Kullback-Leibler Projections, Estimation of Mixing Distributions and Applications

by

R. Martin
Purdue University

S. Tokdar
Carnegie Mellon University

Technical Report #08-06

Department of Statistics
Purdue University

November 2008

Kullback-Leibler Projections, Estimation of Mixing Distributions and Applications

Ryan Martin and Surya T. Tokdar
Purdue University and Carnegie Mellon University

November 15, 2008

Abstract

Newton's recursive estimate [23] of a mixing distribution is an accurate and computationally efficient competitor to the Bayes and ML estimates. It has been shown [20, 33] that if the mixture model is correct, then the recursive estimate is consistent in the weak topology under mild conditions. Here we address the asymptotic behavior of the recursive estimate when the model is *incorrect*. We show that the corresponding estimate of the marginal distribution converges almost surely in L_1 to the "best possible" mixture of the specified form. This generalization allows us to extend the algorithm's scope of application, particularly, to the case where additional unknown parameters are present in the model. A "likelihood-based" estimate of the additional unknown parameters is proposed. Important applications are considered—fitting finite mixtures and large-scale multiple testing—and the performance of the resulting methodology is investigated in several simulated- and real-data examples.

1 Introduction

Historically, mixture distributions have been used primarily to model data that show population heterogeneity or involve latent variables. More recently, the growing popularity of nonparametric empirical Bayes methodology, thanks to scientific advances such as DNA microarrays and medical and satellite imaging technologies, has opened the door to many new and exciting applications of mixture distributions. Typically, the inference problem requires a nonparametric estimate of this underlying mixing distribution which, in addition to being a challenging computational problem, often requires rather strict assumptions on the sampling model.

Suppose that the pairs $(X_i, \theta_i) \in \mathcal{X} \times \Theta$ are independently distributed according to the following model:

$$\theta_i \sim f \quad \text{and} \quad X_i|\theta_i \sim p(\cdot|\theta_i), \qquad i = 1, \ldots, n \tag{1.1}$$

where $p(x|\theta)$ is a known parametric density on the sample space $\mathcal{X}$ and f is a density on Θ with respect to a σ-finite measure μ. The θ_i's are unobservable, so the marginal density of $X_1, \ldots, X_n$ is the mixture

$$m(x) = \int_\Theta p(x|\theta) f(\theta)\, d\mu(\theta). \tag{1.2}$$

From a Bayesian perspective, f represents the prior distribution for the parameters $\theta_1, \ldots, \theta_n$. In high-dimensional problems, such as nonparametric regression [22, 5] or DNA microarray data analysis [13, 1, 21, 4], where the the number of parameters n is very large, fixing f completely or only up to a few unknown parameters can be unsatisfactory. The empirical Bayes [28] approach—using the data directly to estimate the prior—is a more flexible alternative and has been applied quite successfully in many problems.

In this paper, we focus our attention on the empirical Bayes approach and, in particular, on a fast recursive algorithm for nonparametric estimation of f, due to Newton [23]; see also [24, 25]. Suppose we have an independent sample $X_1, \ldots, X_n$ from the mixture $m(x)$ in (1.2). Newton proposed the following algorithm for estimating the mixing density f.

Recursive Algorithm. Choose an initial density f_0 on Θ and a sequence of weights $w_1, \ldots, w_n \in (0,1)$. For $i = 1, \ldots, n$, compute

$$f_i(\theta) = (1 - w_i) f_{i-1}(\theta) + w_i \frac{p(X_i|\theta) f_{i-1}(\theta)}{\int p(X_i|\theta') f_{i-1}(\theta')\, d\mu(\theta')}, \quad \theta \in \Theta \tag{1.3}$$

and produce f_n as the final estimate.

While f_n is not a Bayesian posterior quantity—it depends on the order of the data—the original motivation [24] was that f_n would be *approximately* Bayes. Indeed, if f is given a Dirichlet process (DP) prior distribution with base measure f_0 and precision $1/w_1 - 1$, then f_1 is exactly the posterior mean of f. This connection breaks down for $n > 1$, but Newton, *et al.* [24] conjectured that, by extending the analogy, f_n would serve as a suitable approximation to DP prior Bayes estimate. Simulation results presented in [15, 20, 33] suggest that this conjecture is false, but it is possible that f_n approximates some other posterior quantity, maybe under a different prior.

Until recently, very little was known about the large-sample behavior of the estimate f_n. Ghosh and Tokdar [15] used a novel martingale argument to prove, in the case where Θ is a finite set, that $f_n \to f$ almost surely. Martin and Ghosh [20] prove a slightly stronger consistency theorem by making use of a stochastic approximation representation of the recursion (1.3). Most recently, Tokdar, *et al.* [33] have handled the case of a more general parameter space Θ by extending the martingale argument [15] to the $\mathcal{X}$-space, proving that the marginal density estimate

$$m_n(x) = \int p(x|\theta) f_n(\theta)\, d\mu(\theta) \tag{1.4}$$

converges a.s. to $m(x)$ in the L_1 topology. From this, consistency of f_n in the weak topology on Θ is deduced.

In these earlier treatments of the recursive estimate, the operating assumption is that the density m is a mixture of the form (1.2). This assumption seems to unnecessarily limit the scope and applications of the algorithm. For example, in the Bayesian context, where existence of prior density f is usually taken for granted, there remains some uncertainty regarding the conditional density of the data $p(x|\theta)$. The theoretical and practical performance of the recursive estimate has been investigated only in the case where θ completely specifies the density $p(x|\theta)$. This excludes, for example, the case where $p(x|\theta)$ is a $N(\theta, \sigma^2)$ density with *unknown* variance σ^2. The natural solution would be to estimate σ^2 from the data and use a plug-in $p(x|\theta,\hat{\sigma}^2)$ in the recursive algorithm. Unfortunately, the present theory can say nothing about the performance of such a procedure because $m(x)$ is *not* a mixture of $p(x|\theta,\hat{\sigma}^2)$ densities.

The assumption that the data follows a mixture distribution is most often based on empirical evidence; *e.g.*, if a histogram of the data shows multiple modes, then a mixture model might be used to analyze the data. This procedure is justified by the well-known fact that any density on Euclidean space can be closely approximated by a suitable mixture; see, for example, DasGupta [8, Theorem 33.1]. Considering the computational speed and simplicity of the recursive estimate, coupled with the approximation theorem just mentioned, one might propose using the recursive estimate in general density estimation problems; see Section 6. Again, the present theory says nothing about the performance of such a procedure.

The goal of this paper is to understand the asymptotic properties of the recursive estimate in the more general case. What kind of asymptotic properties can we hope for? Of course, if m is not a mixture of the form (1.2), then there is no hope that m_n is a consistent estimate of m. This is not a fault of the estimate, but a consequence of the model mis-specification. A natural question arises: if there is no f such that $m(x)$ is of the form (1.2), then what are f_n and m_n estimating? We will show that m_n converges to the "best possible" mixture of the form (1.2) in a sense to be made more precise below.

Problems of this kind have been considered by many authors, for example, Csiszár [6], Csiszár and Tusnády [7], Dykstra [9], Leroux [17], Shyamalkumar [31] and others. In one way or another, the concept of "Kullback-Leibler (KL) projections" is introduced. A KL projection of a probability measure P onto a class of probability measures $\mathbb{Q}$ is the measure $Q^* \in \mathbb{Q}$ which is "closest" to P is a KL sense; *i.e.*,

$$K(P,Q^*) = \inf\{K(P,Q) : Q \in \mathbb{Q}\},$$

where $K(P,Q) = \int \log(dP/dQ)\,dP$ is the KL divergence of Q from P. More details on these KL projections, including sufficient conditions for the existence of Q^*, are given in Section 2.

For a preview of our results, we first define some notation. Recall that $\mathcal{X}$ is the sample space, and Θ is the parameter space equipped with a σ-finite measure μ. Let $\mathbb{M} = \mathbb{M}(\mathcal{X},\nu)$ be the set of all probability densities on $\mathcal{X}$ with respect to a specified σ-finite measure ν and $\mathscr{P} = \{p(\cdot|\theta) : \theta \in \Theta\} \subset \mathbb{M}$ a parametric family. Let $\mathbb{F} = \mathbb{F}(\Theta,\mu)$ be the set of all probability densities on Θ with respect to μ. We assume that the data $X_1,\ldots,X_n \in \mathcal{X}$ are iid observations with common density $m \in \mathbb{M}$. This m is unknown to us but we model it with a mixture (1.2). That is, we assume $m \in \mathbb{M}_\Theta \subset \mathbb{M}$, where

$$\mathbb{M}_\Theta = \{m \in \mathbb{M} : m = m_\varphi \ \exists\, \varphi \in \mathbb{F}\}.$$

where $m_\varphi(x) = \int p(x|\theta)\varphi(\theta)\,d\mu(\theta)$ is a φ-mixture of $\mathscr{P}$. If indeed $m \in \mathbb{M}_\Theta$, then the consistency theorem in [33] would apply. Here we prove, more generally, that

$$K(m,m_n) \to \inf\{K(m,m') : m' \in \mathbb{M}_\Theta\} \quad \text{a.s.} \tag{1.5}$$

where $K(m,m') = \int \log(m/m')\,m\,d\nu$. That is, even when the model is misspecified (*i.e.*, $m \notin \mathbb{M}_\Theta$), Newton's estimate m_n in (1.4) converges a.s. in L_1 to the KL-best mixture in $\mathbb{M}_\Theta$.

