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

Rohit Patra and Bodhisattva Sen
Columbia University, USA, and University of Cambridge, UK

April 1, 2012

Abstract

We consider a two-component mixture model with one known component. We develop methods for estimating the mixing proportion and the other unknown distribution given i.i.d. data from the mixture model. We use ideas from shape restricted function estimation and develop "tuning parameter free" estimators that are easily implementable and have good finite sample performance. We also establish the consistency of our procedures. Honest finite sample lower confidence bounds are developed for the mixing proportion. We also address issues like identifiability of the model and estimation of the density of the unknown mixing distribution. Special emphasis is given to the problem of multiple testing, and our approach is compared to the existing methods in this area through a simulation study. We also analyze two data sets, one arising from an application in astronomy and the other from a microarray experiment.

Keywords: Cramér-von Mises statistic, identifiability, local false discovery rate, lower bound, microarray experiment, shape restricted function estimation.

1 Introduction

We consider a mixture model with two components, i.e.,

$$F(x) = \alpha F_s(x) + (1 - \alpha) F_b(x) \qquad (1.1)$$

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, neuro-imaging) 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., [Efr10, RBHDP07]. Translated to the setting of (1.1), 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 estimate of α is important when we want to assess or control multiple error rates, such as the false discovery rate (FDR) of [BH95]. We discuss this problem in more detail in Section 3.

In contamination problems the distribution F_b, for which reasonable assumptions can be made, maybe contaminated by an arbitrary distribution F_s yielding a sample drawn from F as in (1.1); see e.g., [MP00]. 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 positions of the stars (see [WMO$^+$09]). 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. This problem also arises 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., [Lyo08].

In this paper we provide a methodology to estimate α and F_s (nonparametrically), without assuming any constraint on the form of F_s. We also provide a lower confidence bound for α that is distribution-free (i.e., it does not depend on the particular choice of F_b and F_s). We also propose a nonparametric estimator of f_s, the density of F_s, when f_s is assumed to be non-increasing. Our procedure is completely automated (i.e., tuning parameter free) and easily implementable. We also establish the consistency of the proposed estimators. To the best of our knowledge this is the first attempt to nonparametrically estimate the CDF F_s under no further assumptions.

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 the distributions belong to certain parametric models, which lead to techniques based on maximum likelihood (see e.g., [Coh67] and [Lin83]), minimum chi-square (see e.g., [Day69]), method of moments (see e.g., [LB93]) and moment generating functions (see e.g., [QR78]). Bordes et al. [BMV06] assume that both the components belong to an unknown symmetric location-shift family. In the multiple testing setup, this problem has been addressed by various authors and different estimators and confidence bounds for α have been proposed in the literature under suitable assumptions on F_s and its density, see e.g., [Sto02, GW04, MR06, MB05, CR10, LLF05]. For the sake of brevity, we do not discuss the above references here but come back to this application in Section 3.

The paper is organized as follows. In Section 2 we propose estimators of α and F_s and f_s and investigate their theoretical properties. Connection to the multiple testing problem is developed in Section 3. In Section 4 we compare the finite sample performance of our estimators with other estimators available in the literature through simulation studies. Two real data examples, one arising in astronomy and the other from a microarray experiment, are analyzed in Section 5. We conclude with a brief discussion of our procedure and some open questions in Section 6.

2 Estimation

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 (1.1). If α were known, a naive estimator of F_s would be

$$\hat{F}^{\alpha}_{s,n} = \frac{\mathbb{F}_n - (1-\alpha)F_b}{\alpha}, \tag{2.1}$$

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

$$\int \{W(x) - \hat{F}^{\alpha}_{s,n}(x)\}^2 \, d\mathbb{F}_n(x) \equiv \frac{1}{n}\sum_{i=1}^n \{W(X_i) - \hat{F}^{\alpha}_{s,n}(X_i)\}^2 \tag{2.2}$$

over all DFs W. Let $\check{F}^{\alpha}_{s,n}$ be a DF that minimizes (2.2). The above optimization problem is the same as minimizing $\|\boldsymbol{\theta} - \mathbf{V}\|^2$ over $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \Theta_{inc}$ where

$$\Theta_{inc} = \{\theta \in \mathbb{R}^n : 0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n \leq 1\}, \tag{2.3}$$

$\mathbf{V} = (V_1, V_2, \ldots, V_n)$, $V_i := \hat{F}^{\alpha}_{s,n}(X_{(i)})$, $i = 1, 2, \ldots, n$, $X_{(i)}$ being the i-th order statistic of the sample, and $\|\cdot\|$ denoting 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 in page 88 of [Ber03]); it is the L_2 projection of $\mathbf{V}$ on a closed convex cone in $\mathbb{R}^n$. $\hat{\boldsymbol{\theta}}$ is related to the $\check{F}^{\alpha}_{s,n}$ via $\check{F}^{\alpha}_{s,n}(X_{(i)}) = \hat{\theta}_i$, and can be easily computed using the pool-adjacent-violators algorithm (PAVA); see Section 1.2 of [RWD88]. Thus, $\check{F}^{\alpha}_{s,n}$ is uniquely defined at the data points X_i, for all $i = 1, \ldots, n$, and can be defined on the entire real line by extending it in a piece-wise constant fashion that is right continuous with possible jumps only at the data points. The following result, derived easily from Chapter 1 of [RWD88], characterizes $\check{F}^{\alpha}_{s,n}$.

Lemma 2.1. *Let $\tilde{F}^{\alpha}_{s,n}$ be the isotonic regression (see e.g, page 4 of Chapter 1 of [RWD88]) of the set of points $\{\hat{F}^{\alpha}_{s,n}(X_{(i)})\}_{i=1}^n$. Then $\tilde{F}^{\alpha}_{s,n}$ is characterized as the right-hand slope of the greatest convex minorant of the set of points $\{i/n, \sum_{j=0}^{i} \hat{F}^{\alpha}_{s,n}(X_{(j)})\}_{i=0}^n$. The restriction of $\tilde{F}^{\alpha}_{s,n}$ to $[0,1]$, i.e.,*

$$\check{F}^{\alpha}_{s,n} = \min\{\max\{\tilde{F}^{\alpha}_{s,n}, 0\}, 1\}, \tag{2.4}$$

minimizes (2.2) over all DFs.

Isotonic regression and the PAVA algorithm are very well studied objects in the statistical literature, with many text-book length treatments; see e.g., [RWD88, BBBB72]. If skillfully implemented, PAVA has a computational complexity of $O(n)$ (see [GW84]).

2.2 Identifiability of F_s

When α is *unknown*, the problem is considerably harder; in fact, it is non-identifiable. If (1.1) holds for some F_b and α then the mixture distribution 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 $(\frac{\alpha}{\alpha+\gamma}F_s + \frac{\gamma}{\alpha+\gamma}F_b)$ can be thought of as the nonparametric component. A trivial solution occurs when we take $\alpha + \gamma = 1$, in which case (2.2) is minimized when $W = \mathbb{F}_n$. Hence, α is not uniquely defined. To handle the identifiability issue, we redefine the mixing proportion as

$$\alpha_0 := \inf \left\{ \gamma \in (0,1) : \frac{F - (1-\gamma)F_b}{\gamma} \text{ is a valid DF} \right\}. \tag{2.5}$$

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 (2.5).

Suppose that we start with a fixed F_s, F_b and α. As seen from the above discussion we can only hope to estimate α_0, which, from its definition in (2.5), 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 results provides an answer and is proved in the Appendix.

Lemma 2.2. *Suppose that F_s and F_b are absolutely continuous, i.e., they have densities f_s and f_b, respectively. Then $\alpha_0 < \alpha$ if and only if there exists $c > 0$ such that $f_s(x) \geq c f_b(x)$, for all $x \in \mathbb{R}$.*

The above lemma shows that if there does not exist any $c > 0$ for which $f_s(x) \geq c f_b(x)$, for all $x \in \mathbb{R}$, 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. As in [GW04], we define any distribution G to be *pure* if $\operatorname{essinf}_{t\in\mathbb{R}} g(t) = 0$, where $g(t)$ is the density corresponding to G and $\operatorname{essinf}_{t\in\mathbb{R}} g = \inf\{a \in \mathbb{R} : \mu(\{x : g(x) > a\}) = 0\}$, μ being the Lebesgue measure. They proved that purity of F_s is a sufficient condition for identifiability of the model when F_b is the uniform distribution. This is indeed an easy consequence of the above lemma. A few remarks are in order.

