M\right) < \epsilon$ for large enough n is more complicated. We start with some notation. Let $\mathcal{F}$ be the class of all CDFs and $\mathbb{H}$ be the Hilbert space $L_2(F) := \{f : \mathbb{R} \to \mathbb{R} | \int f^2 dF < \infty\}$. For a closed convex subset $\mathcal{K}$ of $\mathbb{H}$ and $h \in \mathbb{H}$, we define the projection of h onto $\mathcal{K}$ as

$$\Pi(h|\mathcal{K}) := \arg\min_{f\in\mathcal{K}} d(f, h), \tag{18}$$

where d stands for the $L_2(F)$ distance, i.e., if $g, h \in \mathbb{H}$, then $d^2(g, h) = \int (g - h)^2 dF$. We define the tangent cone of $\mathcal{F}$ at $f_0 \in \mathcal{F}$, as

$$T_{\mathcal{F}}(f_0) := \{\lambda(f - f_0) : \lambda \ge 0, f \in \mathcal{F}\}. \tag{19}$$

For any $H \in \mathcal{F}$ and $\gamma > 0$, let us define

$$\hat{H}^\gamma := \frac{H - (1-\gamma)F_b}{\gamma}, \quad \check{H}_n^\gamma := \arg\min_{G\in\mathcal{F}} \gamma d_n(\hat{H}^\gamma, G), \quad \text{and} \quad \bar{H}_n^\gamma := \arg\min_{G\in\mathcal{F}} \gamma d(\hat{H}^\gamma, G).$$

For $H = \mathbb{F}_n$ and $\gamma = \alpha_0$ we define the three quantities above and call them $\hat{F}_{s,n}^{\alpha_0}$, $\check{F}_{s,n}^{\alpha_0}$, and $\bar{F}_{s,n}^{\alpha_0}$ respectively. Note that

$$P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) < -M\right) = P(\sqrt{n}\gamma_n \; d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) < c_n), \tag{20}$$

where $\gamma_n = \alpha_0 - M/r_n$. To study the limiting behavior of $d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n})$ we break it as the sum of $d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n})$ and $d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n})$. The following two lemmas (proved in Sections E.5.2 and E.5.3 respectivley) give the asymptotic behavior of the two terms. The proof of Lemma E.3 uses the functional delta method (cf. Theorem 20.8 of Van der Vaart (1998)) for the projection operator; see Theorem 1 of Fils-Villetard et al. (2008).

Lemma E.2. *If $\sqrt{n}/r_n^2 \to 0$, then $U_n := \sqrt{n}\gamma_n d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - \sqrt{n}\gamma_n d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} 0$.*

Lemma E.3. *If $c_n \to \infty$, then*

$$\frac{\sqrt{n}\gamma_n}{c_n M} d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} \left\{\int V^2 dF\right\}^{1/2} > 0$$

where

$$V := (F_s^{\alpha_0} - F_b) - \Pi\left(F_s^{\alpha_0} - F_b \middle| T_{\mathcal{F}}(F_s^{\alpha_0})\right) \neq 0$$

and

$$F_s^{\alpha_0} := \frac{F - (1-\alpha_0)F_b}{\alpha_0}. \tag{21}$$

Using (20), and the notation introduced in the above two lemmas we see that

$$P\left(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) < -M\right) = P\left(\frac{1}{c_n}U_n + \frac{\sqrt{n}\gamma_n}{c_n}d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) < 1\right). \tag{22}$$

However, $U_n \xrightarrow{P} 0$ (by Lemma E.2) and $\frac{\sqrt{n}\gamma_n}{c_n M}d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} \int V^2 dF$ (by Lemma E.3). The result now follows from (22), by taking a large enough M.

E.5.1 Proof of Lemma E.1

Note that

$$\begin{aligned} P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > M) \leq P\left(\hat{\alpha}_0^{c_n} > \alpha_0\right) &= P\left(\sqrt{n}\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) > c_n\right) \\ &\leq P\left(\sqrt{n}\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, F_s^{\alpha_0}) > c_n\right) \\ &= P\left(\sqrt{n}d_n(\mathbb{F}_n, F) > c_n\right) \to 0, \end{aligned}$$

as $c_n \to \infty$, since $\sqrt{n}d_n(\mathbb{F}_n, F) = O_P(1)$. Therefore, the result holds for sufficiently large n.