A general discussion of Kullback-Leibler projections, along with a theorem giving sufficient conditions for the existence of such a projection, can be found in Section 2. In Section 3, after establishing some preliminary results, we prove our main theorem and show that the consistency results of Tokdar, *et al.* [33] follow as special cases. Our "likelihood-based" extension of the recursive algorithm is presented in Section 4. The resulting methodology is applied to two important statistical problems, namely fitting finite mixture models and large-scale simultaneous hypothesis testing. For the former, we can estimate both the mixture complexity as well as the mixture characteristics and, for the latter, we use a more flexible *zero-assumption* than Efron [12]. We assess the performance of these procedures in real- and simulated-data examples. Some final remarks are given in Section 6.

2 Kullback-Leibler projections

For probability measures $P, Q \in \mathbb{P}$, with $P \ll Q$, the Kullback-Leibler (KL) divergence of Q from P is

$$K(P, Q) = \int \log(dP/dQ)\, dP,$$

where dP/dQ is the Radon-Nikodym derivative of P with respect to Q. If both P and Q have densities p and q, respectively, with respect to a common measure ν, then $K(P, Q) = \int p \log(p/q)\, d\nu$. KL divergences arise quite frequently in statistics, primarily due to the ubiquity of likelihood ratios. In our case, the occurrence of the KL divergence is quite natural. Martin and Ghosh [20] show that the KL divergence acts as a Lyapunov function, controlling the dynamics of the algorithm and forcing the estimates to converge to a stable equilibrium.

We will be interested in identifying a probability measure $Q^* \in \mathbb{Q}$, with $\mathbb{Q} \subset \mathbb{P}$, such that Q^* is closest to P in a KL sense; that is,

$$K(P, Q^*) = K(P, \mathbb{Q}) := \inf\{K(P, Q) : Q \in \mathbb{Q}\} \tag{2.1}$$

We call Q^* the KL projection of P onto $\mathbb{Q}$. In this section we prove a simple theorem which gives sufficient conditions for the existence of a KL projection in our special case of mixtures. Many interesting details on more general divergence measures can be found in Liese and Vadja [18].

The most important issue here is *existence*—uniqueness is automatic from convexity of $\mathbb{Q}$ and of the KL divergence mapping $Q \mapsto K(P, Q)$. The following theorem gives conditions which imply the existence of a density $f \in \mathbb{F}$ such that $K(m, m_f) = K(m, \mathbb{M}_\Theta)$.

Theorem 2.1. *If $\mathbb{F}$ is weakly compact and $\theta \mapsto p(x|\theta)$ is bounded and continuous for ν-a.e. x, then $\exists\ f \in \mathbb{F}$ such that $K(m, m_f) = K(m, \mathbb{M}_\Theta)$.*

Proof. Choose any $\varphi \in \mathbb{F}$ and a sequence $\{\varphi_k\} \subset \mathbb{F}$ such that $\varphi_k \to \varphi$ weakly. Let $m_\varphi(x) = \int p(x|\theta)\varphi(\theta)\, d\mu(\theta)$ define the corresponding sequence $m_k = m_{\varphi_k}$ of mixture densities in $\mathbb{M}_\Theta$. Then weak convergence of φ_k to φ and the assumption that $\theta \mapsto p(x|\theta)$ is ν-a.e. x bounded and continuous implies, by Scheffé's theorem, that $m_k \to m_\varphi$ in $L_1(\nu)$. Let $\kappa(\varphi) = K(m, m_\varphi)$. Since $m_k \to m_\varphi$ in

ν-measure, it follows from Fatou's lemma that

$$\begin{aligned}\kappa(\varphi) &= \int \lim_k \log(m/m_k)\, m\, d\nu \\ &\leq \liminf_k \int \log(m/m_k)\, m\, d\nu = \liminf_k \kappa(\varphi_k)\end{aligned}$$

Thus κ is lower semi-continuous with respect to the weak topology and, therefore, must attain its infimum on the compact $\mathbb{F}$. □

Theorem 2.1 shows that m_f is the KL projection of m onto $\mathbb{M}_\Theta$. Convexity implies m_f is unique, but *identifiability* is required for uniqueness of f.

3 Convergence of the recursive estimate

in this section we prove the claim (1.5) that m_n converges a.s. in L_1 to the KL projection of m onto $\mathbb{M}_\Theta$, where $\mathbb{M}_\Theta \subset \mathbb{M}$ is the convex hull of $\mathscr{P} = \{p(\cdot|\theta) : \theta \in \Theta\}$. Here, and in what follows, we will consider the following conditions:

A1. $\sum_n w_n = \infty$ and $\sum_n w_n^2 < \infty$.

A2. $\mathbb{F}$ is weakly compact.

A3. $\theta \mapsto p(x|\theta)$ bounded and continuous for ν-a.e. $x \in \mathscr{X}$.

A4. There exists a constant $B < \infty$ such that, for every $\theta, \theta' \in \Theta$

$$\int_{\mathscr{X}} \left[\frac{p(x|\theta)}{p(x|\theta')}\right]^2 m(x)\, d\nu(x) < B.$$

The condition A1 on the weight sequence $\{w_n\}$ is natural, given the stochastic approximation representation of the recursive algorithm presented in Martin and Ghosh [20]. Conditions A2 and A3 together imply the existence of a density $f \in \mathbb{F}$ such that $K(m, m_f) = K(m, \mathbb{M}_\Theta)$; see Theorem 2.1. Note that this f need not be unique without identifiability. A sufficient condition for A2 is that Θ be a compact metric space. The square-integrability condition A4 is by far our strongest assumption. Since we are not assuming $m(x)$ is of the form (1.2), condition A4 cannot be written as a property of the parametric family $\mathscr{P}$ alone as in Tokdar, *et al.* [33]. A4 does hold for many common families $\mathscr{P}$, such as

- Normal with mean θ and any fixed variance $\sigma^2 > 0$
- Gamma with rate θ and any fixed shape $\alpha > 0$
- Poisson with mean θ

provided that Θ is compact and $m(x)$ admits a moment-generating function on Θ. The following proposition gives an important implication of A4.

Proposition 3.1. *For any* $\varphi, \psi \in \mathbb{F}$, A4 *implies*

$$\int_{\mathscr{X}} \left[\frac{m_\varphi(x)}{m_\psi(x)}\right]^2 m(x)\, d\nu(x) < B. \tag{3.1}$$

Proof. Follows from Jensen's inequality. □

The proof of our main result will be partially based on the development in Tokdar, *et al.* [33]. We proceed by recording a few of these results for future reference. First, let $R(x)$ be the remainder term of a first-order Taylor expansion of $\log(1+x)$ at $x=0$; that is,

$$\log(1+x) = x - x^2 R(x), \quad x > -1. \tag{3.2}$$

Two useful inequalities will be needed:

$$0 \le R(x) \le \max\{1, (1+x)^{-2}\}, \quad x > -1 \tag{3.3}$$

and, for $a > 0$ and $b \in (0,1)$,

$$(a-1)^2 \max\{1, (1+b(a-1))^{-2}\} \le \max\{(a-1)^2, (1/a-1)^2\} \tag{3.4}$$

The mixture density $m_n \in \mathbb{M}_\Theta$ in (1.4) corresponding to Newton's estimate $f_n \in \mathbb{F}$ is of the form

$$m_n(x) = (1-w_n)m_{n-1}(x) + w_n h_{n,X_n}(x),$$

where

$$h_{n,x'}(x) = \int_\Theta \frac{p(x|\theta)p(x'|\theta)f_{n-1}(\theta)}{m_{n-1}(x')}\, d\mu(\theta), \quad x, x' \in \mathcal{X}.$$

For notational convenience, also define the function

$$H_{n,x'}(x) = \frac{h_{n,x'}(x)}{m_{n-1}(x)} - 1, \quad x, x' \in \mathcal{X}.$$

Then the KL divergence $K_n = K(m, m_n)$ satisfies

$$\begin{aligned} K_n - K_{n-1} &= \int_{\mathcal{X}} m \log(m_{n-1}/m_n)\, d\nu \\ &= -\int_{\mathcal{X}} m \log(1 + w_n H_{n,X_n})\, d\nu \\ &= -w_n \int_{\mathcal{X}} m\, H_{n,X_n}\, d\nu + w_n^2 \int_{\mathcal{X}} m\, H_{n,X_n}^2 R(w_n H_{n,X_n})\, d\nu \end{aligned}$$

where $R(x)$ is defined in (3.2). Let $\mathscr{A}_{n-1}$ be the σ-algebra generated by the data sequence $X_1, \ldots, X_{n-1}$. Since K_{n-1} is $\mathscr{A}_{n-1}$-measurable, upon taking conditional expectation with respect to $\mathscr{A}_{n-1}$ we get

$$\mathsf{E}(K_n|\mathscr{A}_{n-1}) - K_{n-1} = -w_n T(f_{n-1}) + w_n^2 \mathsf{E}(Z_n|\mathscr{A}_{n-1}) \tag{3.5}$$

where $T(\cdot)$ and Z_n are defined as

$$T(\varphi) = \int_\Theta \left\{ \int_{\mathcal{X}} \frac{m(x)}{m_\varphi(x)} p(x|\theta)\, d\nu(x) \right\}^2 \varphi(\theta)\, d\mu(\theta) - 1, \quad \varphi \in \mathbb{F} \tag{3.6}$$

$$Z_n = \int_{\mathcal{X}} m\, H_{n,X_n}^2 R(w_n H_{n,X_n})\, d\nu \tag{3.7}$$

Note that $T(f_{n-1})$ is exactly M_n^* defined in [33]. The following property of $T(\cdot)$ will be critical in the proof of our main result.