Remark 2.3. *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 $\sigma_s \leq \sigma_b$. Now consider a mixture of exponentials, 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 $a_s \leq a_b$, the problem is identifiable if and only if $\sigma_s \leq \sigma_b$.*

Remark 2.4. *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 for each individual test. 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, since the value of f_θ is exponentially small in m, $\alpha_0 - \alpha$ is very small. In many practical situations, where m is not too small, the difference between α and α_0 is negligible. It should be noted that the problem may actually be identifiable if we have some restrictions on F_s, e.g., if we require F_s to be normal.

2.3 Estimation of the mixing proportion α_0

Note that when $\gamma = 1$, $\hat{F}_{s,n}^\gamma = \mathbb{F}_n = \check{F}_{s,n}^\gamma$ where $\hat{F}_{s,n}^\gamma$ and $\check{F}_{s,n}^\gamma$ are defined in (2.1) and using (2.2), respectively. Whereas, when γ is much smaller than α_0 the regularization of $\hat{F}_{s,n}^\gamma$ modifies it, and thus $\hat{F}_{s,n}^\gamma$ and $\check{F}_{s,n}^\gamma$ are quite different. We would like to compare the naive and isotonized estimators $\hat{F}_{s,n}^\gamma$ and $\check{F}_{s,n}^\gamma$, respectively, and choose the smallest γ for which their distance is still small. This leads to the following estimator of α_0:

$$\hat{\alpha}_n = \inf \left\{ \gamma \in (0,1) : \gamma \; d_n(\hat{F}_{s,n}^\gamma, \check{F}_{s,n}^\gamma) \leq \frac{c_n}{\sqrt{n}} \right\}, \tag{2.6}$$

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

$$d_n(g,h) = \sqrt{\int \{g(x) - h(x)\}^2 \; d\mathbb{F}_n(x)}.$$

It is easy to see that $d_n(\mathbb{F}_n, \gamma \check{F}_{s,n}^\gamma + (1-\gamma)F_b) = \gamma \; d_n(\hat{F}_{s,n}^\gamma, \check{F}_{s,n}^\gamma)$.

The choice of c_n is important, and in the following 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 choices of c_n will lead to lower bounds for α_0.

2.4 Consistency of $\hat{\alpha}_n$

In this section we prove the consistency of $\hat{\alpha}_n$ through a series of elementary results that are proved in the Appendix.

Lemma 2.5. *For* $1 \geq \gamma \geq \alpha_0$*,*

$$\gamma \; d_n(\check{F}^{\gamma}_{s,n}, \hat{F}^{\gamma}_{s,n}) \leq d_n(F, \mathbb{F}_n).$$

Lemma 2.6. *The set*

$$A_n := \left\{ \gamma \in [0,1] : \gamma \; d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \leq \frac{c_n}{\sqrt{n}} \right\}$$

is convex. Thus, $A_n = [\hat{\alpha}_n, 1]$.

Lemma 2.7.

$$d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \overset{a.s.}{\rightarrow} \begin{cases} 0, & \gamma - \alpha_0 \geq 0, \\ > 0, & \gamma - \alpha_0 < 0. \end{cases} \tag{2.7}$$

Theorem 2.1. *If* $c_n/\sqrt{n} \to 0$ *and* $c_n \to \infty$*, then* $\hat{\alpha}_n \overset{P}{\to} \alpha_0$*.*

The above result shows that for a broad range of choices of c_n, our estimation procedure is consistent.

2.5 Lower bound for α_0

Our goal in this sub-section is to construct a finite sample lower bound $\hat{\alpha}_L$ with the property

$$P(\alpha_0 \geq \hat{\alpha}_L) \geq 1 - \beta \tag{2.8}$$

for a specified confidence level $(1-\beta)$ $(0 < \beta < 1)$, that is valid for any n. 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 $\hat{\alpha}_L > 0$.

Theorem 2.2. *Let* H_n *be the CDF of* $\sqrt{n} d_n(\mathbb{F}_n, F)$*. Let* $\hat{\alpha}_L$ *be defined as in* (2.6) *with* c_n *defined as the* $(1-\beta)$*-quantile of* H_n*. Then* (2.8) *holds.*

Note that H_n is distribution-free (i.e., it does not depend on F_s and F_b) 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 2.3.

$$\sup_{x\in\mathbb{R}} |H_n(x) - G_n(x)| \to 0 \ \textit{as}\ n \to \infty. \tag{2.9}$$

Hence, for moderately large n, we can take c_n to be $(1-\beta)$-quantile of G_n or its asymptotic limit, which are readily available (e.g., see [AD52]). Note that the asymptotic 95% quantile of G_n is 0.6792, and is used in our data analysis.

2.6 A tuning parameter free estimator of α_0

Point estimators of α_0 can be developed by choosing particular values for c_n, e.g., in applications we may choose c_n to be the median of the asymptotic limit of H_n. In this sub-section we propose another method to estimate α_0 that is completely automated and has better finite sample performance (see Section 4). We start with a lemma that describes the shape of our criterion function, and will motivate our procedure.

Lemma 2.8. $\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ *is a non-increasing convex function of* γ *in* $(0,1)$.

Writing

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

we see that for $\gamma \geq \alpha_0$, the second term in the RHS is a DF. Thus, for $\gamma \geq \alpha_0$, $\hat{F}^{\gamma}_{s,n}$ is very close to a DF, 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 DF, and thus the distance $\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is appreciably large. Figure 1 shows two typical such plots 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 $U(0,1)$ (setting II). Note that in both the settings we have used $\alpha_0 = 0.1$ and $n = 5000$. We will use these two settings to illustrate our methodology in the rest of this section and also in Section 4.1.

Thus 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 non-increasing segment to the right of α_0 and a steeply non-increasing segment to the left of α_0. 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 slope of the function changes rapidly.

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, a plot similar to Figure 1a would immediately convey to the practitioner

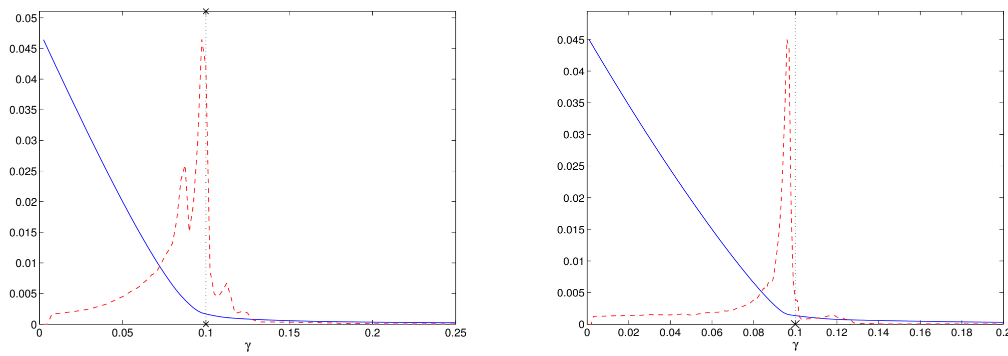


Figure 1: Plot of $\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red) for $n = 5000$ (in solid blue) for setting I (left panel) and setting II (right panel).

the most appropriate choice of $\hat{\alpha}_n$. 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 detech an "elbow" or "knee" of a function is not uncommon; see e.g., [SC04]. In our simulation studies we have used this method to estimate α_0.

2.7 Estimation of F_s

Once we have obtained a consistent estimator $\check{\alpha}_n$ (which may or may not be $\hat{\alpha}_n$ 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 minimizer of (2.2). In the following we show that, indeed, $\check{F}^{\check{\alpha}_n}_{s,n}$ is consistent (in the sup-norm) for estimating F_s.

Theorem 2.4. *Suppose that $\check{\alpha}_n \xrightarrow{P} \alpha_0$. Then, as $n \to \infty$,*

$$\sup_{x \in \mathbb{R}} |\check{F}^{\check{\alpha}_n}_{s,n}(x) - F_s(x)| \xrightarrow{P} 0.$$

An immediate consequence of Theorem 2.4 is that $d_n(\check{F}^{\check{\alpha}_n}_{n,s}, \hat{F}^{\check{\alpha}_n}_{n,s}) \xrightarrow{P} 0$ as $n \to \infty$. Figure 2 shows our estimator $\check{F}^{\check{\alpha}_n}_{s,n}$ along with the true F_s for the same data sets used in Figure 1.

