Electronic Journal of Statistics
Vol. 12 (2018) 1181–1209
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1417

An MM algorithm for estimation of a two component semiparametric density mixture with a known component

Zhou Shen and Michael Levine*

Department of Statistics, Purdue University, 250 N. University St, West Lafayette, IN 47907,
e-mail: shen171@purdue.edu; mlevins@purdue.edu

Zuofeng Shang

Department of Mathematical Sciences, IUPUI, 402 N. Blackford St., Indianapolis, IN 46202,
e-mail: shangzf@iu.edu

Abstract: We consider a semiparametric mixture of two univariate density functions where one of them is known while the weight and the other function are unknown. We do not assume any additional structure on the unknown density function. For this mixture model, we derive a new sufficient identifiability condition and pinpoint a specific class of distributions describing the unknown component for which this condition is mostly satisfied. We also suggest a novel approach to estimation of this model that is based on an idea of applying a maximum smoothed likelihood to what would otherwise have been an ill-posed problem. We introduce an iterative MM (Majorization-Minimization) algorithm that estimates all of the model parameters. We establish that the algorithm possesses a descent property with respect to a log-likelihood objective functional and prove that the algorithm, indeed, converges. Finally, we also illustrate the performance of our algorithm in a simulation study and apply it to a real dataset.

MSC 2010 subject classifications: Primary 62G07; secondary 62G99.
Keywords and phrases: Penalized smoothed likelihood, MM algorithm, regularization.

Received July 2017.

Contents

*The work of Michael Levine has been partially funded by the NSF-DMS grant 1208994

1. Introduction

We consider a general case of a two-component univariate mixture model where one component distribution is known while the mixing proportion and the other component distribution are unknown. Such a model can be defined at its most general as

$$g(x) = (1-p)f_0(x) + pf(x), \tag{1.1}$$

where f_0 is a known density component, while $p \in (0,1)$ and $f(x)$ are the unknown weight and the unknown density component, respectively. The semiparametric mixtures of density functions have been considered by now in a number of publications. The earliest seminal publications in this area are Hall and Zhou [16] and Hall et al. [15]. From the practical viewpoint, the model (1.1) is related to the multiple testing problem where p-values are uniformly distributed on $[0,1]$ under the null hypothesis but their distribution under the alternative is unknown. In the setting of model (1.1), this means that the known distribution is uniform while the goal is to estimate the proportion of the false null hypothesis p and the distribution of the p-values under the alternative. More detailed descriptions in statistical literature can be found in e.g. Efron [11] and Robin et al. [27]. Historically, whenever a two-component mixture model with a known component was considered, some assumptions were imposed on the unknown density function $f(x)$. Most commonly, it was assumed that an unknown distribution belongs to a particular parametric family. In such a case, Cohen [7] and Lindsay [22] used the maximum likelihood-based method to fit it; Day [10] used the minimum χ^2 method, while Lindsay and Basak [23] used the method of moments. Jin [20] and Cai and Jin [5] used empirical characteristic functions to estimate the unknown cumulative density function under a semiparametric normal mixture model. A somewhat different direction was taken by Bordes et al. [3] who considered a special case of the model (1.1) where the unknown component belonged to a location family. In other words, their model is defined as

$$g(x) = (1-p)f_0(x) + pf(x-\mu), \tag{1.2}$$

where f_0 is known while $p \in (0,1)$, the non-null location parameter μ and the pdf f that is symmetric around μ are the unknown parameters. The model of Bordes et al. [3] was motivated by the problem of detection of differentially expressed genes under two or more conditions in microarray data. Typically, a test statistic is built for each gene. Under the null hypothesis (which corresponds to a lack of difference in expression), such a test statistic has a *known* distribution (commonly, Student's or Fisher). Then, the response of thousands of genes is observed; such a response can be thought of as coming from a mixture of two distributions: the known distribution f_0 (under the null hypothesis) and the unknown distribution f under the alternative hypothesis. Once the parameters

p, μ, and f are estimated, we can estimate the probability that a gene belongs to a null component distribution conditionally on observations.