Lemma 3.2. $T(\varphi) \geq 0$ *with equality* iff $K(m, m_\varphi) = K(m, \mathbb{M}_\Theta)$.

Proof. Treating $\theta \sim \varphi$ as a random element in Θ, define

$$g_\varphi(\theta) = \int_{\mathcal{X}} \frac{m(x)}{m_\varphi(x)} p(x|\theta)\, d\nu(x), \quad \varphi \in \mathbb{F}. \tag{3.8}$$

Then $\mathsf{E}_\varphi[g_\varphi(\theta)] = \int g_\varphi \varphi \, d\mu = 1$ and $T(\varphi) = \mathsf{V}_\varphi[g_\varphi(\theta)] \geq 0$, with equality iff $g_\varphi = 1$ μ-a.e., where E_φ and V_φ denote expectation and variance under φ, respectively. To show that $T(\varphi) = 0$ implies $K(m, m_\varphi) = K(m, \mathbb{M}_\Theta)$, we follow Shyamalkumar [31]. Define

$$G(\varphi) = \log \left\{ \int_\Theta g_\varphi(\theta)\, f(\theta)\, d\mu(\theta) \right\},$$

where $f \in \mathbb{F}$ is such that $K(m, m_f) = K(m, \mathbb{M}_\Theta)$. Note that $T(\varphi) = 0$ implies $G(\varphi) = 0$. By Jensen's inequality

$$\begin{aligned} G(\varphi) &= \log \left\{ \int_\Theta \left[\int_{\mathcal{X}} \frac{m(x)}{m_\varphi(x)} p(x|\theta)\, d\nu(x) \right] f(\theta)\, d\mu(\theta) \right\} \\ &= \log \left\{ \int_{\mathcal{X}} \frac{m_f(x)}{m_\varphi(x)} m(x)\, d\nu(x) \right\} \\ &\geq \int_{\mathcal{X}} \log \left(\frac{m_f(x)}{m_\varphi(x)} \right) m(x)\, d\nu(x) \\ &= K(m, m_\varphi) - K(m, m_f) \geq 0 \end{aligned}$$

so that $G(\varphi) = 0$ implies $K(m, m_\varphi) = K(m, m_f)$. □

From (3.5) we see the makings of a supermartingale in K_n. Indeed, if it were not for the term involving Z_n we would have a non-negative supermartingale and convergence of K_n would follow immediately from the martingale convergence theorem [2]. Fortunately, the presence of Z_n causes only minor difficulties.

Lemma 3.3. *Under* A4, $\{Z_n\}$ *is uniformly bounded* a.s.

Proof. Inequalities (3.3) and (3.4) show that

$$H_{n,X_n}^2 R(w_n H_{n,X_n}) \leq \max \left\{ \left(\frac{h_{n,X_n}}{m_{n-1}} - 1 \right)^2, \left(\frac{m_{n-1}}{h_{n,X_n}} - 1 \right)^2 \right\}$$

and, since $h_{n,X_n}, m_{n-1} \in \mathbb{M}_\Theta$ for each n, the claim follows from A4 and Proposition 3.1. In particular, $Z_n < 1 + B$ a.s. □

Our last preliminary result, Lemma 3.4 below, establishes the necessary smoothness properties of the mapping $T(\cdot)$ on $\mathbb{F}$.

Lemma 3.4. *Under* A3–A4, $T(\cdot)$ *is continuous with respect to the weak topology on* $\mathbb{F}$.

Proof. Take a sequence $\{\varphi_n\} \subset \mathbb{F}$ such that $\varphi_n \to \varphi$. Since the weak topology is metrizable, it suffices to show that $T(\varphi_n) \to T(\varphi)$. From weak convergence of φ_n to φ we can conclude $L_1(\nu)$ convergence of m_{φ_n} to m_φ by A3 and Scheffé's Theorem. Recall the definition of g_φ and g_{φ_n} in (3.8). The integrand in $g_{\varphi_n}(\theta)$ is bounded by $\sqrt{B}$ for μ-a.e. θ by A4, so it follows from the dominated convergence theorem that $g_{\varphi_n} \to g_\varphi$ and, hence, $g_{\varphi_n}^2 \to g_\varphi^2$ μ-a.e. Moreover, $g_{\varphi_n}^2 < B$ by A4, so $\{g_{\varphi_n}^2\}$ is uniformly integrable. Then $\mathsf{E}_{\varphi_n}[g_{\varphi_n}^2] \to \mathsf{E}_\varphi[g_\varphi^2]$ and it follows that $T(\varphi_n) \to T(\varphi)$. But $\{\varphi_n\}$ and φ were arbitrary so T must, therefore, be weakly continuous on $\mathbb{F}$. □

At last, we now have the notation and machinery to precisely state and prove our main result, namely, that m_n converges to the KL-projection of m onto the space $\mathbb{M}_\Theta$ of mixtures.

Theorem 3.5. *Under* A1–A4, $K_n \to K(m, \mathbb{M}_\Theta)$ a.s.

Proof. The proof begins with (3.5). Define the random variables

$$\delta_n = \sum_{i=n+1}^{\infty} w_i^2 \mathsf{E}(Z_i | \mathscr{A}_n) \tag{3.9}$$

and set $\widetilde{K}_n = K_n + \delta_n$. The sequence $\{\delta_n\}$ in (3.9) has three important properties, which we establish below.

(i) Obviously $\delta_n \geq 0$ by definition, so $\widetilde{K}_n \geq 0$ too.

(ii) Under A1, Lemma 3.3 implies that $\sum_n w_n^2 \mathsf{E}(Z_n) < \infty$ and it follows that the δ_n's are bounded. Since $\sum_{i=n+1}^{\infty} w_i^2 Z_i \to 0$ as $n \to \infty$, we conclude from the bounded convergence theorem that $\delta_n \to 0$ a.s.

(iii) Since $\mathscr{A}_{n-1} \subset \mathscr{A}_n$, we have

$$\mathsf{E}(\delta_n | \mathscr{A}_{n-1}) = \sum_{i=n+1}^{\infty} w_i^2 \mathsf{E}(Z_i | \mathscr{A}_{n-1})$$

which implies

$$\mathsf{E}(\delta_n | \mathscr{A}_{n-1}) - \delta_{n-1} = -w_n^2 \mathsf{E}(Z_n | \mathscr{A}_{n-1}). \tag{3.10}$$

Combining (3.5) and (3.10), we get

$$\begin{aligned}
\mathsf{E}(\widetilde{K}_n | \mathscr{A}_{n-1}) - \widetilde{K}_{n-1} &= \mathsf{E}(K_n | \mathscr{A}_{n-1}) - K_{n-1} + \mathsf{E}(\delta_n | \mathscr{A}_{n-1}) - \delta_{n-1} \\
&= -w_n T(f_{n-1}) + w_n^2 \mathsf{E}(Z_n | \mathscr{A}_{n-1}) - w_n^2 \mathsf{E}(Z_n | \mathscr{A}_{n-1}) \\
&= -w_n T(f_{n-1}) \\
&\leq 0
\end{aligned}$$

Therefore, $\widetilde{K}_n$ forms a non-negative supermartingale and, hence, there is a $K_\infty \geq K(m, \mathbb{M}_\Theta)$ such that $\widetilde{K}_n \to K_\infty$ a.s. In fact, $K_n \to K_\infty$ by (ii). It remains to show $K_\infty = K(m, \mathbb{M}_\Theta)$ a.s.