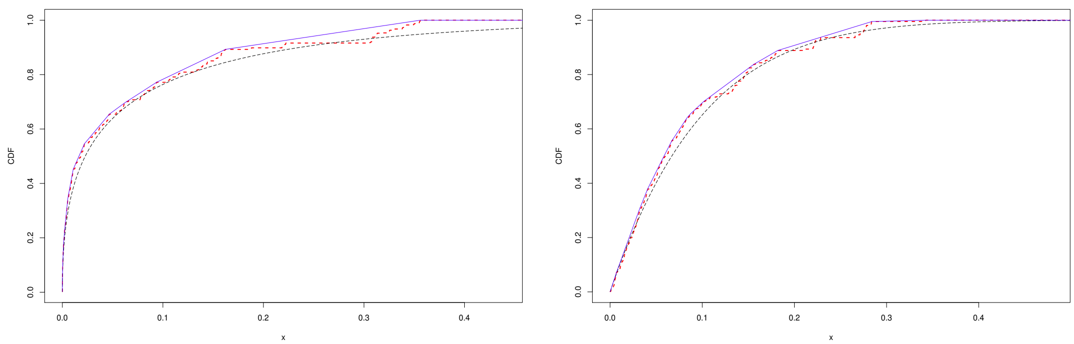


Figure 2: Plot of the estimates $\check{F}_{s,n}^{\check{\alpha}_n}(x)$ (in dotted red), $F_{s,n}^{\dagger}(x)$ (in solid blue) and F_s (in dashed black) for setting I (left panel) and setting II (right panel).

2.8 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, and especially so in our setup, as it requires smoothing and usually involves the choice of tuning parameter(s) (e.g., smoothing bandwidths).

In this section we describe a tuning parameter free approach to estimating f_s, under the additional assumption that f_s is a *non-increasing* density. Without loss of generality, we assume that f_s is non-increasing on $[0, \infty)$. The assumption that f_s is non-increasing, i.e., F_s is concave on its support, is natural in many situations (see Section 3 for an application in the multiple testing problem) and has been investigated by several authors, including [Gre56, WS93, LLF05, GW04].

For a bounded function $g : [0, \infty) \to \mathbb{R}$, let us represent the least concave majorant (LCM) of g by $LCM[g]$. Define $F_{s,n}^{\dagger} := LCM[\check{F}_{s,n}^{\check{\alpha}_n}]$. Note that $F_{s,n}^{\dagger}$ is a valid DF. 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 we show that both $F_{s,n}^{\dagger}$ and $f_{s,n}^{\dagger}$ are consistent estimators of their population versions.

Theorem 2.5. *Assume that $F_s(0) = 0$ and that F_s is a concave on $[0, \infty)$. If $\check{\alpha}_n \xrightarrow{P} \alpha_0$, then, as $n \to \infty$,*

$$\sup_{x \in \mathbb{R}} |F_{s,n}^{\dagger}(x) - F_s(x)| \xrightarrow{P} 0. \qquad (2.10)$$

Further, if for any $x > 0$, $f_s(x)$ is continuous at x, then, as $n \to \infty$,

$$f_{s,n}^{\dagger}(x) \xrightarrow{P} f_s(x).$$

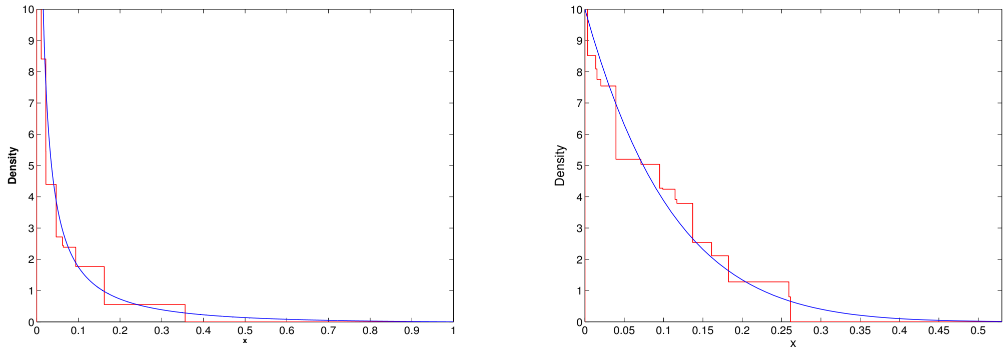


Figure 3: Plot of the estimate $f^{\dagger}_{s,n}(x)$ (in solid red) and f_s (in solid blue) for setting I (left panel) and setting II (right panel).

Computing $F^{\dagger}_{s,n}$ and $f^{\dagger}_{s,n}$ are straightforward, an application of the PAVA gives both the estimators; see e.g., [RWD88, GW84]. Figure 2 shows the LCM $F^{\dagger}_{s,n}$ whereas Figure 3 shows its derivative $f^{\dagger}_{s,n}$ along with the true density f_s for the same data sets as in Figure 1.

Another alternative procedure for estimating F_s and f_s, that will again crucially use estimation under shape constraints, as in (2.1), is provided below, and involves solving an optimization problem. For a fixed α, consider minimizing (2.2) where W is now restricted to the class of all DFs with $F(0) = 0$ that are *concave* on $[0, \infty)$. The new estimator $\bar{F}^{\alpha}_{s,n}$ can be taken as the piece-wise linear concave function such that $\bar{F}^{\alpha}_{s,n}(X_{(i)}) = \hat{\theta}_i$ where

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n) = \arg \min_{\theta \in \Theta \subset \mathbb{R}^n} \|\boldsymbol{\theta} - \mathbf{V}\|^2 \tag{2.11}$$

where $\mathbf{V} = (V_1, V_2, \ldots, V_n)$, $V_i := \hat{F}^{\alpha}_{s,n}(X_{(i)})$, $i = 1, 2, \ldots, n$, $\Theta = \Theta_{inc} \cap \Theta_{con}$, with Θ_{inc} as in (2.3) and

$$\Theta_{con} = \left\{ \theta \in \mathbb{R}^n : \frac{\theta_1}{X_{(1)}} \geq \frac{\theta_2 - \theta_1}{X_{(2)} - X_{(1)}} \geq \frac{\theta_3 - \theta_2}{X_{(3)} - X_{(2)}} \geq \ldots \geq \frac{\theta_n - \theta_{n-1}}{X_{(n)} - X_{(n-1)}} \right\}.$$

The estimator $\hat{\boldsymbol{\theta}}$ is uniquely defined as it is the L_2 projection of $\mathbf{V}$ on a closed convex cone in $\mathbb{R}^n$ and can be easily computed using any standard optimization toolbox (e.g., the cvx package in MATLAB; see http://cvxr.com/cvx/). Note that Θ_{con} guarantees that the fitted $\hat{\boldsymbol{\theta}}$ will be the evaluation of a concave function on $[0, \infty)$.

3 Multiple testing problem

The problem of estimating the proportion, α_0, of true null hypotheses is of interest in situations where a large number of hypotheses tests are performed. Recently, various such situations have arisen in applications. One major motivation is in estimating the proportion of genes that are not 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 (e.g., [TSS01]) and source detection in astrophysics (e.g., [MGN+01]).

Suppose that we wish to test n null hypothesis $H_{01}, H_{02}, \ldots, H_{0n}$ 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. Here $\mathbb{X}$ can be a 6033×102 matrix of expression values in the prostate data example (see Section 5 for more details; also see Section 2.1 of [Efr10]) giving rise to n p-values $X_1, X_2, \ldots, X_n$ and $\mathcal{D}$ might be the rule that rejects H_{0i} if $X_i < 0.001$ and accepts H_{0i} otherwise.

Our estimator of the mixing proportion α_0 can also be used to form the decision rule $\mathcal{D}$. The traditional measure of error in this context is the familywise error rate (FWER). This is defined as FWER = Prob (# of false rejections ≥ 1), the probability of committing at least one type I error. But to control FWER i.e., to guard against any single false positive occurring is often too strict and will lead to many missed findings. In their seminal work [BH95], Benjamini and Hochberg argued that a better quantity to control is the false discovery rate (FDR), defined as the expectation of the proportion of false rejections; more precisely,

$$FDR = E\left\{\frac{V}{R}\mathbf{1}(R > 0)\right\},$$

where V is the number of false rejections and R is the number of total rejections. They also described a method to control FDR, at level β, using the following strategy: reject the null hypotheses corresponding to the (ordered) p-values $p_{(1)}, p_{(2)}, ..., p_{(\hat{k})}$, where $\hat{k} = \max\{k : p_{(k)} \leq \beta k/m\}$. In fact, under identifiability, it can be shown that the above procedure guarantees FDR $\leq \beta\alpha_0$. When α_0 is significantly smaller than 1 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 [BH00]). See [Sto02, Bla04, LLF05] for a discussion on the importance of efficient estimation of

α_0 and some proposed estimators.