Bordes et al. [3] establish some sufficient identifiability conditions for their proposed model; they also suggest two estimation methods for it, both of which rely heavily on the fact that the density function of the unknown component is symmetric. A sequel paper, Bordes and Vandekerkhove [4], also establishes a joint central limit theorem for estimators that are based on one of these methods. There is no particular practical reason to make the unknown component symmetric and Bordes et al. [3] themselves note that "In our opinion, a challenging problem would be to consider model (1.2) without the symmetry assumption on the unknown component". This is the goal we set for ourselves in this manuscript. Our approach is based on, first, defining the (joint) estimator of f and p as a minimizer of the log-likelihood type objective functional of p and f. Such a definition is implicit in nature; however, we construct an MM (Majorization-Minimization) iterative algorithm that possesses a descent property with respect to that objective functional. Moreover, we also show that the resulting algorithm actually converges. Our simulation studies also show that the algorithm is rather well-behaved in practice.

Just as we were finishing our work, a related publication Patra and Sen [26] came to our attention. They also consider a two-component mixture model with one unknown component. Their identifiability approach is somewhat more general as our discussion mostly concerns sufficient conditions for specific functional classes containing the function f. They propose some general identifiability criteria for this model and obtain a separate estimator of the weight p; moreover, they also construct a distribution free finite sample lower confidence bound for the weight p. Patra and Sen [26] start with estimating parameter p first; then, a completely automated and tuning parameter free estimate of f can be constructed when f is decreasing. When f is not decreasing, one can start with e.g. estimating $\hat{g}$ based on observations $X_1, \ldots, X_n$; then, one can construct e.g. a kernel estimate of unknown f that is proportional to $\max(\hat{g} - (1-\hat{p})f_0, 0)$. In contrast to their approach, our approach estimates both p and f jointly and the algorithm works the same way regardless of the shape of f.

The rest of our manuscript is structured as follows. Section 2 discusses identifiability of the model (1.1). Section 3 introduces our approach to estimation of the model (1.1). Section 4 suggests an empirical version of the algorithm first introduced in Section 3 that can be implemented in practice. Section 5 provides simulated examples of the performance of our algorithm. Section 6 gives a real data example. Finally, Section 7 rounds out the manuscript with a discussion of our result and delineation of some related future research.

2. Identifiability

In general, the model (1.1) is not identifiable. In what follows, we investigate some special cases. For an unknown density function f, let us denote its mean μ_f and its variance σ_f^2. To state a sufficient identifiability result, we consider a

general equation

$$(1-p)f_0(x)+pf(x)=(1-p_1)f_0(x)+p_1f_1(x). \tag{2.1}$$

We also denote variance of the distribution $f(x)$ as a function of its mean μ_f as $V(\mu_f)$.

Lemma 2.1. *Consider the model (1.1) with the unknown density function f. Without loss of generality, assume that the first moment of f_0 is zero while its second moment is finite. We assume that the function f belongs to a set of density functions whose first two moments are finite, whose means are not equal to zero and that are all of the same sign; that is, $f\in\mathcal{F}=\{f:\int x^2f(x)\,dx<+\infty;\mu_f>0$ or $\mu_f<0\}$. Moreover, we assume that for any $f\in\mathcal{F}$ the function $G(\mu_f)=\frac{V(\mu_f)}{\mu_f}$ is strictly increasing. Then, the equation (2.1) has the unique solution $p_1=p$ and $f_1=f$.*