E.5.2 Proof of Lemma E.2

It is enough to show that

$$W_n := n\gamma_n^2 d_n^2(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - n\gamma_n^2 d^2(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} 0, \tag{23}$$

since $U_n^2 \leq |W_n|$. Note that

$$\begin{aligned} \check{F}_{s,n}^{\gamma_n} &= \arg\min_{G \in \mathcal{F}} d_n(\mathbb{F}_n, \gamma_n G + (1-\gamma_n)F_b), \\ \bar{F}_{s,n}^{\gamma_n} &= \arg\min_{G \in \mathcal{F}} d(\mathbb{F}_n, \gamma_n G + (1-\gamma_n)F_b). \end{aligned}$$

For each positive integer n and $c > 0$, we introduce the following classes of functions:

$$\begin{aligned} \mathcal{G}_c(n) &= \left\{\sqrt{n}(G - (1-\gamma_n)F_b - \gamma_n \check{G}_n^{\gamma_n})^2 : G \in \mathcal{F},\ \|G - F\| < \frac{c}{\sqrt{n}}\right\}, \\ \mathcal{H}_c(n) &= \left\{\sqrt{n}(H - (1-\gamma_n)F_b - \gamma_n \bar{H}_n^{\gamma_n})^2 : H \in \mathcal{F},\ \|H - F\| < \frac{c}{\sqrt{n}}\right\}. \end{aligned}$$

Let us also define

$$D_n := \sup_{t \in \mathbb{R}} \sqrt{n}|\mathbb{F}_n(t) - F(t)| = \|\mathbb{F}_n - F\|.$$

From the definition of the minimisers $\check{F}_{s,n}^{\gamma_n}$ and $\bar{F}_{s,n}^{\gamma_n}$, we see that

$$\begin{aligned} \gamma_n^2\ |d_n^2(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) - d^2(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n})| \leq \max\big\{&|(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \check{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)|\,, \\ &|(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \bar{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)|\big\}. \end{aligned} \tag{24}$$

Observe that

$$n\gamma_n^2\ [(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \check{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)] = \sqrt{n}(\mathbb{P}_n - P)[g_n] = \nu_n(g_n),$$

where $g_n := \sqrt{n}\{\mathbb{F}_n - \gamma_n \check{F}_{s,n}^{\gamma_n} - (1-\gamma_n)F_b\}^2$, $\mathbb{P}_n$ denotes the empirical measure of the data, and $\nu_n := \sqrt{n}(\mathbb{P}_n - P)$ denotes the usual empirical process. Similarly,

$$n\gamma_n^2\ [(d_n^2 - d^2)(\mathbb{F}_n, \gamma_n \bar{F}_{s,n}^{\gamma_n} + (1-\gamma_n)F_b)] = \sqrt{n}(\mathbb{P}_n - P)[h_n] = \nu_n(h_n),$$

where $h_n := \sqrt{n}\{\mathbb{F}_n - \gamma_n \bar{F}_{s,n}^{\gamma_n} - (1-\gamma_n)F_b\}^2$. Thus, combining (23), (24) and the above two displays, we get, for any $\delta > 0$,

$$P(|W_n| > \delta) \le P\left(|\nu_n(g_n)| > \delta\right) + P\left(|\nu_n(h_n)| > \delta\right). \tag{25}$$