Our method can be directly used to yield a consistent estimator of α_0, that does not require the specification of any tuning parameters, as discussed in Section 2.6. Note that to formulate the problem of multiple testing in our setting we would usually take F_b to the uniform distribution and F_s to be the unknown distribution of interest. Our procedure also gives a completely nonparametric estimator of F_s, the distribution of the p-values arising from the alternative hypotheses.

Suppose now 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 applications in the multiple testing setup we will take f_b to be the uniform distribution on $(0, 1)$. For identifiability in this setup, if F_b is taken to be the uniform distribution on $[0, 1]$, we only need to assume that $\inf_{x\in[0,1]} f_s(x) = 0$; see Lemma 2.2. This is indeed the standard assumption made in the literature; see e.g., [NM11].

The *local false discovery rate* (LFDR) is defined as the function $l : (0, 1) \to [0, \infty)$, where

$$l(x) = P\{H_i = 0|X_i = x\} = \frac{(1-\alpha_0)f_b(x)}{f(x)} \tag{3.1}$$

where $f(x) = \alpha_0 f_s(x) + (1-\alpha_0)f_b(x)$ 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., $l(x) \le 0.20$); see Section 5 of [Efr10]. Obtaining good estimates of l can be tricky as it involves the estimation of an unknown density, usually requiring smoothing methods; see Section 5 of [Efr10] for a discussion on estimation and interpretation of l.

We now describe a tuning parameter free approach to estimating the function l, under the additional assumption that f_s is a non-increasing density on $[0, \infty)$. The assumption that the f_s is non-increasing, i.e., F_s is concave, is quite intuitive and natural and has been investigated by several authors, including [GW04, LLF05]. When the alternative hypothesis is true the p-values are generally small and from the discussion in Section 2.8 we have a natural tuning parameter free estimator $\hat{l}$ of the local false discovery rate:

$$\hat{l}(x) = \frac{(1-\hat{\alpha}_n)f_b(x)}{\hat{\alpha}_n f_{s,n}^{\dagger}(x) + (1-\hat{\alpha}_n)f_b(x)}, \tag{3.2}$$

for $x \in (0, 1)$.

4 Simulation

To investigate the finite sample performance of the estimators discussed in this paper we carry out a few simulation experiments. We also compare the performance with other existing methods.

4.1 Lower bound for α_0

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

α	Setting I			Setting II		
	$\hat{\alpha}_L^0$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L^0$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$
0.01	0.973	0.976	0.965	0.975	0.977	0.974
0.03	0.979	0.979	0.976	0.973	0.976	0.972
0.05	0.980	0.980	0.979	0.979	0.977	0.977
0.1	0.991	0.986	0.987	0.987	0.981	0.982

Though there have been some work on the estimating α_0 in the multiple testing setting, [MR06, GW04] are the only papers we found that discuss methodology to construct a lower confidence bound for α_0. These procedures are intellectually connected and the methods in [MR06] are extensions of those proposed by [GW04]. The lower bound $\hat{\alpha}_L$ proposed in both the papers satisfies (2.8) and has the form

$$\hat{\alpha}_L = \sup_{t \in (0,1)} \frac{\mathbb{F}_n(t) - t - \eta_{n,\beta}\delta(t)}{1 - t},$$

where $\eta_{n,\beta}$ is a *bounding sequence* for the *bounding function* $\delta(t)$ at level β (see [MR06]). A *constant* bounding function, $\delta(t) = 1$, is used in [GW04] with $\eta_{n,\beta} = \sqrt{\frac{1}{2n} \log \frac{2}{\beta}}$, whereas [MR06] suggest a class of *bounding functions* but observe that *standard deviation-proportional* bounding function $\delta(t) = \sqrt{t(1-t)}$ has optimal properties among a large calls of possible bounding functions. We have used this bounding function and a bounding sequence suggested by the authors. Note that to use these methods we have to choose the tuning parameters $\delta(t)$ and $\eta_{n,\beta}$ whereas the procedure suggested in Section 2.5 is completely automated. We denote the lower bound proposed by [MR06] as $\hat{\alpha}_L^{MR}$, the bound by [GW04] as $\hat{\alpha}_L^{GW}$ and the lower bound in Section 2.5 by $\hat{\alpha}_L^0$.

We take $\alpha \in \{0.01, 0.03, 0.05, 0.10\}$ and compare the performance of the three lower bounds in two different simulation settings mentioned in Section 2.6. For each setting we have used a sample size (n) of 1000 and 5000. We present the estimated coverage probabilities, obtained by averaging over 5000 simulations, of the lower bounds for the different settings. The results are summarized in Tables 1 and 2.

Table 2: Coverage probabilities of nominal 95% lower confidence bounds for the three methods when $n = 5000$.

α	Setting I			Setting II		
	$\hat{\alpha}_L^0$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L^0$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$
0.01	0.977	0.978	0.973	0.977	0.976	0.970
0.03	0.988	0.985	0.981	0.980	0.979	0.979
0.05	0.987	0.983	0.983	0.984	0.982	0.980
0.1	0.994	0.990	0.988	0.990	0.983	0.986

4.2 Performance of the estimate of α_0

There has been quite a bit of work on the estimation of α_0, as discussed in the Introduction. Some of them use shape constraint on F_s (see e.g., [LLF05]), while others do not assume any constraint (see e.g.,[MR06]). In this sub-section we compare the performance of our estimator, proposed in Section 2.6, with four other estimators available in the literature. Storey [Sto02] proposed an estimate of α_0 which we denote by $\hat{\alpha}_0^{st}$. Due to space constraints we do not discuss the estimation procedure of [Sto02], but we would like to mention that he uses bootstrapping to choose the tuning parameter involved. [LLF05] proposed an estimator which is tuning parameter free but crucially uses the known shape restriction on f_s (convex and non-increasing); we denote it by $\hat{\alpha}_0^L$. We also use the estimator proposed in [MR06] for two bounding functions ($\delta(t) = \sqrt{t(1-t)}$ and $\delta(t) = 1$). For its implementation we have to choose a sequence $\{\beta_n\}$ going to zero as $n \to \infty$. In their paper [MR06] were not specific about the choice of $\{\beta_n\}$ but required the sequence to satisfy some conditions. We choose $\beta_n = \beta/\sqrt{n}$, where $\beta = 0.05$. We denote the estimator proposed by [MR06] by $\hat{\alpha}_0^{MR}$ when $\delta(t) = \sqrt{t(1-t)}$ and by $\hat{\alpha}_0^{GW}$ when $\delta(t) = 1$. We denote our estimator by $\hat{\alpha}_0^0$.

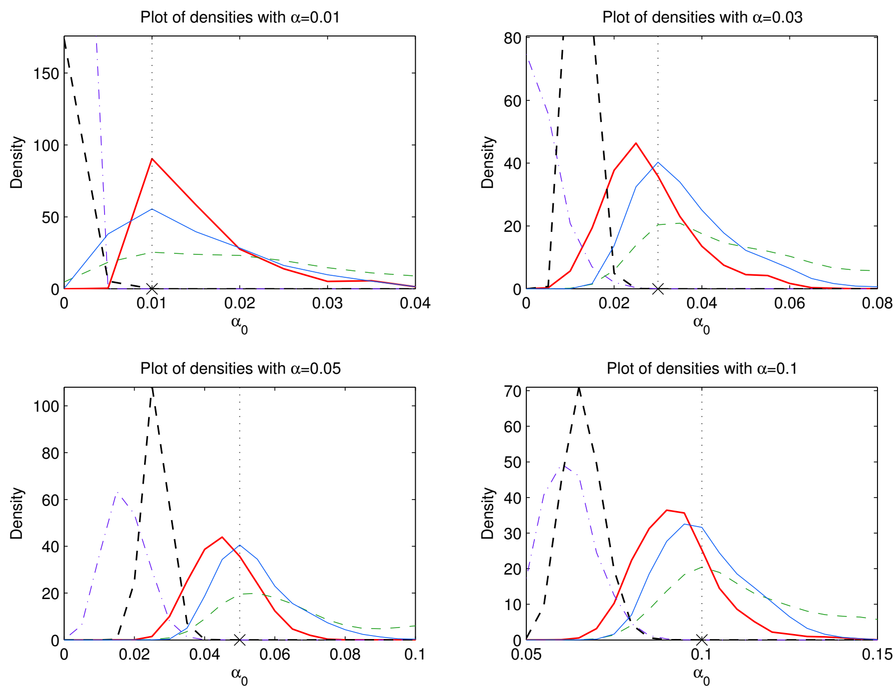


Figure 4: Density plots of the estimators of α_0 ($\alpha \in \{.01, .03, .05, .10\}$): $\hat{\alpha}_0^0$ (in solid red), $\hat{\alpha}_0^{GW}$ (in dash-dotted purple), $\hat{\alpha}_0^{MR}$ (in dashed black), $\hat{\alpha}_0^{st}$ (in dashed green) and $\hat{\alpha}_0^{L}$ (in solid blue). The vertical line (in doted black) denotes the true mixing proportion.