Proof. First, let us assume that the mean $\mu_f>0$. Then, the assumption of our lemma implies that the function $V:(0,\infty)\mapsto(0,\infty)$ is strictly increasing. Let us use the notation θ_0 for the second moment of f_0. If we assume that there are distinct $p_1\neq p$ and $f_1\neq f$ such that $(1-p)f_0(x)+pf(x)=(1-p_1)f_0(x)+p_1f_1(x)$, the following two moment equations are easily obtained,

$$\zeta=p_1\mu_{f_1}=p\mu_f \tag{2.2}$$

and

$$(p_1-p)\theta_0=\zeta(\mu_{f_1}-\mu_f)+p_1V(\mu_{f_1})-pV(\mu_f), \tag{2.3}$$

where $\zeta>0$. Our task is now to show that if (2.2) and (2.3) are true, then $p=p_1$ and $f=f_1$. To see this, let us assume $p_1>p$ (the case $p_1<p$ can be treated in exactly the same way). Then from the first equation we have immediately that $\mu_{f_1}<\mu_f$; moreover, since the function $G(\mu_f)$ is a strictly increasing one, then so is the function $\mu_f+G(\mu_f)$. With this in mind, we have

$$\mu_{f_1}+\frac{V(\mu_{f_1})}{\mu_{f_1}}<\mu_f+\frac{V(\mu_f)}{\mu_f}.$$

On the other hand, $(p_1-p)\theta_0\geq 0$ which implies

$$0\leq\zeta(\mu_{f_1}-\mu_f)+p_1V(\mu_{f_1})-pV(\mu_f)=\zeta(\mu_{f_1}-\mu_f)+\zeta\left(\frac{V(\mu_{f_1})}{\mu_{f_1}}-\frac{V(\mu_f)}{\mu_f}\right).$$

Therefore, this implies that

$$\mu_{f_1}+\frac{V(\mu_{f_1})}{\mu_{f_1}}\geq\mu_f+\frac{V(\mu_f)}{\mu_f}.$$

and we end up with a contradiction. Therefore, we must have $p=p_1$. This, in turn, implies immediately that $f=f_1$.

The case where $\mu_f<0$ proceeds similarly. Let us now consider the case where the variance function $V:(-\infty,0)\to(0,\infty)$ and is strictly monotonically

increasing. As a first step, again take $p_1 > p$. Clearly, the first moment equation is yet again (2.2) where now $\zeta < 0$. If $p_1 > p$, we now have $\mu_{f_1} > \mu_f$ and, due to the strict monotonicity of $G(\mu)$, we have $\mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}} > \mu_f + \frac{V(\mu_f)}{\mu_f}$. On the other hand, since $(p_1 - p)\theta_0 \geq 0$, we have

$$\begin{aligned} 0 &\leq \zeta(\mu_{f_1} - \mu_f) + p_1 V(\mu_{f_1}) - pV(\mu_f) \\ &= \zeta \left(\left\{ \mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}} \right\} - \left\{ \mu_f + \frac{V(\mu_f)}{\mu_f} \right\} \right). \end{aligned}$$

Because $\zeta < 0$, the above implies that $\left\{ \mu_{f_1} + \frac{V(\mu_{f_1})}{\mu_{f_1}} \right\} - \left\{ \mu_f + \frac{V(\mu_f)}{\mu_f} \right\} < 0$ which contradicts the assumption that the function $G(\mu)$ is strictly increasing. □

Remark 1. *To understand better what is going on here, it is helpful if we can suggest a more specific density class which satisfies the sufficient condition in Lemma 2.1. The form of Lemma 2.1 suggests one such possibility - a family of natural exponential families with power variance functions (NEF-PVF). For convenience, we give the definition due to Bar-Lev and Stramer [2]: "A natural exponential family (NEF for short) is said to have a power variance function if its variance function is of the form $V(\mu) = \alpha\mu^\gamma$, $\mu \in \Omega$, for some constants $\alpha \neq 0$ and γ, called the scale and power parameters, respectively". This family of distributions is discussed in detail in Bar-Lev and Enis [1] and Bar-Lev and Stramer [2]. In particular, they establish that the parameter space Ω can only be $\mathbb{R}$, $\mathbb{R}^+$ and $\mathbb{R}^-$; moreover, we can only have $\gamma = 0$ iff $\Omega = \mathbb{R}$. The most interesting for us property is that (see Theorem 2.1 from Bar-Lev and Stramer [2] for details) is that for any NEF-PVF, it is necessary that $\gamma \notin (-\infty, 0) \cup (0, 1)$; in other words, possible values of γ are 0, corresponding to the normal distribution, 1, corresponding to Poisson, and any positive real numbers that are greater than 1. In particular, the case $\gamma = 2$ corresponds to gamma distribution. Out of these choices, the only one that does not result in a monotonically increasing function $G(\mu)$ is $\gamma = 0$ that corresponds to the normal distribution; thus, we have to exclude it from consideration. With this exception gone, the NEF-PVF framework includes only density families with either strictly positive or strictly negative means; due to this, it seems a rather good fit for the description of the family of density functions f in the Lemma 2.1.*