The first term in the right hand side of (25) can be bounded above as

$$\begin{aligned}
P(|\nu_n(g_n)| > \delta) &= P(|\nu_n(g_n)| > \delta, g_n \in \mathcal{G}_c(n)) + P(|\nu_n(g_n)| > \delta, g_n \notin \mathcal{G}_c(n)) \\
&\le P(|\nu_n(g_n)| > \delta, g_n \in \mathcal{G}_c(n)) + P(g_n \notin \mathcal{G}_c(n)) \\
&\le P\left(\sup_{g \in \mathcal{G}_c(n)} |\nu_n(g)| > \delta\right) + P(g_n \notin \mathcal{G}_c(n)) \\
&\le \frac{1}{\delta} E\left(\sup_{g \in \mathcal{G}_c(n)} |\nu_n(g)|\right) + P(g_n \notin \mathcal{G}_c(n)) \\
&\le J_{[\,]} \frac{P[G_{c,n}^2]}{\delta} + P(g_n \notin \mathcal{G}_c(n)),
\end{aligned} \tag{26}$$

where $G_{c,n} := 6c^2/\sqrt{n} + 16\sqrt{n}\frac{M^2}{r_n^2}\|F_s^{\alpha_0} - F_b\|^2$ is an envelope for $\mathcal{G}_c(n)$ and $J_{[\,]}$ is a constant. Note that to derive the last inequality, we have used the maximal inequality in Corollary (4.3) of Pollard (1989); the class $\mathcal{G}_c(n)$ is "manageable" in the sense of Pollard (1989) (as a consequence of equation (2.5) of Van de Geer (2000)).

To see that $G_{c,n}$ is an envelope for $\mathcal{G}_c(n)$, observe that for any $G \in \mathcal{F}$,

$$G - (1-\gamma_n)F_b \;=\; G - F + \frac{M}{r_n}(F_s^{\alpha_0} - F_b) + \gamma_n F_s^{\alpha_0}.$$

Hence,

$$F_s^{\alpha_0} - \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| - \frac{\|G-F\|}{\gamma_n} \le \frac{G - (1-\gamma_n)F_b}{\gamma_n} \le F_s^{\alpha_0} + \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| + \frac{\|G-F\|}{\gamma_n}.$$

As the two bounds are monotone, from the properties of isotonic estimators (see e.g., Theorem 1.3.4 of Robertson et al. (1988)), we can always find a version of $\check{G}_s^{\gamma_n}$ such that

$$F_s^{\alpha_0} - \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| - \frac{\|G-F\|}{\gamma_n} \le \check{G}_s^{\gamma_n} \le F_s^{\alpha_0} + \frac{M}{r_n\gamma_n}\|F_s^{\alpha_0} - F_b\| + \frac{\|G-F\|}{\gamma_n}.$$

Therefore,

$$-2\frac{M}{r_n}\|F_s^{\alpha_0}-F_b\|-\|G-F\| \leq \gamma_n\check{G}_s^{\gamma_n}-\gamma_n F_s^{\alpha_0}-\frac{M}{r_n}(F_s^{\alpha_0}-F_b) \leq 2\frac{M}{r_n}\|F_s^{\alpha_0}-F_b\|+\|G-F\|. \tag{27}$$

Thus, for $\sqrt{n}(G-(1-\gamma_n)F_b-\gamma_n\check{G}_s^{\gamma_n})^2 \in \mathcal{G}_c(n)$,

$$\begin{aligned}
(G-(1-\gamma_n)F_b-\gamma_n\check{G}_s^{\gamma_n})^2 &= \left[(G-F)+\left(\gamma_n\check{G}_s^{\gamma_n}-\gamma_n F_s^{\alpha_0}-\frac{M}{r_n}(F_b-F_s^{\alpha_0})\right)\right]^2 \\
&\leq 2(G-F)^2+2\left(\gamma_n\check{G}_s^{\gamma_n}-\gamma_n F_s^{\alpha_0}-\frac{M}{r_n}(F_b-F_s^{\alpha_0})\right)^2 \\
&\leq 2\|G-F\|^2+2\left(2\frac{M}{r_n}\|F_s^{\alpha_0}-F_b\|+\|G-F\|\right)^2 \\
&\leq 6\|G-F\|^2+16\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2 \\
&\leq 6c^2+16\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2=\frac{G_{c,n}}{\sqrt{n}},
\end{aligned}$$

where the second inequality follows from (27). From the definition of g_n and D_n^2, we have $|g_n(t)| \leq \frac{6}{\sqrt{n}}D_n^2+16\sqrt{n}\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2$, for all $t \in \mathbb{R}$. As $D_n = O_P(1)$, for any given $\epsilon > 0$, there exists $c > 0$ (depending on ϵ) such that

$$P(g_n \notin \mathcal{G}_c(n)) = P\left(\|\mathbb{F}_n-F\| \geq \frac{c}{\sqrt{n}}\right) = P(D_n \geq c) \leq \epsilon, \tag{28}$$

for all sufficiently large n.