We have used the simulation setting used in [LLF05]. A total of $n = 5000$ features were simulated for each $J = 10$ samples. Let these random variables be denoted by X_{ij} , $i = 1, ..., n$, $j = 1, ..., J$, and the corresponding realizations x_{ij}. Let $X_j = (X_{1j}, X_{2j}, ..., X_{nj})$, and assume that each $X_j \sim N(\mu_{n\times 1}, I_{n\times n})$, and that $X_1, X_2, ..., X_J$ are independent. We test $H_{0i} : \mu_i = 0$ vs $H_{0i} : \mu_i \neq 0$ for each i, and calculate a two-sided p-value p_i based on a one-sample t-test using $p_i = 2P(T_{J-1} \geq |\bar{x}_i/\sqrt{s_i/J}|)$. Here $\bar{x}_i = \sum_{j=1}^{J} x_{ij}/J$ and $s_i = \sum_{j=1}^{J}(x_{ij} - \bar{x}_i)^2/(J-1)$ are the sample mean and variance, respectively, and T_{J-1} is a random variable having the t-distribution with $J - 1$ degrees of freedom.

As before, four different choices of α are considered, namely $0.01, 0.03, 0.05, 0.10$. The μ_i's were set to zero for the true null hypotheses, whereas for the false null hypotheses they were drawn from symmetric bi-triangular density with parameters $a = \log_2(1.2) = 0.263$ and $b = \log_2(4) = 2$ (see page 568 of [LLF05]). We drew $N = 5000$ sets of independent 5000-dimensional vectors from the multivariate

Gaussian distribution $N(\mu_{n\times 1}, I_{n\times n})$, and calculated the corresponding 5000 sets of vectors of p-values.

The mixing proportion α_0 is estimated, using the five different estimates described above, for each set of p-values, and the empirical kernel density of the estimates are shown in Figure 4, for the different choice of α_0. In Table 3 we give the average of the 5000 estimates of the mixing proportion for the five methods along with their root mean squared errors (RMSE). It is clearly evident that our procedure has the least RMSE and has the least bias.

Table 3: Average and RMSE of the five estimators discussed in Section 4.2.

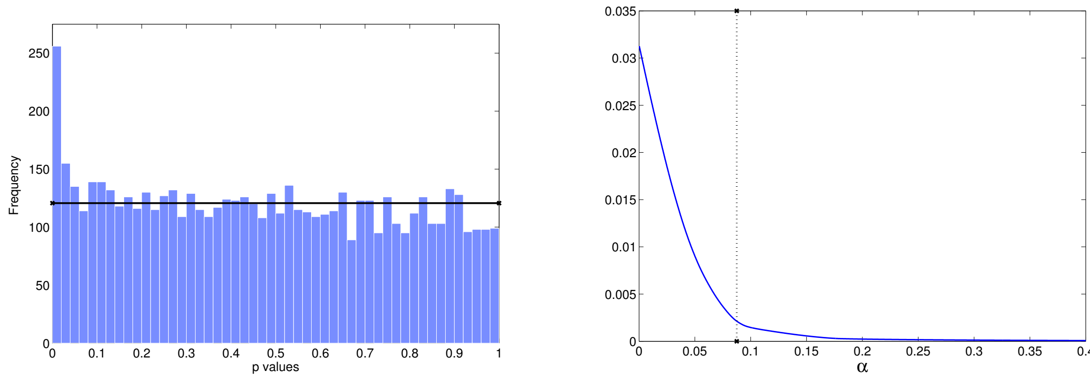
α	Average of the estimators					RMSE of the estimators				
	$\hat{\alpha}_0^0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{st}$	$\hat{\alpha}_0^{L}$	$\hat{\alpha}_0^0$	$\hat{\alpha}_0^{GW}$	$\hat{\alpha}_0^{MR}$	$\hat{\alpha}_0^{st}$	$\hat{\alpha}_0^{L}$
0.01	0.013	0.000	0.001	0.033	0.015	0.008	0.010	0.009	0.036	0.010
0.03	0.028	0.002	0.012	0.061	0.037	0.010	0.028	0.018	0.050	0.014
0.05	0.046	0.017	0.026	0.079	0.055	0.010	0.033	0.024	0.045	0.012
0.10	0.093	0.062	0.066	0.121	0.101	0.014	0.039	0.034	0.038	0.013

5 Real data analysis

5.1 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 biological details, 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, $i = 1, 2, .., n$, and $j = 1, 2, .., m$; with $j = 1, 2, .., 50$, for the normal controls and $j = 51, 52, ..., 102$, 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 for gene i. The two-sample t-statistic for testing significance of gene i is

$$t_i = \frac{\bar{x}_i(1) - \bar{x}_i(2)}{s_i},$$



(a) Histogram of the p-values. (b) Plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue).

Figure 5: The horizontal line (in solid black) in the left panel indicates the $U(0,1)$ distribution. The vertical line (in dotted black) in the right panel indicates the point of maximum curvature ($\hat{\alpha}^0_L$).

where s_i is an estimate of the standard error of $\bar{x}_i(1) - \bar{x}_i(2)$, i.e.,

$$s_i^2 = \frac{\sum_1^{50}\{x_{ij} - \bar{x}_i(1)\}^2 + \sum_{51}^{102}\{x_{ij} - \bar{x}_i(2)\}^2}{100}\left(\frac{1}{50} + \frac{1}{52}\right).$$

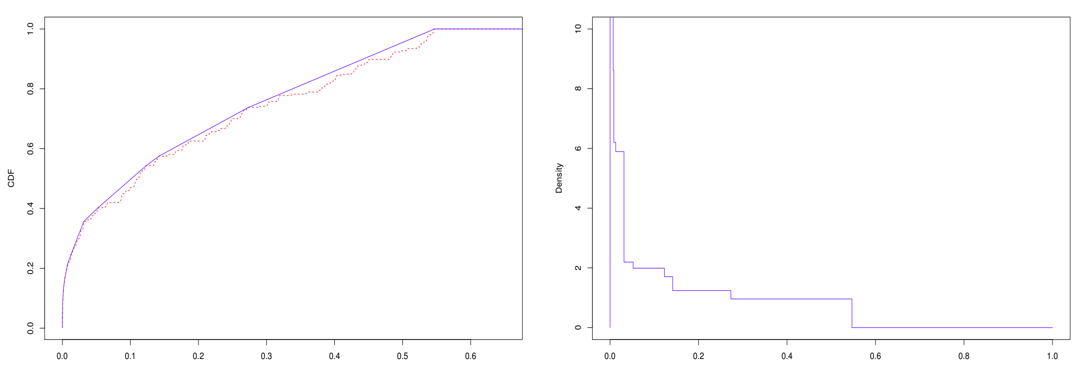


Figure 6: The left panel shows the estimates $\check{F}^{\check{\alpha}_n}_{s,n}(x)$ (in dotted red) and $F^{\dagger}_{s,n}(x)$ (in solid blue). The right panel shows the density $f^{\dagger}_{s,n}$.