Suppose $K_\infty > K(m, \mathbb{M}_\Theta)$ with positive probability. From the previous display we have

$$\widetilde{K}_{i-1} - \mathsf{E}(\widetilde{K}_i | \mathscr{A}_{i-1}) = w_i T(f_{i-1}), \quad \forall\, i. \tag{3.11}$$

Fix any two integers N and n. Taking conditional expectation with respect to $\mathscr{A}_{N-1}$ in (3.11) and summing gives

$$\begin{aligned}
\sum_{i=N}^{N+n} w_i \mathsf{E}[T(f_{i-1})|\mathscr{A}_{N-1}] &= \sum_{i=N}^{N+n} \mathsf{E}[\widetilde{K}_{i-1} - \mathsf{E}(\widetilde{K}_i|\mathscr{A}_{i-1}) \mid \mathscr{A}_{N-1}] \\
&= \sum_{i=N}^{N+n} \Big\{ \mathsf{E}(\widetilde{K}_{i-1}|\mathscr{A}_{N-1}) - \mathsf{E}[\mathsf{E}(\widetilde{K}_i|\mathscr{A}_{i-1}) \mid \mathscr{A}_{N-1}] \Big\} \\
&= \sum_{i=N}^{N+n} \mathsf{E}[\widetilde{K}_{i-1} - \widetilde{K}_i \mid \mathscr{A}_{N-1}] \\
&= \widetilde{K}_{N-1} - \mathsf{E}(\widetilde{K}_{N+n}|\mathscr{A}_{N-1}) \\
&\le \widetilde{K}_{N-1}
\end{aligned} \tag{3.12}$$

If $K_\infty > K(m, \mathbb{M}_\Theta)$, then there exists $\varepsilon > 0$ such that

$$K(m, m_n) > K(m, \mathbb{M}_\Theta) + \varepsilon$$

for all but finitely many n. In the proof of Theorem 2.1 it was shown that the mapping $\kappa(\varphi) = K(m, m_\varphi)$ is lower semi-continuous with respect to the weak topology on $\mathbb{F}$. Consequently,

$$O_\varepsilon = \{\varphi \in \mathbb{F} : \kappa(\varphi) > K(m, \mathbb{M}_\Theta) + \varepsilon\} \subset \mathbb{F}$$

is a weakly open set. Its closure $\overline{O}_\varepsilon$ is compact and

$$\overline{O}_\varepsilon \cap \{\varphi \in \mathbb{F} : \kappa(\varphi) = K(m, \mathbb{M}_\Theta)\} = \emptyset.$$

Since $f_n \in \overline{O}_\varepsilon$ for all but finitely many n, it follows from Lemma 3.4 that $T(f_n)$ is bounded away from zero. This, together with A2, implies that, with positive probability, the left-hand side of (3.12) goes to ∞ as $n \to \infty$. But $\widetilde{K}_{N-1}$ is finite with probability 1—a contradiction! Therefore, $K_\infty = K(m, \mathbb{M}_\Theta)$ a.s., completing the proof. □

Theorem 3.5 states that the recursive estimates f_n converge to the f at which the infimum $K(m, \mathbb{M}_\Theta)$ is attained. But to conclude weak convergence of f_n from L_1 convergence of m_n, we need two additional conditions:

A5. Identifiability: $m_\varphi = m_\psi$ ν-a.e. implies $\varphi = \psi$ μ-a.e.

A6. For any $\varepsilon > 0$ and any compact $\mathcal{X}_0 \subset \mathcal{X}$, there exists a compact $\Theta_0 \subset \Theta$ such that $\int_{\mathcal{X}_0} p(x|\theta)\, d\nu(x) < \varepsilon$ for all $\theta \notin \Theta_0$.

With conditions A5–A6 and Theorem 3 of Tokdar, *et al.* [33], the next two corollaries follow immediately from Theorem 3.5.

Corollary 3.6. *If conditions* A1–A6 *hold, then* $f_n \to f$ a.s. *in the weak topology, where* $f \in \mathbb{F}$ *satisfies* $K(m, m_f) = K(m, \mathbb{M}_\Theta)$.

Corollary 3.7. *If* $m \in \mathbb{M}_\Theta$, *then under* A1–A4, $m_n \to m$ a.s. *in the* L_1 *topology. Moreover, if* A5–A6 *hold then* f_n *converges* a.s. *in the weak topology to the mixing density* $f \in \mathbb{F}$ *that satisfies* (1.2).

4 Extensions: the RE+ algorithm

Thus far, we have assumed that the sampling densities $p(x|\theta)$ are completely specified by the parameter θ. However, greater modeling flexibility can be achieved by choosing a richer family of sampling densities and using the data to specify the additional unknowns. A serious drawback of the recursive algorithm is that it cannot directly handle additional unknown parameters [3]. In this section, we extend the recursive algorithm to find the best possible mixture over the augmented space of mixing distributions and the additional parameters.

Let $\xi \in \Xi$ represent this "non-mixing" parameter and let $p(x|\theta,\xi)$ be the corresponding family of sampling densities on $\mathcal{X}$. In this section, we will assume that the density m is a mixture of the form

$$m(x) = \int p(x|\theta,\xi^*) f^*(\theta)\, d\mu(\theta), \tag{4.1}$$

where both f^* and ξ^* are unknown and to be estimated. Even the dominating measure μ can depend on ξ^* (*i.e.*, $\mu = \mu_{\xi^*}$) but we shall generally suppress this dependence in the notation. Let $m_{f,\xi}$ denote a mixture of the form (4.1) with (f,ξ) in place of the unknown (f^*,ξ^*). Martin and Ghosh [20] considered the problem of estimating the pair (f^*,ξ^*) in the case of a finite Θ. They assumed replicates $X_{i1},\ldots,X_{ir}$ were available from $p(x|\theta_i,\xi)$ for each $i = 1,\ldots,n$ and proposed a simple modification of the algorithm which could recursively estimate both f and ξ. Our approach here is different.

Our jumping off point is that, for fixed $\xi \in \Xi$, the marginal density $m_{n,\xi} = m_{f_n,\xi}$ based on the recursive algorithm will converge, by Theorem 3.5, to the mixture of $p(x|\theta,\xi)$ densities that minimizes the KL divergence. This observation suggests that we estimate the pair (f^*,ξ^*) by minimizing $K(m, m_{f,\xi})$ over $\mathbb{F} \times \Xi$. Unfortunately, the density m is unknown so this quantity cannot be calculated. The natural modification would be to replace expectation with respect to m by expectation with respect to the empirical distribution of $X_1,\ldots,X_n$. That is, instead minimize

$$L_n(\xi) = \frac{1}{n}\sum_{i=1}^{n} \log\{m(X_i)/m_{i-1,\xi}(X_i)\} \tag{4.2}$$

over Ξ. Note that minimizing $L_n(\xi)$ is equivalent to maximizing a pseudo-likelihood function $\widetilde{L}_n(\xi)$, given by

$$\widetilde{L}_n(\xi) = \sum_{i=1}^{n} \log m_{i-1,\xi}(X_i), \tag{4.3}$$

so evaluation of m is not required. To rigorously justify our intuition, we need a modification of assumption A4 used in the proof of Theorem 3.5. Recall that we are assuming that m is itself a mixture, so condition A4 can be rewritten in terms of the sampling densities $p(x|\theta,\xi)$. Specifically,

A4′. There exists $B < \infty$ such that

$$\int_{\mathcal{X}} \left[\frac{p(x|\theta_1,\xi_1)}{p(x|\theta_2,\xi_2)}\right]^2 p(x|\theta_3,\xi_3)\, d\nu(x) < B$$

for any $(\theta_k,\xi_k) \in \Theta \times \Xi$, $k = 1,2,3$.

Theorem 4.1. *Under conditions* A1–A3 *of Theorem 3.5 and* A4′ *above,*

$$L_n(\xi) \to \inf\{K(m, m_{\varphi,\xi}) : \varphi \in \mathbb{F}\} \quad \text{a.s.}$$

as $n \to \infty$ for fixed $\xi \in \Xi$.