Note that the exclusion of the normal distribution is also rather sensible from the practical viewpoint because it belongs to a location family; therefore, it can be treated in the framework of Bordes et al. [3]. More specifically, Proposition 1 *of Bordes et al. [3] suggests that, when $f(x)$ is normal, the equation (2.1) has at most two solutions if f_0 is an even pdf and at most three solutions if f_0 is not an even pdf.*

Remark 2. *It is also of interest to compare our Lemma 2.1 with the Lemma 4 of Patra and Sen [26] that also establishes an identifiability result for the model (1.1). The notions of identifiability that are considered in the two results differ: whereas we discuss the identifiability based on the first two moments, Lemma 4 of Patra and Sen [26] looks at a somewhat different definition of identifiability.*

At the same time, the interpretation given in the previous Remark, suggests an interesting connection. For example, the case where the unknown density function f is gamma corresponds to the power parameter of the NEF-PVF family being equal to 2. *According to our identifiability result Lemma 2.1, the mixture model (1.1) is, then, identifiable with respect to the first two moments. On the other hand, let us assume that the known density function f_0 is the standard normal. Since its support fully contains the support of any density from the gamma family, identifiability in the sense of Patra and Sen [26] now follows from their Lemma 4.*

Remark 3. *We only assumed that the first moment of f_0 is equal to zero for simplicity. It is not hard to reformulate the Lemma 2.1 if this is not the case. The proof is analogous.*

Corollary 2.1.1. *Consider the model (1.1) with the unknown density function f. We assume that the known density f_0 has finite first two moments and denote its first moment μ_{f_0}. We also assume that the function f belongs to a set of density functions whose first two moments are finite, and whose means are all either greater than μ_{f_0} or less than μ_{f_0}:*

$$f \in \mathcal{F} = \{f : \int x^2 f(x)\,dx < +\infty; \mu_f > \mu_{f_0} \text{ or } \mu_f < \mu_{f_0}\}.$$

Let us assume that $G(\mu_f) = \frac{V(\mu_f)}{\mu_f - \mu_{f_0}}$ is a strictly increasing function in μ_f for a fixed, known f_0. Then, the equation (2.1) has the unique solution $p_1 = p$ and $f_1 = f$.

3. Estimation

3.1. Possible interpretations of our approach

Let h be a positive bandwidth and K a symmetric positive-valued kernel function that is also a true density; as a technical assumption, we will also assume that K is continuously differentiable. The rescaled version of this kernel function is denoted $K_h(x) = K(x/h)/h$ for any $x \in \mathbb{R}$. We will also need a linear smoothing operator $\mathcal{S}f(x) = \int K_h(x-u)f(u)du$ and a nonlinear smoothing operator $\mathcal{N}f(x) = \exp(\mathcal{S}\log f(x))$ for any generic density function f. For simplicity, let us assume that our densities are defined on a closed interval, e.g. $[0,1]$. This assumption is here for technical convenience only when proving algorithmic convergence related results. Simulations in the Section 5 show that the algorithm works well also when the support of the density f is e.g. half the real line. In the future, we will omit these integration limits whenever doing so doesn't cause confusion. A simple application of Jensen's inequality and Fubini's theorem suggests that $\int \mathcal{N}f(x)\,dx \le \int Sf(x)\,dx = \int f(x)\,dx = 1$.