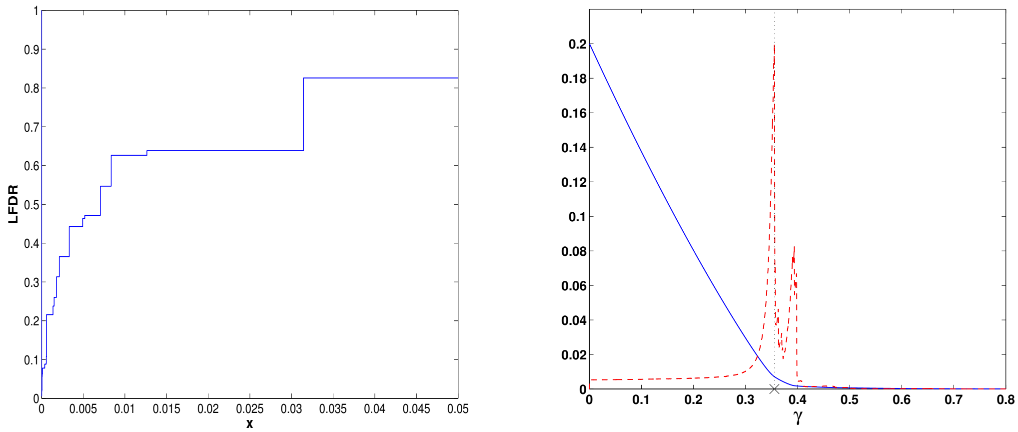
If we had only data from gene i to consider, we could use t_i in the usual way to test the null hypothesis H_{0i}: gene i has no effect, i.e., x_{ij} has the same distribution for the normal and cancer patients; rejecting H_{0i} if t_i looked too big in absolute value. The usual 5% rejection criterion, based on normal theory assumptions, would reject H_{0i} if $|t_i|$ exceeded 1.98, the two-tailed 5% point for a Student-t random variable with 100 degrees of freedom.

We will work with the p-values instead of the "t-values" as then the distribution under the alternative will have a non-increasing density which we can estimate using results from Section 2.8. We have plotted the histogram of the p-values in Figure 5a. Figure 5b shows the plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$, as γ varies from 0 to 1, along with our estimator $\hat{\alpha}_0^0$, which turns out be 0.0877. The lower bound $\hat{\alpha}_L^0$ for this data is found to be 0.0512. The other estimates perform similarly except the one proposed by [LLF05], which does not detect any "signal". In Figure 6 we plot the estimate of the distribution of the p-values under the alternative $\check{F}^{\check{\alpha}_n}_{s,n}(x)$, and its LCM $F^{\dagger}_{s,n}(x)$, along with the estimate of the density f_s, found using theory developed in Section 2.8.

5.2 An Astronomy Example

In this sub-section we analyze the radial velocity (RV) distribution of stars in Carina, a dwarf spheroidal (dSph) galaxy. The dSph galaxies are low luminosity galaxies that are companions of the Milky Way. The data have been obtained by Magellan and MMT telescopes (see [WOG$^+$07]) and consist of radial (line of sight) velocity measurements for $n = 1215$ stars from Carina, contaminated with Milky Way stars in the field of view. We would like to understand the distribution of the line of sight velocity. For the contaminating stars from the Milky Way, we assume a non-Gaussian velocity distribution F_b that we estimate from the Besancon Milky Way model ([RRDP03]), calculated along the line of sight to Carina.

Our estimator for α_0 for this data set turns out to be 0.356, while the lower bound for α_0 is found to be 0.322. Figure 7b shows the plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ along with the estimated α_0. The left panel of Figure 7b shows the estimate of F_s and the closest (in terms of minimizing the $L_2(\check{F}^{\hat{\alpha}_n}_{s,n})$ distance) Gaussian distribution. Astronomers usually assume that the distribution of the radial velocities for these dSph galaxies is Gaussian in nature. Indeed we see that the estimated F_s is close to a normal distribution (with mean 222.9 and variance 7.51), although a formal test of this hypothesis is beyond the scope of the present paper. The right panel of 8 shows the density of the original data and the known f_b, obtained from the Besancon Milky Way model.



(a) Plot of the estimated LFDR $\hat{l}$.

(b) Plot of $\gamma\ d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$.

Figure 7: The left panel shows the plot of the estimate LFDR for p-values less than 0.05 for the prostate data. The right panel plot of $\gamma\ d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ (in solid blue) overlaid with its (scaled) second derivative (in dashed red).

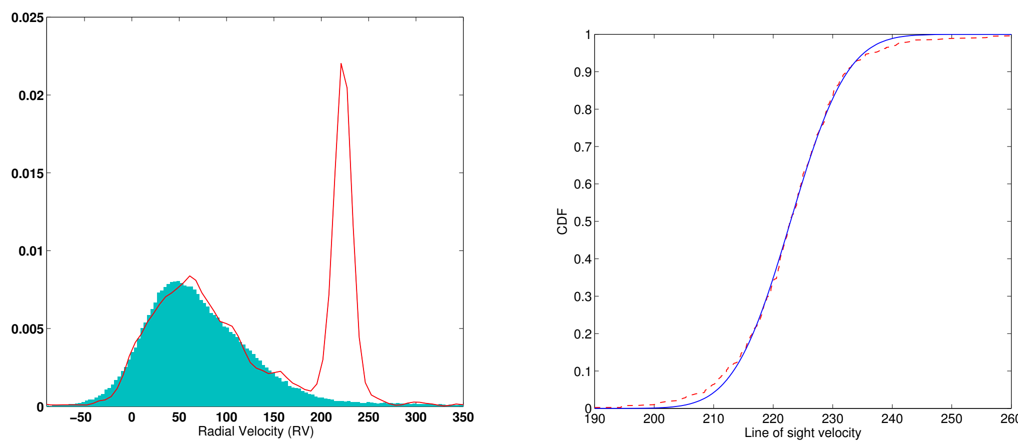


Figure 8: In the left panel we have the histogram of the foreground radial velocity of stars overlaid with the (scaled) kernel density estimator of the Carina dataset. The right panel shows the nonparametric estimator $\check{F}_{s,n}^{\hat{\alpha}_n}$ (in dashed red) overlaid with the closest Gaussian distribution (in solid blue).

6 Conclusion

In this paper we have developed procedure for estimating the mixing proportion and the unknown distribution in a two component mixture model using ideas from shape restricted statistical inference. Our procedures have good finite sample performance and is completely tuning parameter free.

It should be noted that although the methods developed in [GW04, MR06, LLF05] for estimating α_0 and f_s under the multiple testing setting, can, in fact, be generalized to handle situations where F_b is not the uniform distribution by transforming the observed X_i's to $Y_i := F_b^{-1}(X_i)$; the "background" distribution of Y_i becomes uniform on $(0, 1)$. However, apart from the fact that the methods developed in this paper use different techniques and have better finite sample performance, the main advantage of our procedures is that we do not have to choose any tuning parameters in the implementation.

We have established the consistency properties of the estimators developed in the paper. However nothing is presently known about the rates of convergence of $\hat{\alpha}_n$ and the estimators of F_s. Construction of confidence intervals for α_0 can be carried out if we can find the limiting distribution of $\hat{\alpha}_n$. It must be mentioned here that investigating such asymptotic properties of these estimators is expected to be a hard exercise.

As we have observed in the astronomy application, goodness-of-fit tests for F_s are important as it can help the practitioner to use appropriate parametric models for further modelling and study.

A Appendix 1

A.1 Proof of Lemma 2.2

Proof. Suppose that $\alpha_0 < \alpha$. Then there exists $\alpha^* \in (\alpha_0, \alpha)$ such that $[F - (1 - \alpha^*)F_b]/\alpha^*$ is a valid DF. Using the fact that $F = \alpha F_s + (1 - \alpha)F_b$ and letting $\eta := \alpha/\alpha^* > 1$, we see that $F_\eta := \eta F_s - (\eta - 1)F_b$ must be a valid DF. For F_η to be non-decreasing, we must have $\eta f_s(x) - (\eta - 1)f_b(x) \geq 0$ for all $x \in \mathbb{R}$. This implies that we must have $f_s(x) \geq (1 - 1/\eta)f_b(x)$ for all $x \in \mathbb{R}$, which completes the argument. Retracing the steps backwards we can see that if for some $c > 0$ (which necessarily has to be less than 1) $f_s(x) \geq c f_b(x)$, for all $x \in \mathbb{R}$, then there exists $\alpha^* := \alpha(1 - c)$ for which $[F - (1 - \alpha^*)F_b]/\alpha^*$ is a valid DF. Now, from the definition of α_0, it follows that $\alpha_0 < \alpha^* < \alpha$. □

A.2 Proof of Lemma 2.5

Proof. Letting

$$F_s^\gamma = \frac{F - (1-\gamma)F_b}{\gamma}, \tag{A.1}$$

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 DF for $\gamma \geq \alpha_0$. As $\check{F}_{s,n}^\gamma$ is defined as the function that minimizes the $L_2(\mathbb{F}_n)$ distance of $\hat{F}_{s,n}^\gamma$ over all DFs,

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

□

A.3 Proof of Lemma 2.6

Proof. Assume that $\gamma_1 \leq \gamma_2$ and $\gamma_1, \gamma_2 \in A_n$. If

$$\gamma_3 = \eta\gamma_1 + (1-\eta)\gamma_2$$

for $0 \leq \eta \leq 1$, it is easy to observe from (2.1) that

$$\eta(\gamma_1 \hat{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \hat{F}_{s,n}^{\gamma_2}) = \gamma_3 \hat{F}_{s,n}^{\gamma_3}. \tag{A.2}$$