Proof. Define the random variables

$$U_i = \log[m(X_i)/m_{i-1,\xi}(X_i)] - K(m, m_{i-1,\xi}), \quad i \geq 1$$

and note that $\mathsf{E}[U_i|\mathscr{A}_{i-1}] = 0$, where $\mathscr{A}_{i-1} = \sigma(X_i, \ldots, X_{i-1})$. Therefore, $\{(U_n, \mathscr{A}_n) : n \geq 1\}$ forms a zero mean martingale sequence. Furthermore, if we let $\mathcal{E} = \{m < m_{i-1,\xi}\} \subset \mathcal{X}$, then by A4′ and several applications of Jensen's inequality we get

$$\begin{aligned}
\mathsf{E}[U_i^2|\mathscr{A}_{i-1}] &\leq \int_{\mathcal{X}} \left(\log \frac{m}{m_{i-1,\xi}}\right)^2 m\, d\nu \\
&= \int_{\mathcal{E}} \left(\log \frac{m_{i-1,\xi}}{m}\right)^2 m\, d\nu + \int_{\mathcal{E}^c} \left(\log \frac{m}{m_{i-1,\xi}}\right)^2 m\, d\nu \\
&\leq \int_{\mathcal{E}} \left(\frac{m_{i-1,\xi}}{m} - 1\right)^2 m\, d\nu + \int_{\mathcal{E}^c} \left(\frac{m}{m_{i-1,\xi}} - 1\right)^2 m\, d\nu \\
&< C := 2(B+1)
\end{aligned}$$

Then the so-called *conditional variance* of the process $\{U_n\}$, defined as

$$V_n^2 = \sum_{i=1}^n \mathsf{E}[U_i^2|\mathscr{A}_{i-1}],$$

satisfies $a_n\sqrt{V_n^2} \leq n^{1/2} a_n \sqrt{C} \to 0$, where $a_n = n^{-1} \log\log n$. Also, by Chebyshev's inequality,

$$\sum_{n=1}^{\infty} \mathsf{P}\{|U_n| > a_n^{-1} \mid \mathscr{A}_{n-1}\} \leq C \sum_{n=1}^{\infty} a_n^2 < \infty \quad \text{a.s.}$$

and it follows from Corollary 2 of Teicher [32] (with $b_n = n$ and $\beta = 1$) that $n^{-1}\sum_{i=1}^n U_i \to 0$ a.s. Therefore, we can conclude that

$$\left| L_n(\xi) - \frac{1}{n}\sum_{i=1}^n K(m, m_{i-1,\xi}) \right| \to 0 \quad \text{a.s.}$$

However, we know from Theorem 3.5 that $K(m, m_{i-1,\xi})$ and, hence, the sample average $n^{-1}\sum_{i=1}^n K(m, m_{i-1,\xi})$ converges to $\inf_\varphi K(m, m_{\varphi,\xi})$. The claim now follows immediately. □

Remark 4.2. Consider the very general Bayesian model

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(\cdot|f, \xi),$$

where f—possibly a mixing density—has a prior distribution $f \sim \Pi$ and ξ is a hyperparameter. The marginal likelihood of ξ, after integrating out f

with respect to Π, equals $\prod_{i=1}^n m_{i-1,\xi}(X_i)$ where $m_{i-1,\xi}(x)$ equals the posterior predictive density of X_i given $X_1,\ldots,X_{i-1}$. Therefore $\widetilde{L}_n(\xi)$ in (4.3) can be explained as the (log) marginal likelihood of ξ under the Bayesian formulation $f \sim \text{DP}(1/w_1 - 1, f_0)$ obtained by an approximate filtering algorithm in which at step i, the conditional posterior distribution of f given $X_1,\ldots,X_{i-1}$ is approximated by $\text{DP}(1/w_i - 1, f_{i-1,\xi})$.

Remark 4.3. Theorem 4.1 does not imply convergence of the sequence of estimates $\hat{\xi}_n = \arg\max \widetilde{L}_n(\xi)$ unless ξ is restricted to a *finite* set; in general, uniform convergence is needed.

Remark 4.4. In many cases, when additional unknown parameters are introduced, the model may become unidentifiable. For example, a location mixture of normal densities with unknown variance σ^2 is unidentifiable. This can be problematic when the parameters of interest are "real-world" quantities. However, the mixture model structure is often artificially imposed—for modeling simplicity and flexibility—so any set of parameters that provide an adequate fit to the data would suffice.

Remark 4.5. Evaluation of the pseudo-likelihood $\widetilde{L}_n(\xi)$ in (4.3) is performed by passing through the recursive algorithm, either with or without averaging over permutations of the data, with the specified ξ in the sampling density $p(x|\theta,\xi)$. Maximization can then be performed using any available optimization procedure. In our experience, maximization of $\widetilde{L}_n(\xi)$ is relatively fast, usually requiring only a few, relatively inexpensive function evaluations.

Next we present a simple simulation study that demonstrates the performance of the above procedure, which we call RE+ or PARE+, dependeing on whether an average over permuations is included. In this example, we maximize the pseudo-likelihood function $\widetilde{L}_n(\xi)$ in (4.3) numerically using the `nlm` routine in the R statistical software package [26].

Example 4.6. In this example we revisit one of the simulation studies highlighted in Tokdar, *et al.* [33], namely, when $m(x)$ is a location mixture of normals. Specifically, take $\Theta = [0,1]$ and

$$\theta_i \sim \tfrac{1}{3}\text{Beta}(3,30) + \tfrac{2}{3}\text{Beta}(4,4), \quad X_i|\theta_i \sim \text{Normal}(\theta_i, \sigma^2),$$

where sampling is independent across $i = 1,\ldots,n$. In [33] it was assumed that σ was *known*. Here we compare the performance of the recursive algorithm when $\sigma = 0.1$ is known versus the extended version, as described above, that treats σ as *unknown* and estimates it from the observed data.

In this case, the variance is the unknown parameter and we want choose σ^2 to maximize $\widetilde{L}_n(\sigma^2)$. Evaluation of $\widetilde{L}_n(\sigma^2)$ is done via a single pass through the recursive algorithm with weights $w_i = (i+1)^{-1}$ and initial guess f_0 taken to be a $\text{Unif}(\Theta)$ density. Once the maximizing $\hat{\sigma}^2$ is found, we estimate the mixing density with the plug-in recursive estimate, averaged over 25 permutations of the data. For $N = 100$ samples of size $n = 250$, the estimates of the mixing and mixture densities are displayed in Figure 4.1. The top row (RE) shows the estimates when $\sigma = 0.1$ is *known* and the bottom row (RE+) shows the estimates under the extended algorithm with estimated σ^2. There is no noticeable difference between the estimates with σ known versus those with σ estimated. Figure 4.2 shows the L_1 distance $L_1(m,\hat{m})$ of the RE and RE+ estimates from

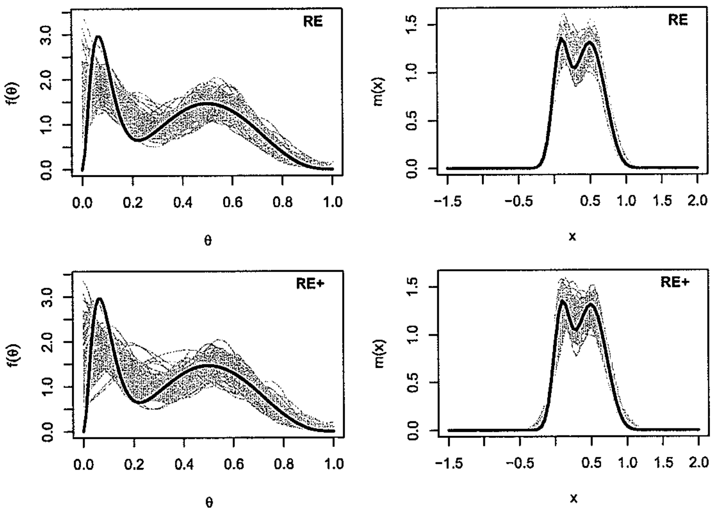

Fig 4.1: Plots of the estimated mixing and mixture densities for RE (top) and the RE+ extension (bottom) for the Beta-Normal simulation in Example 4.6.

the truth, as well as a summary of the $N = 100$ estimates of σ^2. From the L_1 plots we see that the RE+ estimates are slightly worse, the loss of RE+ being about 15% larger than that of RE, on average. However, with the exception of just a few extreme cases, the RE+ estimates of σ^2 are quite accurate, suggesting that there are no identifiability problems in this example. The optimization required between 6 and 7 evaluations of $\widetilde{L}_n(\xi)$ on average. Overall we find that little efficiency is lost—computational or otherwise—when σ is unknown and estimated versus when σ is known.

5 Applications

In this section we consider two important problems and demonstrate how the extension of the recursive algorithm described in Section 4 can be used to solve these problems.

5.1 Fitting finite mixtures

Finite mixture models play an important role in applied statistics. When the number of mixture components is known, numerical methods such as the EM algorithm can be used to estimate the various parameters. When the number of mixture components is unknown, as is generally the case, the problem becomes much more difficult. In such cases, often the goal is to find the most parsimonious mixture—the one with the fewest components—that provides an acceptable fit to the observed data. A nice procedure, based on minimizing a Hellinger distance, is given by Woo and Sriram [37].

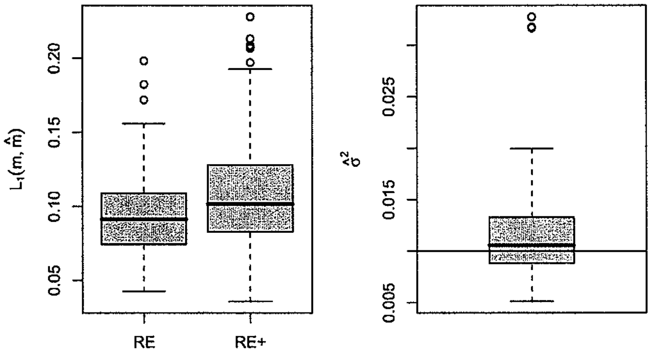

Fig 4.2: Summary of the L_1 loss (left) for RE and RE+ and of the RE+ estimates of σ^2 (right) for the Beta-Normal simulation in Example 4.6.