Our estimation approach is based on selecting p and f that minimize the following log-likelihood type objective functional:

$$\ell(p,f) = \int g(x) \log \frac{g(x)}{(1-p)f_0(x) + p\mathcal{N}f(x)}\,dx. \tag{3.1}$$

The reason the functional (3.1) is of interest as an objective functional is as follows. First, recall that $KL(a(x), b(x)) = \int \left[a(x) \log \frac{a(x)}{b(x)} + b(x) - a(x)\right] dx$ is a Kullback-Leibler distance between the two arbitrary positive integrable functions (not necessarily densities) $a(x)$ and $b(x)$; as usual, $KL(a, b) \geq 0$. This version of the Kullback-Leibler distance is a special case of the so-called Bregman divergence; one can find its definition in e.g. Eggermont et al. [12] p. 16. Note that the functional (3.1) can be represented as a penalized Kullback-Leibler distance between the target density $g(x)$ and the smoothed version of the mixture $(1-p)f_0(x) + p\mathcal{N}f(x)$; indeed, we can represent $\ell(p, f)$ as

$$\ell(p, f) = KL(g(x), (1-p)f_0(x) + p\mathcal{N}f(x)) + p\left\{1 - \int \mathcal{N}f(x)\, dx\right\}. \tag{3.2}$$

The quantity $1 - \int \mathcal{N}f(x)\, dx = \int [f(x) - \mathcal{N}f(x)]\, dx$ is effectively the penalty on the smoothness of the unknown density. Thus, the functional (3.1) can be interpreted as a penalized smoothed likelihood functional.

Of course, it is a matter of serious interest to find out if the optimization problem (3.2) has a solution at all. This problem can be thought of as a kind of generalized Tikhonov regularization problem; these problems have recently become an object of substantial interest in the area of ill-posed inverse problems. A very nice framework for these problems is described in the monograph Flemming [14] and we will follow it here. First of all, we define the domain of the operator $\mathcal{N}$ to be a set of square integrable densities, i.e. all densities on $[0, 1]$ that belong in $L_2[0, 1]$. We also define $L_2^+([0, 1])$ to be the set of all non-negative functions that belong to $L_2([0, 1])$. Define a nonlinear smoothing operator $A : \mathcal{D}(A) \subseteq L_2(D) \to L_2(D)$ as

$$Af(x) := (1-p)f_0(x) + p\mathcal{N}f(x)$$

where $\mathcal{D}(A) = \{f(x) : f \in L_2^+([0, 1]),\ f(x) \geq \eta > 0,\ \int_0^1 f(x)\, dx = 1,\ \exists F \in \mathbb{R}^+ \ s.t.\ ||f||_2 \leq F\}$. In optimization theory, A is commonly called a *forward operator*. Note that, as long as $||K||_2^2 := \int K^2(u)\, du < \infty$, it is easy to show that $\mathcal{N}f(x) \in L_2([0, 1])$ if $f(x) \in L_2([0, 1])$ and, therefore, this framework is justified.

Next, for any two functions $a(x)$ and $b(x)$, we define a *fitting functional* $Q : L_2([0, 1]) \times L_2([0, 1]) \to [0, \infty)$ as $Q(a(x), b(x)); = KL(a(x), b(x))$. Finally, we also define a non-negative *stabilizing functional* $\Omega : \mathcal{D}(\Omega) \subseteq L_2([0, 1]) \to [0, 1]$ as $\Omega(f) := \left\{1 - \int_0^1 \mathcal{N}f(x)\, dx\right\}$ where $\mathcal{D}(\Omega) = \mathcal{D}(A)$. Now, we are ready to define the minimization problem

$$T_{p,g}(f) = Q_p(g, Af) + p\Omega(f) \to \min \tag{3.3}$$

where p plays the role of *regularization parameter*. We use the subscript p for Q to stress the fact that the fitting functional is dependent on the regularization parameter; this doesn't seem to be common in optimization literature but we can still obtain the existence result that we need. The following set of assumptions is

needed to establish existence of the solution of this problem; although a version of these assumptions is given in Flemming [14] , we give them here in full for ease of reading.