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 DF, and thus from the definition of $\check{F}_{s,n}^{\gamma_3}$, we have

$$\begin{aligned} d_n(\hat{F}_{s,n}^{\gamma_3}, \check{F}_{s,n}^{\gamma_3}) &\leq d_n\left(\hat{F}_{s,n}^{\gamma_3}, \frac{\eta(\gamma_1 \check{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \check{F}_{s,n}^{\gamma_2})}{\gamma_3}\right) \\ &= d_n\left(\frac{\eta(\gamma_1 \hat{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \hat{F}_{s,n}^{\gamma_2})}{\gamma_3}, \frac{\eta(\gamma_1 \check{F}_{s,n}^{\gamma_1}) + (1-\eta)(\gamma_2 \check{F}_{s,n}^{\gamma_2})}{\gamma_3}\right) \\ &\leq \frac{\eta\gamma_1}{\gamma_3} d_n(\hat{F}_{s,n}^{\gamma_1}, \check{F}_{s,n}^{\gamma_1}) + \frac{(1-\eta)\gamma_2}{\gamma_3} d_n(\hat{F}_{s,n}^{\gamma_2}, \check{F}_{s,n}^{\gamma_2}). \end{aligned} \tag{A.3}$$

But as $\gamma_1, \gamma_2 \in A_n$, the above inequality yields

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

Thus $\alpha_3 \in A_n$. □

A.4 Proof of Lemma 2.7

Proof. For $\gamma \geq \alpha_0$ the result follows from Lemma 2.5 and the fact that $d_n(F, \mathbb{F}_n) \overset{a.s.}{\to} 0$ as $n \to \infty$.

For $\gamma < \alpha_0$, F_s^γ is not a valid DF, by the definition of α_0. And as $n \to \infty$, $\hat{F}_{s,n}^\gamma \overset{a.s.}{\to} F_s^\gamma$ point-wise. So for large enough n, $\hat{F}_{s,n}^\gamma$ is not a valid DF, whereas $\check{F}_{s,n}^\gamma$ is always a DF. Thus, $d_n(\hat{F}_{s,n}^\gamma, \check{F}_{s,n}^\gamma)$ converges to something positive. □

A.5 Proof of Theorem 2.1

Proof. We need to show that $P(|\hat{\alpha}_n - \alpha_0| > \epsilon) \to 0$ for any $\epsilon > 0$. So let us first show that

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

Suppose $\hat{\alpha}_n - \alpha_0 < -\epsilon$, i.e., $\hat{\alpha}_n < \alpha_0 - \epsilon$. Then by the definition of $\hat{\alpha}_n$ 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 $\hat{\alpha}_n(< \alpha_0 - \epsilon) \in A_n$), and thus

$$d_n(\hat{F}_{s,n}^{\alpha_0-\epsilon}, \check{F}_{s,n}^{\alpha_0-\epsilon}) \leq \frac{c_n}{\sqrt{n}(\alpha_0 - \epsilon)}. \tag{A.4}$$

But by (2.7) the L.H.S. of (A.4) goes to a non-zero constant in probability. Hence, if $\frac{c_n}{\sqrt{n}} \to 0$,

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

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

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

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

The first implication follows from definition of $\hat{\alpha}_n$, while the second implication is true by Lemma 2.5. The R.H.S. of the last inequality is (asymptotically similar to) the Cramér–von Mises statistic for which the asymptotic distribution is well-known and thus if $c_n \to \infty$ then the result follows. □

A.6 Proof of Theorem 2.2

Proof. Note that

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

where we have used the fact that $d_n(\hat{F}_{s,n}^{\alpha_0}, F_s^{\alpha_0}) = d_n(\mathbb{F}_n, F)/\alpha_0$. □

A.7 Proof of Theorem 2.3

Proof. 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

$$\sqrt{n} d_n(\mathbb{F}_n, F) - \sqrt{n} d(\mathbb{F}_n, F) \stackrel{P}{\to} 0. \tag{A.5}$$

We now prove (A.5). Observe that

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

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 $\nu_n(\hat{g}_n) \stackrel{P}{\to} 0$, which will prove (A.6).

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

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

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 $\hat{g}_n(t) \leq \frac{1}{\sqrt{n}} D_n^2$, 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

$$P\{\hat{g}_n \notin \mathcal{G}_c(n)\} = P\{\sqrt{n} \sup_t |\hat{g}_n(t)| \geq c^2\} = P(D_n^2 \geq c^2) \leq \epsilon, \tag{A.6}$$

for all sufficiently large n. Therefore, for any $\delta > 0$,

$$\begin{aligned}
P\{|\nu_n(\hat{g}_n)| > \delta\} &= P\{|\nu_n(\hat{g}_n)| > \delta, \hat{g}_n \in \mathcal{G}_c(n)\} + P\{|\nu_n(\hat{g}_n)| > \delta, \hat{g}_n \notin \mathcal{G}_c(n)\} \\
&\leq P\{|\nu_n(\hat{g}_n)| > \delta, \hat{g}_n \in \mathcal{G}_c(n)\} + P\{\hat{g}_n \notin \mathcal{G}_c(n)\} \\
&\leq P\left\{\sup_{g\in\mathcal{G}_c(n)} |\nu_n(g)| > \delta\right\} + P\{\hat{g}_n \notin \mathcal{G}_c(n)\} \\
&\leq \frac{1}{\delta} E\left\{\sup_{g\in\mathcal{G}_c(n)} |\nu_n(\hat{g}_n)|\right\} + P\{\hat{g}_n \notin \mathcal{G}_c(n)\} \\
&\leq J\frac{P[G_c^2(n)]}{\delta} + P\{\hat{g}_n \notin \mathcal{G}_c(n)\}, \qquad \text{(A.7)}
\end{aligned}$$

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 [Pol89] (as a consequence of Eq. (2.5) of [vdG00]).

Therefore, for any given $\delta > 0$ and $\epsilon > 0$, for large enough n and $c > 0$ we can make both $Jc^4/(\delta n)$ and $P\{\hat{g}_n \notin \mathcal{G}_c(n)\}$ less than ϵ, using (A.6) and (A.7), in thus, $P\{|\nu_n(\hat{g}_n)| > \delta\} \leq 2\epsilon$. The result now follows. □

A.8 Proof of Lemma 2.8

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

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

which shows that $\gamma\ d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ is a non-increasing function.

To show that $\gamma\ d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ is convex, let $0 < \gamma_1 < \gamma_2 < 1$ and $\gamma_3 = \eta\gamma_1 + (1-\eta)\gamma_2$, for $0 \leq \eta \leq 1$. Then, by (A.3) we have the desired result. □

A.9 Proof of Theorem 2.4

Proof. Note that from (2.1),

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

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

$$\frac{\alpha_0}{\breve{\alpha}_n}F_s(x) - \frac{|\breve{\alpha}_n - \alpha_0|}{\breve{\alpha}_n} - \frac{D_n}{\breve{\alpha}_n} \leq \hat{F}_{n,s}^{\breve{\alpha}_n}(x) \leq \frac{\alpha_0}{\breve{\alpha}_n}F_s(x) + \frac{|\breve{\alpha}_n - \alpha_0|}{\breve{\alpha}_n} + \frac{D_n}{\breve{\alpha}_n}, \qquad \text{(A.8)}$$

where $D_n = \sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)|$, and both the upper and lower bounds are non-decreasing functions in x. Thus, from the characterization of $\check{F}_{s,n}^{\breve{\alpha}_n}$ and properties of isotonic estimators (see e.g., Theorem 1.3.4 of [RWD88]), we know that for all $i = 1, 2, \ldots, n$,

$$\frac{\alpha_0}{\breve{\alpha}_n}F_s(X_i) - \frac{|\breve{\alpha}_n - \alpha_0|}{\breve{\alpha}_n} - \frac{D_n}{\breve{\alpha}_n} \leq \check{F}_{n,s}^{\breve{\alpha}_n}(X_i) \leq \frac{\alpha_0}{\breve{\alpha}_n}F_s(X_i) + \frac{|\breve{\alpha}_n - \alpha_0|}{\breve{\alpha}_n} + \frac{D_n}{\breve{\alpha}_n}. \qquad \text{(A.9)}$$

Therefore, for all $i = 1, 2, \ldots, n$,

$$\begin{aligned} |\check{F}_{n,s}^{\breve{\alpha}_n}(X_i) - F_s(X_i)| &\leq \frac{|\alpha_0 - \breve{\alpha}_n|}{\breve{\alpha}_n}F_s(X_i) + \frac{|\breve{\alpha}_n - \alpha_0|}{\breve{\alpha}_n} + \frac{D_n}{\breve{\alpha}_n} \\ &\leq 2\frac{|\alpha_0 - \breve{\alpha}_n|}{\breve{\alpha}_n} + \frac{D_n}{\breve{\alpha}_n} \stackrel{P}{\to} 0, \end{aligned}$$

as $n \to \infty$, using the fact $\breve{\alpha}_n \stackrel{P}{\to} \alpha_0 \in (0, 1)$. As the X_is are dense in the support of F, we have the desired result. □

A.10 Proof of Theorem 2.5

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

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

and therefore,

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

By Theorem 2.4, as $\epsilon_n \stackrel{P}{\to} 0$, we must also have (2.10).