Suppose $m(x)$ is a finite mixture of the form

$$m(x) = \sum_{r=1}^{R} p(x|\vartheta_r) f^*(\vartheta_r),$$

where f^* has an *unknown* finite support $\{\vartheta_1, \ldots, \vartheta_R\}$ of unknown size R within some known bounded set $\overline{\Theta}$. Martin and Ghosh [20] considered this problem of unknown support of a finite mixing distribution and suggested the following approach: choose a suitably fine grid of points $\Theta = \{\theta_1, \ldots, \theta_S\}$ from $\overline{\Theta}$ as a candidate support and estimate f^* with Newton's estimate f_n on Θ. This intuitive approach will produce a decent answer very quickly but, unfortunately, there are two serious drawbacks.

- f_n and f^* possibly have different (fixed) supports so the consistency theorem in [20] cannot be applied.
- f_n is too smooth in the sense that too many points in Θ are given positive mass; see, *e.g.*, Figure 3.1 in [20] or Figure 19.1 in [15].

For the first problem, Martin and Ghosh [20] conjectured that, for large n, the recursive estimate would produce the KL-best mixture over the candidate support Θ. Theorem 3.5 confirms their conjecture. For the second problem above, it would appear that a modification of the recursive algorithm is required. Instead of modifying the algorithm, we design an estimation procedure that favors small supports. This can be done using the method described in Section 4, treating the support as an unknown non-mixing "parameter" ξ.

Recall that $\Theta = \{\theta_1, \ldots, \theta_S\}$ is the candidate support grid. Let ξ be a binary S-vector, with ξ_s indicating whether or not θ_s receives positive mass. That is, ξ controls which points of Θ are included in the mixture. More precisely, we consider mixtures of the form

$$m_{f,\xi}(x) = \frac{\sum_{s=1}^{S} p(x|\theta_s) f(\theta_s) \xi_s}{\sum_{s=1}^{S} f(\theta_s) \xi_s}.$$

It follows from Theorem 4.1 that $L_n(\xi) \to \inf_f K(m, m_{f,\xi})$ for fixed ξ as $n \to \infty$, which justifies choosing $\xi \in \Xi := \{0,1\}^S$ to maximize $\widetilde{L}_n(\xi)$. For this optimization problem, the solution space has 2^S elements so an exhaustive search procedure cannot be used, even for relatively small S. Instead, a simulated annealing procedure [29, Section 5.2.3] is used to maximize the pseudo-likelihood $\widetilde{L}(\xi)$ over the space Ξ of possible solutions. The initial state $\xi^{(0)}$ of the simulated annealing procedure is taken to be the full vector of 1's, which corresponds to the support of Newton's original estimate f_n. The majority of the early stages of the algorithm will eliminate candidate support points, causing the algorithm to favor smaller supports. At later stages, candidate support points can be added in the move $\xi^{(t)} \to \xi^{(t+1)}$, but this would typically require an increase in the pseudo-likelihood function; *i.e.*, $\widetilde{L}_n(\xi^{(t)}) < \widetilde{L}_n(\xi^{(t+1)})$.

Example 5.1. Consider a simple five-component mixture, with mixture components centered within $\overline{\Theta} = [-5, 5]$ and we choose a grid of candidate support points $\Theta = \{-5.0, -4.75, \ldots, 4.75, 5.0\}$. Each mixture component $p(x|\theta)$ is taken to be a $N(\theta, 0.5^2)$ density. 100 datasets $X_1, \ldots, X_n$ were simulated from the true mixture for $n = 100, 250, 500$. Figure 5.1 shows the RE and RE+ estimates from one particular dataset of size $n = 250$. Note that the RE+ estimate of the mixing distribution closely matches the truth in both the support points and the weights it assigns. Also, the RE+ estimate of the mixture is visually better than that of the RE estimate, although its KL divergence $K(m, \hat{m}_{\mathrm{RE+}})$ is slightly larger. Figure 5.2 summarizes these KL divergences for the two estimates over the three sample sizes considered. We find that the RE+ estimate is a bit unstable for small n but, on average, outperforms RE. Also, in terms of the estimated mixture complexity, RE estimated a constant 41 mixture components while RE+ averaged 9.47, 5.80 and 4.98 mixture components for the three sample sizes, respectively.

In the following example, we apply the above methodology to the well-known galaxy data studied in [30, 14, 27, 35] to name a few.

Example 5.2. Empirical evidence shows that the universe is in a never-ending process of expansion, giving motivation to the Big Bang explanation of how the universe and all its matter came to be. The attraction of matter to other matter—the cause of planet, star, and galaxy formation—also suggests that galaxies themselves should be attracted to one another. Therefore, under the Big Bang model, galaxies should form clusters and the relative velocities of the galaxies should be similar within clusters. Roeder [30] considers data on the measured velocities of $n = 82$ galaxies, relative to our own galaxy. She models this data as a finite mixture of normal densities, with the number and location of mixture components unknown, and assumes that each galactic cluster is a single component of the normal mixture. Multiple mixture components is consistent with the hypothesis of galaxy clustering.

We apply the methodology outlined above to estimate the mixing distribution itself, which immediately gives an estimate of the mixture complexity. Other authors, including Escobar and West [14] and Richardson and Green [27], have fit Bayesian hierarchical models that require a fairly complex Monte Carlo sampling scheme for posterior inference on the number of mixture components. To illustrate our approach, we will consider a simple mixture of normals model in which each normal component has variance $\sigma^2 = 1$. The choice of $\sigma^2 = 1$ is

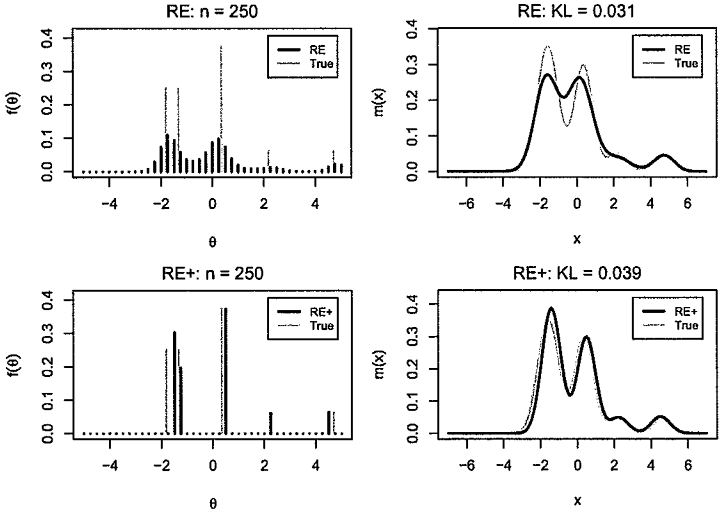

Fig 5.1: Plots of the estimated mixing (left) and mixture densities (right) for the recursive estimate (RE, top) and the extension (RE+, bottom) based on a simulated annealing optimization for one particular dataset of size $n = 250$ in Example 5.1.

based on the *a priori* considerations of Escobar and West [14]: their common prior for the variance of each normal component has unit mean.

From the observed velocities, it is apparent that the mixture components must be centered somewhere in the interval $\overline{\Theta} = [5, 40]$, so we choose a grid of candidate support points $\Theta = \{5.0, 5.5, 6.0, \ldots, 39.5, 40.0\}$; here $S = 71$. Figure 5.3 shows the estimates of the mixing and mixture based on the recursive algorithm and the extended RE+ algorithm via a simulated annealing optimization. The recursive estimate of the mixing density is much too smooth to accurately assess the number of mixture components, and the corresponding mixture density estimate fails to adequately capture the shape of the observed distribution of velocities, as depicted by the histogram. On the other hand, the RE+ estimate of the mixing distribution clearly identifies six, or possibly seven, mixture components, closely matching the conclusions in [30, 14, 27]. The positive weight given to two neighboring θ's by the RE+ is likely an attempt by the algorithm to account for the heavier tails of the component with the second highest peak. Finally, note that, compared to RE, the RE+ mixture density provides a much better fit to the observed galactic velocities.