Assumption 3.1. *Assumptions on* $A : \mathcal{D}(A) \subset L_2([0,1]) \to L_2([0,1])$*:*

1. *A is sequentially continuous with respect to the weak topology of the space* $L_2([0,1])$*, i.e. if* $f_k \rightharpoonup f$ *for* $f, f_k \in \mathcal{D}(A)$*, then we have* $A(f_k) \rightharpoonup A(f)$*.*
2. $\mathcal{D}(A)$ *is sequentially closed with respect to the weak topology on* $L_2([0,1])$*, that is* $f_k \rightharpoonup f$ *for* $\{f_k\} \in \mathcal{D}(A)$ *implies that* $f \in \mathcal{D}(A)$*.*

Assumptions on $Q : L_2([0,1]) \times L_2([0,1]) \to [0,\infty)$*:*

3. $Q_p(g,v)$ *is sequentially lower semi-continuous with respect to the weak topology on* $L_2([0,1]) \times L_2([0,1])$*, that is if* $p_k \to p$*,* $g_k \rightharpoonup g$ *and* $v_k \rightharpoonup v$*, then* $Q_p(g,v) \le \liminf_{k\to\infty} Q_{p_k}(g_k, v_k)$*.*
4. *If* $Q_p(g, v_k) \to 0$ *then there exists some* $v \in L_2([0,1])$ *such that* $v_k \rightharpoonup v$*.*
5. *If* $v_k \rightharpoonup v$ *and* $Q_p(g,v) < \infty$*, then* $Q_p(g, v_K) \to Q_p(g,v)$*.*

Assumptions on $\Omega : \mathcal{D}(A) \to [0,1]$*:*

6. Ω *is sequentially lower semicontinuous with respect to the weak topology in* $L_2([0,1])$*, that is, if* $f_k \rightharpoonup f$ *for* $f, f_k \in L_2([0,1])$*, we have* $\Omega(f) \le \liminf_{k\to\infty} \Omega(f_k)$*.*
7. *The sets* $M_\Omega(c) := \{f \in \mathcal{D}(A) : \Omega(f) \le c\}$ *are sequentially pre-compact with respect to the weak topology on* $L_2([0,1])$ *for all* $c \in \mathbb{R}$*, that is each sequence in* $M_\Omega(c)$ *has a subsequence that is convergent in the weak topology on* $L_2([0,1])$*.*

Lemma 3.1. *Assume that the kernel function* K *is bounded from above:* $K(x) \le \mathcal{K}$*. Then, the optimization problem* (3.3) *satisfies all of Assumption 3.1.*

Proof. We start with the Assumption 3.1(1). The space dual to $L_2([0,1])$ is again $L_2([0,1])$; therefore, the weak convergence $f_k \rightharpoonup f$ in $L_2([0,1])$ means that, for any $q \in L_2([0,1])$, we have $\int_0^1 f_k(x)q(x)\,dx \to \int_0^1 f(x)q(x)\,dx$ as $k \to \infty$. To show that the Assumption 3.1(1) is, indeed, true, we first note that $\{f_k\}$ and f are bounded away from 0 which tells

$$\int_0^1 |\log f_k(x) - \log f(x)| q(x)dx \le \int_0^1 |f_k(x) - f(x)| \frac{1}{\eta} q(x)dx \to 0$$

as $k \to \infty$ for some positive η that does not depend on k. Therefore, $f_k \rightharpoonup f$ implies $\log f_k \rightharpoonup \log f$. Second,

$$\begin{aligned} \int S \log f_k(x) q(x) dx &= \int q(x) \int_0^1 K_h(x-u) \log f_k(u) du\, dx \\ &= \int_0^1 \log f_k(u) \int K_h(x-u) q(x) dx\, du. \end{aligned}$$