The second part of the result follows immediately from the lemma is page 330 of [RWD88], and is similar to the result in Theorem 7.2.2 of that book. □

References

[AD52] T. W. Anderson and D. A. Darling, *Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes*, Ann. Math. Statistics **23** (1952), 193–212. MR 0050238 (14,298h)

[BBBB72] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical inference under order restrictions. The theory and application of isotonic regression*, John Wiley & Sons, London-New York-Sydney, 1972, Wiley Series in Probability and Mathematical Statistics. MR 0326887 (48 #5229)

[Ber03] Dimitri P. Bertsekas, *Convex analysis and optimization*, Athena Scientific, Belmont, MA, 2003, With Angelia Nedić and Asuman E. Ozdaglar. MR 2184037 (2006j:90001)

[BH95] Yoav Benjamini and Yosef Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. Roy. Statist. Soc. Ser. B **57** (1995), no. 1, 289–300. MR 1325392 (96d:62143)

[BH00] ———, *On the adaptive control of the false discovery rate in multiple testing with independent statistics.*, J. Educational and Behavioral Statistics **25** (2000), 60–83.

[Bla04] M. A. Black, *A note on the adaptive control of false discovery rates*, J. R. Stat. Soc. Ser. B Stat. Methodol. **66** (2004), no. 2, 297–304. MR 2062377

[BMV06] Laurent Bordes, Stéphane Mottelet, and Pierre Vandekerkhove, *Semiparametric estimation of a two-component mixture model*, Ann. Statist. **34** (2006), no. 3, 1204–1232. MR 2278356 (2008e:62064)

[Coh67] A. Clifford Cohen, *Estimation in mixtures of two normal distributions*, Technometrics **9** (1967), 15–28. MR 0216626 (35 #7455)

[CR10] Alain Celisse and Stéphane Robin, *A cross-validation based estimation of the proportion of true null hypotheses*, J. Statist. Plann. Inference **140** (2010), no. 11, 3132–3147. MR 2659843 (2011h:62131)

[Day69] N. E. Day, *Estimating the components of a mixture of normal distributions*, Biometrika **56** (1969), 463–474. MR 0254956 (40 #8163)

[Efr10] Bradley Efron, *Large-scale inference*, Institute of Mathematical Statistics Monographs, vol. 1, Cambridge University Press, Cambridge, 2010, Empirical Bayes methods for estimation, testing, and prediction. MR 2724758 (2012a:62006)

[Gre56] Ulf Grenander, *On the theory of mortality measurement. I*, Skand. Aktuarietidskr. **39** (1956), 70–96. MR 0086459 (19,188d)

[GW84] S. J. Grotzinger and C. Witzgall, *Projections onto order simplexes*, Appl. Math. Optim. **12** (1984), no. 3, 247–270. MR 768632 (86d:90128)

[GW04] Christopher Genovese and Larry Wasserman, *A stochastic process approach to false discovery control*, Ann. Statist. **32** (2004), no. 3, 1035–1061. MR 2065197 (2005k:62149)

[LB93] Bruce G. Lindsay and Prasanta Basak, *Multivariate normal mixtures: a fast consistent method of moments*, J. Amer. Statist. Assoc. **88** (1993), no. 422, 468–476. MR 1224371 (94a:62045)

[Lin83] Bruce G. Lindsay, *The geometry of mixture likelihoods: a general theory*, Ann. Statist. **11** (1983), no. 1, 86–94. MR 684866 (85m:62008a)

[LLF05] Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad, *Estimating the proportion of true null hypotheses, with application to DNA microarray data*, J. R. Stat. Soc. Ser. B Stat. Methodol. **67** (2005), no. 4, 555–572. MR 2168204

[Lyo08] Louis Lyons, *Open statistical issues in particle physics*, Ann. Appl. Stat. **2** (2008), no. 3, 887–915. MR 2516798

[MB05] Nicolai Meinshausen and Peter Bühlmann, *Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures*, Biometrika **92** (2005), no. 4, 893–907. MR 2234193

[MGN+01] C. J. Miller, C. Genovese, R. C. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, and A. Schneider, J.and Moore, *Controlling the false-discovery rate in astrophysical data analysis*, Astron. J. **122** (2001), no. 6, 3492–3505.

[MP00] Geoffrey McLachlan and David Peel, *Finite mixture models*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, 2000. MR 1789474 (2002b:62025)

[MR06] Nicolai Meinshausen and John Rice, *Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses*, Ann. Statist. **34** (2006), no. 1, 373–393. MR 2275246 (2008h:62145)

[NM11] Van H. Nguyen and Catherine Matias, *Is it possible to construct an asymptotically efficient estimator of the proportion of true null hypotheses in a multiple testing setup?*, Tech. report, Statistics for Systems Biology Group, Jouy-en-Josas/Paris/Evry, France, December 2011.

[Pol89] David Pollard, *Asymptotics via empirical processes*, Statist. Sci. **4** (1989), no. 4, 341–366, With comments and a rejoinder by the author. MR 1041762 (91e:60112)

[QR78] Richard E. Quandt and James B. Ramsey, *Estimating mixtures of normal distributions and switching regressions*, J. Amer. Statist. Assoc. **73** (1978), no. 364, 730–752, With comments and a rejoinder by the authors. MR 521324 (80a:62040)

[RBHDP07] Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin, and Laurent Pierre, *A semi-parametric approach for mixture models: application to local false discovery rate estimation*, Comput. Statist. Data Anal. **51** (2007), no. 12, 5483–5493. MR 2407654

[RRDP03] A. C. Robin, C. Reyl, S. Derrire, and S. Picaud, *A synthetic view on structure and evolution of the milky way*, Astronomy and Astrophysics **409** (2003), no. 1, 523–540.

[RWD88] Tim Robertson, F. T. Wright, and R. L. Dykstra, *Order restricted statistical inference*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Ltd., Chichester, 1988. MR 961262 (90b:62001)

[SC04] Stan Salvador and Philip Chan, *Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms*, Proc. 16th IEEE Intl. Conf. on Tools with AI **25** (2004), 576–584.

[Sto02] John D. Storey, *A direct approach to false discovery rates*, J. R. Stat. Soc. Ser. B Stat. Methodol. **64** (2002), no. 3, 479–498. MR 1924302 (2003h:62029)

[TSS01] F. Turkheimer, C.B Smith, and K. Schmidt, *Estimation of the number of "true null hypotheses in multivariate analysis of neuroimaging data*, NeuroImage **13** (2001), no. 5, 920–930. MR 1924302 (2003h:62029)

[vdG00] Sara A. van de Geer, *Applications of empirical process theory*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 6, Cambridge University Press, Cambridge, 2000. MR 1739079 (2001h:62002)

[WMO+09] M.G. Walker, M. Mateo, E.W. Olszewski, B. Sen, and M. Woodroofe, *Clean kinematic samples in dwarf spheroidals: An algorithm for evaluating membership and estimating distribution parameters when contamination is present*, The Astronomical Journal **137** (2009), 3109.

[WOG+07] Mario Walker, Matthew G.and Mateo, Edward W. Olszewski, Oleg Y. Gnedin, Xiao Wang, Bodhisattva Sen, and Michael Woodroofe, *Velocity dispersion profiles of seven dwarf spheroidal galaxies*, Astrophysical J. **667** (2007), no. 1, L53–L56.

[WS93] Michael Woodroofe and Jiayang Sun, *A penalized maximum likelihood estimate of $f(0+)$ when f is nonincreasing*, Statist. Sinica **3** (1993), no. 2, 501–515. MR 1243398 (94i:62044)