Therefore, for any given $\delta > 0$ and $\epsilon > 0$, we can make both $J\{6\frac{c^2}{\sqrt{n}}+16\sqrt{n}\frac{M^2}{r_n^2}\|F_s^{\alpha_0}-F_b\|^2\}^2$ and $P(g_n \notin \mathcal{G}_c(n))$ less than ϵ for large enough n and $c(>0)$, using the fact that $\sqrt{n}/r_n^2 \to 0$ and (28). Thus, $P(|\nu_n(g_n)| > \delta) \leq 2\epsilon$ by (26).

A similar analysis can be done for the second term of (25). The result now follows.

E.5.3 Proof of Lemma E.3

Note that

$$\frac{\sqrt{n}\gamma_n}{c_n}(\hat{F}_{s,n}^{\gamma_n}-\bar{F}_{s,n}^{\gamma_n}) = \frac{\sqrt{n}\gamma_n}{c_n}(\hat{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0})-\frac{\sqrt{n}\gamma_n}{c_n}(\bar{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0}).$$

However, a simplification yields

$$\frac{\sqrt{n}\gamma_n}{c_n}(\hat{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0}) = \frac{1}{c_n}\sqrt{n}(\mathbb{F}_n-F)+\frac{\sqrt{n}M}{c_n r_n\alpha_0}(F-F_b).$$

Since $\sqrt{n}(\mathbb{F}_n-F)/c_n$ is $o_P(1)$, $\sqrt{n} = c_n r_n$, and $F-F_b = \alpha_0(F_s^{\alpha_0}-F_b)$, we have

$$\frac{\sqrt{n}\gamma_n}{c_n M}(\hat{F}_{s,n}^{\gamma_n}-F_s^{\alpha_0}) \xrightarrow{P} F_s^{\alpha_0}-F_b \quad \text{in } \mathbb{H}. \tag{29}$$

By applying the functional delta method (see Theorem 20.8 of Van der Vaart (1998)) for the projection operator (see Theorem 1 of Fils-Villetard et al. (2008)) to (29), we have

$$\frac{\sqrt{n}\gamma_n}{c_n M}(\bar{F}_{s,n}^{\gamma_n} - F_s^{\alpha_0}) \xrightarrow{P} \Pi\left(F_s^{\alpha_0} - F_b \middle| T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}. \tag{30}$$

By combining (29) and (30), we have

$$\frac{\sqrt{n}\gamma_n}{c_n M}(\hat{F}_{s,n}^{\gamma_n} - \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} (F_s^{\alpha_0} - F_b) - \Pi\left(F_s^{\alpha_0} - F_b \middle| T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}. \tag{31}$$

The result now follows by applying the continuous mapping theorem to (31). We prove $V \neq 0$ by contradiction. Suppose that $V = 0$, i.e., $(F_s^{\alpha_0} - F_b) \in T_{\mathcal{F}}(F_s^{\alpha_0})$. Therefore, for some distribution function G and $\eta > 0$, we have $V = (\eta + 1)F_s^{\alpha_0} - F_b - \eta G$, by the definition of $T_{\mathcal{F}}(F_s^{\alpha_0})$. By the discussion leading to (5), it can be easily seen that ηG is a sub-CDF, while $(\eta + 1)F_s^{\alpha_0} - F_b$ is not (as that would contradict (5)). Therefore, $V \neq 0$ and thus $\int V^2 dF > 0$.

§E.6 Proof of Theorem 3.8

The constant c defined in the statement of the theorem can be explicitly expressed as

$$c = -\left\{\int V^2 dF\right\}^{-\frac{1}{2}},$$

where

$$V = (F_s - F_b) - \Pi(F_s - F_b | T_{\mathcal{F}}(F_s)),$$

and Π and $T_{\mathcal{F}}(\cdot)$ are defined in (18) and (19), respectively.