5.2 Large-scale simultaneous hypothesis testing

Performing numerous statistical tests simultaneously is an important statistical problem these days. Such situations routinely arise in genetics, proteomics, astrophysics, education science, etc. An abstract representation of these situations

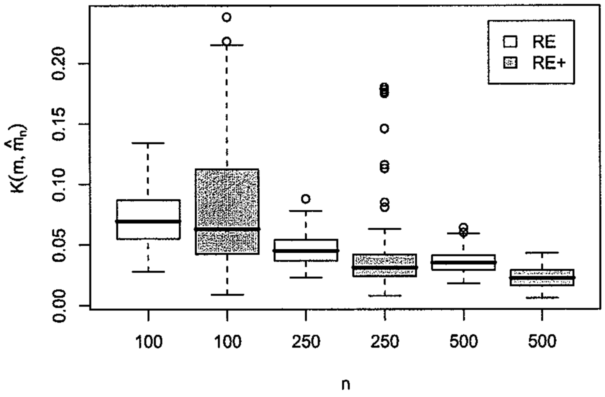

Fig 5.2: Summary of the KL divergences $K(m, \hat{m})$ for RE and RE+ over the 100 simulated datasets of sizes $n = 100$, 250, and 500 in Example 5.1.

is testing a large set of hypotheses

$$H_{0i} : \text{the } i^{\text{th}} \text{ case manifests a ``null'' behavior}, \quad i = 1, \ldots, n$$

based on summary test statistics $z_1, \ldots, z_n$. In applications, n can range from a few hundred to several thousand. Recent statistical research in this area has focused on an empirical Bayes framework that allows for information sharing between cases, even though a separate decision is to be made for each case. Our interest is in the two-groups model championed by Efron [10, 11, 12], where the z_i are assumed to arise from a mixture density

$$m(z) = \pi m_0(z) + (1 - \pi) m_1(z), \tag{5.1}$$

with m_0 encoding the null behavior and m_1 describing the often unspecified alternative behavior of the z-scores.

Usually, by design, the null behavior of z is supposed to match that of a standard normal distribution; think of $z_i = \Phi^{-1}(F_i(t_i))$ where Φ is the standard normal CDF, t_i is a suitable statistic and F_i is the corresponding null distribution for testing the i^{th} case in isolation. But as Efron [12] argues, m_0 in reality often appears different from the theoretical null. A number of factors can contribute to this phenomenon, inter-case correlation being one of them. This necessitates estimating m_0 from the data.

Estimating both m_0 and m_1 (as well as π) from data, however, is fraught with many dangers. The most severe of these is a lack of identifiability—exchanging the labels of the null and the alternative produces an identical marginal behavior for the z-scores. To counter this, one needs strong assumptions on the components of $m(z)$. For example, the *zero-assumption* in Efron [12] states that most of the z-scores near zero are null cases.

We consider a different scenario where such a segregation between the observables is deemed unlikely to occur, but the basic essence of Efron's zero-assumption prevails. Instead of focusing directly on the z-scores, we make a

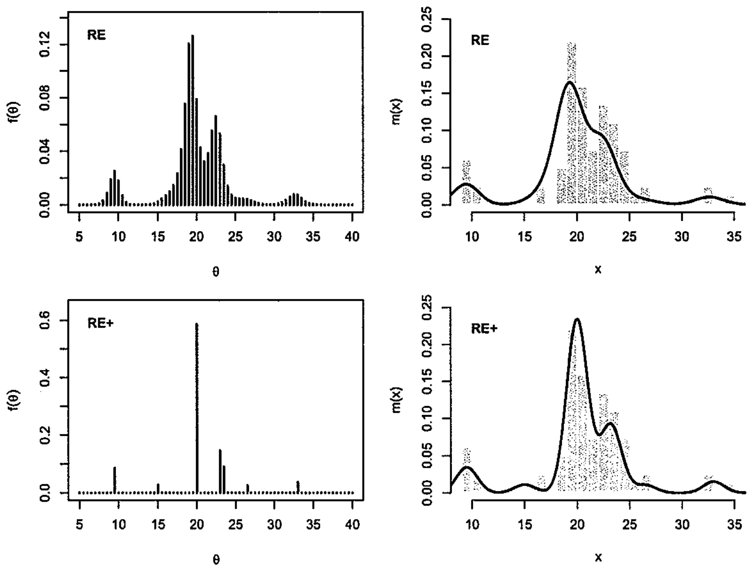

Fig 5.3: Plots of the estimated mixing (left) and mixture densities (right) for the recursive estimate (RE, top) and the extension (RE+, bottom) based on a simulated annealing optimization for the galaxy data in Example 5.2.

zero assumption about m_0 and m_1 as follows:

$$m_1 \text{ has strictly heavier tails than } m_0. \tag{5.2}$$

This is weaker than Efron's zero-assumption in the sense that the z-scores near zero are only *more likely* to have come from m_0 than from m_1.

A simple model that encodes our zero assumption (5.2) into the two-groups model (5.1) is

$$m(z) = \pi p(z|\theta_0, \sigma^2) + (1-\pi)\int p(z|\theta, \sigma^2) g(\theta)\, d\theta, \tag{5.3}$$

where $p(z|\theta, \sigma^2)$ is a Normal(θ, σ^2) density and the parameters θ_0, σ^2, π are unspecified, along with the mixing density $g(\theta)$. The following theorem shows that this model is identifiable in $(\theta_0, \sigma, \pi, g)$.

Theorem 5.3. *Let $\mathscr{D}$ denote the space of probability densities with respect to Lebesgue measure on $\mathbb{R}$; i.e., $\mathscr{D} = \{f \in \mathcal{L}_1(\mathbb{R}) : f \geq 0,\ \|f\|_1 = 1\}$. Then the map $M : \mathbb{R} \times (0,\infty) \times (0,1) \times \mathscr{D} \to \mathscr{D}$, given by*

$$M(\theta, \sigma, \pi, g)(z) = \pi p(z|\theta, \sigma^2) + (1-\pi)\int p(z|\theta', \sigma^2) g(\theta')\, d\theta' \tag{5.4}$$

is one-to-one.

Proof. Assume $M(\theta_1, \sigma_1, \pi_1, g_1) = M(\theta_2, \sigma_2, \pi_2, g_2)$. Then, in terms of characteristic functions, we must have

$$e^{-\sigma_1^2 t^2/2}\left[\pi_1 e^{it\theta_1} + (1-\pi_1)\psi_1(t)\right] = e^{-\sigma_2^2 t^2/2}\left[\pi_2 e^{it\theta_2} + (1-\pi_2)\psi_2(t)\right] \quad (5.5)$$

for every $t \in \mathbb{R}$, where ψ_j is the characteristic function of g_j, $j = 1, 2$. Since $g_j \in \mathscr{D}$, we know that

$$\psi_j(t) \to 0 \quad \text{as} \quad t \to \pm\infty. \quad (5.6)$$

Now, suppose $\sigma_1 > \sigma_2$. Choose a sequence $\{t_n\} \subset \mathbb{R}$ such that $t_n \to \infty$ and $e^{it_n\theta_2} = 1$ for all n. Then, for large enough n, (5.6) would imply that $\pi_2 + (1-\pi_2)\psi_2(t_n) \neq 0$. On rearranging the terms in (5.5) we get

$$e^{t_n^2(\sigma_1^2 - \sigma_2^2)/2} = \frac{\pi_1 e^{it_n\theta_1} + (1-\pi_1)\psi_1(t_n)}{\pi_2 + (1-\pi_2)\psi_2(t_n)}. \quad (5.7)$$

As $n \to \infty$, the left-hand side of (5.7) blows up to infinity while the right-hand side is bounded. Therefore, to avoid contradiction, we need $\sigma_1 \leq \sigma_2$; by symmetry, it follows that $\sigma_1 = \sigma_2$. With this equality, relation (5.5) easily leads to the equalities $\theta_1 = \theta_2$, $\pi_1 = \pi_2$ and $g_1 = g_2$, completing the proof. □

For a given configuration $\xi = (\theta_0, \sigma)$, one can write

$$m(z) = \int p(z|\theta, \xi) f(\theta)\, d\mu_\xi(\theta) \quad (5.8)$$

where $f = \pi\delta(\theta_0) + (1-\pi)g$ is a probability density with respect to the dominating measure $\mu_\xi = \delta(\theta_0) + \lambda$—a point mass at θ_0 plus Lebesgue measure. The formulation (5.8) is ideally suited for the estimation theory developed in Section 4. That is, by treating $\xi = (\theta_0, \sigma)$ as the "non-mixing" parameter we can employ the PARE+ method to produce estimates $\hat{\xi} = (\hat{\theta}_0, \hat{\sigma})$. The corresponding mixing density $f_{n,\xi} = \hat{\pi}\delta(\hat{\theta}_0) + (1-\hat{\pi})\hat{g}$ then provides estimates of π and g. The *empirical null* is Normal$(\hat{\theta}_0, \hat{\sigma}^2)$.

In implementing the above procedure, one needs to take care in specifying the initial estimate $f_0 = \pi_0\delta(\theta_0) + (1-\pi_0)g_0$ in the recursive algorithm. In spite of having large n, the initial guess π_0 can have a substantial effect on the final estimate $f_{n,\hat{\xi}}$ when one of the two groups is scarce. In most modern applications, the non-null group consists of a very small proportion of the total sample. This motivates us to include π_0 as one more (tuning) parameter and carry out the PARE+ maximization over the vector $\xi^+ = (\theta_0, \sigma, \pi_0)$.

Example 5.4. Gene expressions for four HIV+ males are compared to the same in four normal males in van't Wout, *et al.* [34]. The histogram in the left panel of Figure 5.4 shows the z-scores for 7680 genes under investigation. The z-score z_i was calculated by suitably transforming a two-sample t-statistic that compares the expression levels for gene i in the HIV+ patients against the normal subjects. The genes which had similar expression levels in the two groups were likely to produce z-scores close to zero, while the differentially expressed ones were likely the produce z-scores away from zero. The goal is to identify the genes that are differentially expressed.