Let $x > 0$. Obviously,

$$P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) \leq x) = 1 - P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > x).$$

By Lemma E.1, we have that $P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) > x) \to 0$ if $c_n \to \infty$. Now let $x \leq 0$. In this case the left hand side of the above display equals $P(\sqrt{n}\gamma_n d_n(\hat{F}_{s,n}^{\gamma_n}, \check{F}_{s,n}^{\gamma_n}) \leq c_n)$, where $\gamma_n = \alpha_0 + x/r_n$. A simplification yields

$$\frac{\sqrt{n}}{c_n}\gamma_n(\hat{F}_{s,n}^{\gamma_n} - F_s^{\alpha_0}) \xrightarrow{P} -x(F_s^{\alpha_0} - F_b), \text{ in } \mathbb{H}, \tag{32}$$

since $\sqrt{n}(\mathbb{F}_n - F)/c_n$ is $o_P(1)$; see the proof of Lemma E.3 (Section E.5.3) for the details. By applying the functional delta method (cf. Theorem 20.8 of Van der Vaart (1998)) for the projection operator (see Theorem 1 of Fils-Villetard et al. (2008)) to (32), we have

$$\frac{\sqrt{n}}{c_n}\gamma_n(\bar{F}_{s,n}^{\gamma_n} - F_s^{\alpha_0}) \xrightarrow{d} \Pi\left(-x(F_s^{\alpha_0} - F_b) \middle| T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}. \tag{33}$$

Adding (32) and (33), we get

$$\frac{\sqrt{n}}{c_n}\gamma_n(\hat{F}_{s,n}^{\gamma_n} - \bar{F}_{s,n}^{\gamma_n}) \to -x(F_s^{\alpha_0} - F_b) - \Pi\left(-x(F_s^{\alpha_0} - F_b)\,|\, T_{\mathcal{F}}(F_s^{\alpha_0})\right) \quad \text{in } \mathbb{H}.$$

By the continuous mapping theorem, we get $\sqrt{n}/c_n\gamma_n d(\hat{F}_{s,n}^{\gamma_n}, \bar{F}_{s,n}^{\gamma_n}) \xrightarrow{P} |x| \left\{\int V^2 dF\right\}^{1/2}$. Hence, by Lemma E.2,

$$P(r_n(\hat{\alpha}_0^{c_n} - \alpha_0) \le x) \to \begin{cases} 1, & \text{if } x > 0, \\ 1, & \text{if } x \le 0 \text{ and } |x| \le \left\{\displaystyle\int V^2 dF\right\}^{-1/2}, \\ 0, & \text{otherwise.} \end{cases}$$

§E.7 Proof of Theorem 4.2

It is enough to show that $\sup_x |H_n(x) - G(x)| \to 0$, where G is the limiting distribution of the Cramér-von Mises statistic, a continuous distribution. As $\sup_x |G_n(x) - G(x)| \to 0$, it is enough to show that

$$\sqrt{n} d_n(\mathbb{F}_n, F) - \sqrt{n} d(\mathbb{F}_n, F) \xrightarrow{P} 0. \tag{34}$$

We now prove (34). Observe that

$$n(d_n^2 - d^2)(\mathbb{F}_n, F) = \sqrt{n}(\mathbb{P}_n - P)[\hat{g}_n] = \nu_n(\hat{g}_n), \tag{35}$$

where $\hat{g}_n = \sqrt{n}(\mathbb{F}_n - F)^2$, $\mathbb{P}_n$ denotes the empirical measure of the data, and $\nu_n := \sqrt{n}(\mathbb{P}_n - P)$ denotes the usual empirical process. We will show that $\nu_n(\hat{g}_n) \xrightarrow{P} 0$, which will prove (35).

For each positive integer n, we introduce the following class of functions

$$\mathcal{G}_c(n) = \left\{\sqrt{n}(H - F)^2 : H \in \mathcal{F} \text{ and } \sup_{t \in \mathbb{R}} |H(t) - F(t)| < \frac{c}{\sqrt{n}}\right\}.$$

Let us also define

$$D_n := \sup_{t \in \mathbb{R}} \sqrt{n} |\mathbb{F}_n(t) - F(t)|.$$

From the definition of $\hat{g}_n$ and D_n^2, we have $\hat{g}_n(t) \le \frac{1}{\sqrt{n}} D_n^2$, for all $t \in \mathbb{R}$. As $D_n = O_P(1)$, for any given $\epsilon > 0$, there exists $c > 0$ (depending on ϵ) such that

$$P(\hat{g}_n \notin \mathcal{G}_c(n)) = P(\sqrt{n} \sup_t |\hat{g}_n(t)| \ge c^2) = P(D_n^2 \ge c^2) \le \epsilon, \tag{36}$$

for all sufficiently large n. Therefore, for any $\delta > 0$, using the same sequence of steps as in (26),

$$P(|\nu_n(\hat{g}_n)| > \delta) \quad \le \quad J_{[\,]} \frac{E[G_c^2(n)]}{\delta} + P(\hat{g}_n \notin \mathcal{G}_c(n)), \tag{37}$$

where $G_c(n) := \frac{c^2}{\sqrt{n}}$ is an envelope for $\mathcal{G}_c(n)$ and $J_{[\,]}$ is a constant. Note that to derive the last inequality we have used the maximal inequality in Corollary (4.3) of Pollard (1989); the class $\mathcal{G}_c(n)$ is "manageable" in the sense of Pollard (1989) (as a consequence of equation (2.5) of Van de Geer (2000)).

Therefore, for any given $\delta > 0$ and $\epsilon > 0$, for large enough n and $c > 0$ we can make both $J_{[\,]}c^4/(\delta n)$ and $P(\hat{g}_n \notin \mathcal{G}_c(n))$ less than ϵ, using (36) and (37), and thus, $P(|\nu_n(\hat{g}_n)| > \delta) \leq 2\epsilon$. The result now follows.

§E.8 Proof of Theorem 4.3

The random variable U defined in the statement of the theorem can be explicitly expressed as

$$U := \left[\int \{\mathbb{G}_F - \Pi(\mathbb{G}_F | T_{\mathcal{F}}(F_s^{\alpha_0})\}^2 \, dF\right]^{1/2},$$

where $\mathbb{G}_F$ is the F-Brownian bridge.

By the same line of arguments as in the proof of Lemma E.2 (see Section E.5.2), it can be easily seen that $\sqrt{n}\alpha_0 \; d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) - \sqrt{n}\alpha_0 \; d(\hat{F}_{s,n}^{\alpha_0}, \bar{F}_{s,n}^{\alpha_0}) \xrightarrow{P} 0$. Moreover, by Donsker's theorem,

$$\sqrt{n}\alpha_0(\hat{F}_{s,n}^{\alpha_0} - F_s^{\alpha_0}) \xrightarrow{d} \mathbb{G}_F.$$

By applying the functional delta method for the projection operator, in conjunction with the continuous mapping theorem to the previous display, we have

$$\sqrt{n}\alpha_0(\bar{F}_{s,n}^{\alpha_0} - F_s^{\alpha_0}) \xrightarrow{d} \Pi(\mathbb{G}_F | T_{\mathcal{F}}(F_s^{\alpha_0})) \quad \text{in} \quad \mathbb{H},$$

where Π, $T_{\mathcal{F}}(\cdot)$, and $F_s^{\alpha_0}$ are defined in (18), (19), and (21), respectively. Hence, by an application of the continuous mapping theorem, we have $\sqrt{n}\alpha_0 d(\hat{F}_{s,n}^{\alpha_0}, \bar{F}_{s,n}^{\alpha_0}) \xrightarrow{d} U$. The result now follows.

§E.9 Proof of Theorem 6.1

The constant c and the function Q defined in the statement of the theorem can be explicitly expressed as

$$c = d(Q, \Pi\left(Q | T_{\mathcal{F}}(F_s)\right)),$$

and

$$Q := (F_s - F_b)\left\{\alpha_0^2 \int V^2 dF\right\}^{-1/2},$$

where

$$r_n = \sqrt{n}/c_n, \quad V = (F_s - F_b) - \Pi(F_s - F_b | T_{\mathcal{F}}(F_s)),$$

and Π and $T_{\mathcal{F}}(\cdot)$ are defined in (18) and (19), respectively.