On applying the PARE+ procedure described above to the HIV data set, we estimated the empirical null $\hat{m}_0$ to be a Normal$(-0.11, 0.74^2)$ density, shown in

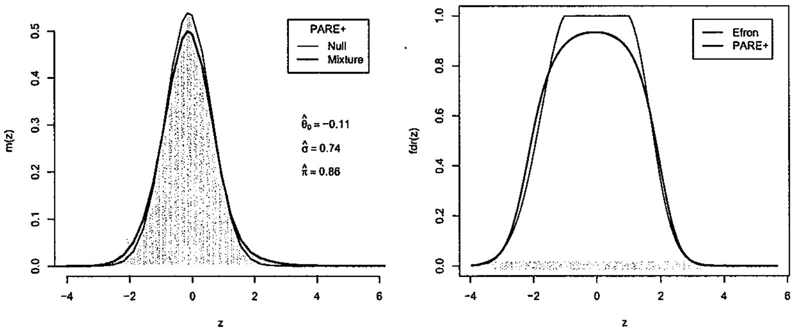

Fig 5.4: Plots associated with the HIV data set described in Example 5.4. Left: histogram of the data with the PARE+ empirical null (thin line) and mixture (thick line) overlaid. Right: Plots of the estimated local fdr—Efron (thin line) and PARE+ (thick line).

the left panel of Figure 5.4. The optimization was carried out numerically using the `optim` routine in R [26], each evaluation of $\widetilde{L}_n(\xi^+)$ was made based on a PARE derived from a fixed set of 25 permutations of the data. The optimum $\hat{\xi}^+ = (-0.11, 0.74, 0.57)$ was then used to estimate $\tilde{\pi} = 0.86$ and $\tilde{g}$ through a longer run of PARE based on 100 permutations. The estimated mixture $\tilde{m}$ is shown in the left panel of Figure 5.4. The right panel shows the estimated local false discovery rate (fdr)

$$\widehat{\mathrm{fdr}}(z) = \tilde{\pi}\tilde{m}_0(z)/\tilde{m}(z).$$

The thin line shows the fdr estimated with the `locfdr` package [26] due to Efron, Turnbull and Balasubramaniam.

Our estimate of the empirical null density closely matches the one reported in Efron [12], namely $\tilde{m}_0 = \mathrm{Normal}(-0.11, 0.75^2)$. But our estimate $\tilde{\pi} = 0.86$ is substantially lower than the $\tilde{\pi} = 0.93$. This is due to the difference in the zero-assumptions underlying the two methods. Our method allows a small fraction of non-null z-scores to be close to zero, while Efron rules out this possibility at the outset. Consequently, the two methods estimated quite different fdr values for the central z-scores. Note, however, the striking similarity to their treatment of the non-central z-scores. In fact, with a cut-off of $\mathrm{fdr}(z) < 0.2$, our method identifies 173 differentially expressed genes, closely matching the 160 genes identified by Efron.

6 Discussion

In this paper, we have shown that Newton's recursive estimate of a mixing distribution can do reasonably well even when the model is in some way incorrectly specified. Since the modeling error incurred by using an incorrectly specified mixture would generally be small relative to the error incurred by using a finite sample, this result greatly extends the potential applicability of the recursive estimate. Indeed, the "likelihood-based" extension can be applied in many important statistical problems that the original recursive estimate could not. While the numerical results in Section 5 are quite promising, they are also somewhat limited—deeper investigation into the performance of the (PA)RE+ algorithm in these and other problems is the focus of ongoing research.

It is interesting that the nonparametric MLE of the mixing distribution behaves similarly to Newton's estimate in the sense that they both try to minimize a KL divergence. Shyamalkumar [31] shows that the nonparametric MLE finds, for every *fixed* n, the mixing distribution that minimizes the KL divergence of the estimated mixture from the empirical distribution. This fixed-n "optimality" of the nonparametric MLE is a nice property, something that Newton's estimate lacks. There are, however, two important drawbacks to the nonparametric MLE. First, the resulting ML estimate of the mixing distribution is almost surely discrete [19], so if the mixing distribution f is known *a priori* to have a continuous density, then using the MLE for inference on f is nonsensical. Secondly, computation of the nonparametric MLE, in general, is quite non-trivial [19, 16, 36]. The recursive estimate, on the other hand, suffers from neither of these drawbacks. Therefore, by combining these nice properties of Newton's estimate with the result of Theorem 3.5, we see that the recursive estimate might serve well as an alternative to the nonparametric MLE.

In Section 1 it was suggested that $m_n(x)$ in (1.4) might be used in a general density estimation problem since almost any density can be well-approximated by a suitable mixture. As a matter of fact, Newton's recursive estimate seems to be a generalization of the popular kernel density estimates, so one would naturally expect that it too would perform well. Indeed, in preliminary experiments we have found that the recursive density estimate, with $p(x|\theta)$ a suitable normal density, performs comparably to and, in some cases, better than a Gaussian kernel estimate. By a "suitable normal density" we mean one with mean θ and standard deviation chosen according to the RE+ algorithm of Section 4. Further investigation into the performance of the RE+ estimate in the general nonparametric density estimation problem is required.

Acknowledgments

We would like to thank Professor Jun Xie for helpful discussions and Professor J.K. Ghosh for his never-ending support and guidance throughout this and other projects. We also acknowledge that Theorem 3.5 was an earlier conjecture of Professor J.K. Ghosh.

References

[1] Allison, D., Gadbury, G., Heo, M., Fernández, J., Lee, C., Prolla, T., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.* **39** 1–20.

[2] Billingsley, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York.

[3] Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg prodecure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, N. Balakrishnan, E. Peña, and M. Silvapulle, Eds. IMS, Beachwood, OH, 211–230.

[4] Carvalho, C., Lucas, J., Wang, Q., Chang, J., Nevins, J., and West, M. (2008). High-dimensional sparse factor modelling – applications in gene expression genomics. *J. Amer. Statist. Assoc.*. To appear.

[5] Clyde, M. A. and George, E. I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian inference in wavelet-based models.* Lecture Notes in Statist., Vol. **141**. Springer, New York, 309–322.

[6] Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158.

[7] Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions* Suppl. 1 205–237.

[8] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer, New York.

[9] Dykstra, R. (1985). An iterative procedure for obtaining *I*-projections onto the intersection of convex sets. *Ann. Probab.* **13** 975–984.

[10] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104.

[11] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103.

[12] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22.

[13] Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160.

[14] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.

[15] Ghosh, J. K. and Tokdar, S. T. (2006). Convergence and consistency of Newton's algorithm for estimating mixing distribution. In *Frontiers in statistics.* Imp. Coll. Press, London, 429–443.

[16] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

[17] Leroux, B. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360.

[18] Liese, F. and Vajda, I. (1987). *Convex statistical distances.* Teubner, Leipzig.

[19] Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications.* IMS, Haywood, CA.

[20] Martin, R. and Ghosh, J. K. (2008). Stochastic approximation and Newton's estimate of a mixing distribution. *Statist. Sci.*—To appear.

[21] McLachlan, G., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18** 413–422.

[22] Muller, P. and Vidakovic, B. (1998). Bayesian inference with wavelets: Density estimation. *J. Comput. Graph. Statist.* **7** 456–468.

[23] Newton, M. (2002). A nonparametric recursive estimator of the mixing distribution. *Sankhyā* **A 64** 306–322.

[24] Newton, M., Quintana, F., and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, D. Dey, P. Muller, and D. Sinha, Eds. Springer, New York.

[25] Quintana, F. and Newton, M. (2000). Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. *J. Comput. Graph. Statist.* **9** 711–737.

[26] R Development Core Team. (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

[27] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792.

[28] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20.

[29] Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York.

[30] Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* 617–624.

[31] Shyamalkumar, N. (1996). Cyclic I_0 projections and its applications in statistics. Tech. Rep. 96-24, Purdue University, Department of Statistics, West Lafayette, IN.

[32] Teicher, H. (1998). Strong laws for martingale differences and independent random variables. *J. Theoret. Probab.* **11** 979–995.

[33] Tokdar, S. T., Martin, R., and Ghosh, J. K. (2008). Consistency of a recursive estimate of mixing distributions. *Ann. Statist.*—accepted.

[34] van't Wout, A., Lehrma, G., Mikheeva, S., O'Keefe, G., Katze, M., Bumgarner, R., Geiss, G., and Mullins, J. (2003). Cellular gene expression upon human immunodeficiency virus type 1 injection of cd$+T-Cell lines. *Journal of Virology* **77** 1392–1402.

[35] Wang, L. and Dunson, D. B. (2007). Fast Bayesian inference in Dirichlet process mixture models. Preprint.

[36] Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 185–198.

[37] Woo, M.-J. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101** 1475–1486.