Recall the notation of Section E.5. Note that from (2),

$$\hat{F}_{s,n}^{\check{\alpha}_n}(x) = \frac{\alpha_0}{\check{\alpha}_n} F_s(x) + \frac{\check{\alpha}_n - \alpha_0}{\check{\alpha}_n} F_b(x) + \frac{(\mathbb{F}_n - F)(x)}{\check{\alpha}_n},$$

for all $x \in \mathbb{R}$. Thus we can bound $\hat{F}_{s,n}^{\check{\alpha}_n}(x)$ as follows:

$$\frac{\alpha_0}{\check{\alpha}_n} F_s(x) - \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} - \frac{D_n'}{\check{\alpha}_n} \leq \hat{F}_{s,n}^{\check{\alpha}_n}(x) \leq \frac{\alpha_0}{\check{\alpha}_n} F_s(x) + \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n},$$

where $D_n' = \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)|$. As both the upper and lower bounds are monotone, we can always find a version of $\check{F}_{s,n}^{\check{\alpha}_n}$ such that

$$\frac{\alpha_0}{\check{\alpha}_n} F_s - \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} - \frac{D_n'}{\check{\alpha}_n} \leq \check{F}_{s,n}^{\check{\alpha}_n} \leq \frac{\alpha_0}{\check{\alpha}_n} F_s + \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n}.$$

Therefore,

$$\begin{aligned} |\check{F}_{s,n}^{\check{\alpha}_n} - F_s| &\leq \frac{|\alpha_0 - \check{\alpha}_n|}{\check{\alpha}_n} F_s + \frac{|\check{\alpha}_n - \alpha_0|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n} \\ &\leq 2\frac{|\alpha_0 - \check{\alpha}_n|}{\check{\alpha}_n} + \frac{D_n'}{\check{\alpha}_n} \xrightarrow{P} 0, \end{aligned}$$

as $n \to \infty$, using the fact $\check{\alpha}_n \xrightarrow{P} \alpha_0 \in (0,1)$. Furthermore, if $q_n(\check{\alpha}_n - \alpha_0) = O_P(1)$, where $q_n/\sqrt{n} \to 0$, it is easy to see that $q_n|\check{F}_{s,n}^{\check{\alpha}_n} - F_s| = O_P(1)$, as $q_n D_n' = o_P(1)$. Note that

$$r_n \hat{\alpha}_0^{c_n} (\hat{F}_{s,n}^{\hat{\alpha}_0^{c_n}} - F_s) = r_n(\mathbb{F}_n - F) + r_n(\alpha_0 - \hat{\alpha}_0^{c_n})\ (F_s - F_b)$$

Thus

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

Hence by an application of functional delta method for the projection operator, in conjunction with the continuous mapping theorem, we have

$$r_n d(\check{F}_{s,n}^{\hat{\alpha}_0^{c_n}}, F_s) \xrightarrow{P} d(Q, \Pi(Q|T_{\mathcal{F}}(F_s))).$$

References

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

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

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

Benjamini, Y., A. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*(3), 491–507.

Bertsekas, D. P. (2003). *Convex analysis and optimization*. Athena Scientific, Belmont, MA. With Angelia Nedić and Asuman E. Ozdaglar.

Black, M. A. (2004). A note on the adaptive control of false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol. 66*(2), 297–304.

Bordes, L., S. Mottelet, and P. Vandekerkhove (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist. 34*(3), 1204–1232.

Cai, T., J. Jin, and M. G. Low (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist. 35*(6), 2421–2449.

Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist. 38*(1), 100–145.

Celisse, A. and S. Robin (2010). A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Planng. Inf. 140*(11), 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 J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist. 32*(3), 962–994.

Efron, B. (2007). Size, power and false discovery rates. *Ann. Statist. 35*(4), 1351–1377.

Efron, B. (2010). *Large-scale inference*, Volume 1 of *Institute of Mathematical Statistics Monographs*. Cambridge: Cambridge University Press. Empirical Bayes methods for estimation, testing, and prediction.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney.

Fils-Villetard, A., A. Guillou, and J. Segers (2008). Projection estimators of Pickands dependence functions. *Canad. J. Statist. 36*(3), 369–382.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *Ann. Statist. 32*(3), 1035–1061.

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

Grotzinger, S. J. and C. Witzgall (1984). Projections onto order simplexes. *Appl. Math. Optim. 12*(3), 247–270.

Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani (2009). *The elements of statistical learning*, Volume 2. Springer.

Hengartner, N. W. and P. B. Stark (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist. 23*(2), 525–550.

Hunter, D. R., S. Wang, and T. P. Hettmansperger (2007). Inference for mixtures of symmetric distributions. *Ann. Statist. 35*(1), 224–251.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B Stat. Methodol. 70*(3), 461–493.

Kulldorff, M., J. Heffernan, R. Hartman, R. Assuncao, and F. Mostashari (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Med. 2*(3), e59.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol. 67*(4), 555–572.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist. 11*(1), 86–94.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics 5*, 1–163.

Lindsay, B. G. and P. Basak (1993). Multivariate normal mixtures: a fast consistent method of moments. *J. Amer. Statist. Assoc. 88*(422), 468–476.

Lyons, L. (2008). Open statistical issues in particle physics. *Ann. Appl. Stat. 2*(3), 887–915.

McLachlan, G. and D. Peel (2000). *Finite mixture models.* Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Meinshausen, N. and P. Bühlmann (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika 92*(4), 893–907.

Meinshausen, N. and J. Rice (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist. 34*(1), 373–393.

Miller, C. J., C. Genovese, R. C. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, and A. Schneider, J.and Moore (2001). Controlling the false-discovery rate in astrophysical data analysis. *Astron. J. 122*(6), 3492–3505.

Nguyen, V. H. and C. Matias (2013). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. arXiv:1205.4097.

Parzen, E. (1960). *Modern probability theory and its applications.* John Wiley & Sons, Incorporated.

Pollard, D. (1989). Asymptotics via empirical processes. *Statist. Sci. 4*(4), 341–366. With comments and a rejoinder by the author.

Quandt, R. E. and J. B. Ramsey (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc. 73*(364), 730–752. With comments and a rejoinder by the authors.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order restricted statistical inference.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons Ltd.

Robin, A. C., C. Reyl, S. Derrire, and S. Picaud (2003). A synthetic view on structure and evolution of the milky way. *Astronomy and Astrophysics 409*(1), 523–540.

Robin, S., A. Bar-Hen, J.-J. Daudin, and L. Pierre (2007). A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput. Statist. Data Anal. 51*(12), 5483–5493.

Salvador, S. and P. Chan (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Proc. 16th IEEE Intl. Conf. on Tools with AI 25*, 576–584.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64*(3), 479–498.

Swanepoel, J. W. H. (1999). The limiting behavior of a modified maximal symmetric $2s$-spacing with applications. *Ann. Statist. 27*(1), 24–35.

Turkheimer, F., C. Smith, and K. Schmidt (2001). Estimation of the number of "true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage 13*(5), 920–930.

Van de Geer, S. A. (2000). *Applications of empirical process theory*, Volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

Walker, M., M. Mateo, E. Olszewski, B. Sen, and M. Woodroofe (2009). Clean kinematic samples in dwarf spheroidals: An algorithm for evaluating membership and estimating distribution parameters when contamination is present. *The Astronomical Journal 137*, 3109.

Walker, M. G., M. Mateo, E. W. Olszewski, O. Y. Gnedin, X. Wang, B. Sen, and M. Woodroofe (2007). Velocity dispersion profiles of seven dwarf spheroidal galaxies. *Astrophysical J. 667*(1), L53–L56.

Walther, G. (2001). Multiscale maximum likelihood analysis of a semiparametric model, with applications. *Ann. Statist. 29*(5), 1297–1319.

Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc. 97*(458), 508–513.