Note that, since $\log f_k \rightharpoonup \log f$, and the function $\tilde{q}(u) = \int K_h(x-u)q(x)dx$ belongs to $L_2([0,1])$, we can claim that

$$\begin{aligned}\int S\log f_k(x)q(x)dx &\to \int_0^1 \log f(u)\int K_h(x-u)q(x)dx\,du \\ &= \int S\log f(x)q(x)dx\end{aligned}$$

as $k\to\infty$. In other words, we just established that $S\log f_k \rightharpoonup S\log f$ as $k\to\infty$. Moving ahead, we find out, using the Cauchy - Schwarz inequality that $\int_0^1 K_h(x-u)\log f_k(u)du < \int_0^1 K_h(x-u)f_k(u)du \le \mathcal{K}\int_0^1 f_k(u)du = \mathcal{K}$. The same is true for $f(x)$ and so $\int |exp\{S\log f_k(x)\} - exp\{S\log f(x)\}|g(x)dx \le \int |S\log f_k(x) - S\log f(x)| \le E\cdot g(x)dx \to 0$ where E is a positive constant that does not depend on k. Therefore, $f_k \rightharpoonup f$ finally implies $\mathcal{N}f_k \rightharpoonup \mathcal{N}f$ and thus $Af_k \rightharpoonup Af$.

Next, we need to prove that the Assumption 3.1(2) is also valid. To show that $\mathcal{D}(A)$ is sequentially closed, we select a particular function $q\equiv 1$ on $[0,1]$. Such a function clearly belongs to $L_2([0,1])$ and so we have $\int_0^1 f_k(x)q(x)\,dx \equiv \int_0^1 f_k(x)\,dx \to \int_0^1 f(x)\,dx$ as $k\to\infty$. Since we know that, for any k, we have $\int f_k(x)\,dx = 1$, it follows that $\int_0^1 f(x)dx = 1$ as well. It is not hard to check that other characteristics of $D(A)$ are preserved under weak convergence as well.

The fitting functional Q is a Kullback-Leibler functional; the fact that it satisfies Assumption 3.1(3)(4)(5) has been demonstrated several times in optimization literature concerned with variational regularization with non-metric fitting functionals. The details can be found in e.g. Flemming [13].

The sequential lower semi-continuity of the stabilizing functional Ω in Assumption 3.1(6) is guaranteed by the weak version of Fatou's Lemma. Indeed, let us define $\phi_k(x) = Sf_k(x) - \mathcal{N}f_k(x)$. Then, due to Jensen's inequality, $\{\phi_k\}$ is a sequence of non-negative measurable functions. We already know that $f_k \rightharpoonup f$ guarantees $\mathcal{N}f_k \rightharpoonup \mathcal{N}f$; therefore, we have $\phi_k \rightharpoonup \phi$ where $\phi(x) = Sf(x) - \mathcal{N}f(x)$. By the weak version of Fatou's lemma, we then have $\int \phi(x)\,dx \le \liminf_{k\to\infty}\int \phi_k(x)\,dx$, or equivalently, $\Omega(f) \le \liminf_{k\to\infty}\Omega(f_k)$. Therefore, $\Omega : \mathcal{D}(A)\to[0,1]$ is lower semi-continuous with respect to the weak topology on $L_2([0,1])$. Finally, the Assumption 3.1(7) is always true simply because $D(A)$ is a closed subset of a closed ball in $L_2([0,1])$; sequential Banach-Alaoglu theorem lets us conclude then that $M_\Omega(c)$ is sequentially compact with respect to the weak topology on $L_2([0,1])$. □

Finally, we can state the existence result. Note that in optimization literature sometimes one can see results of this nature under the heading of *well-posedness*, not existence; see, e.g. Hofmann et al. [18].

Theorem 3.2. (Existence) *For any mixture density $g(x)\in L_2([0,1])$ and any $0<p<1$, the minimization problem (3.3) has a solution. The minimizer $f^*\in\mathcal{D}(A)$ satisfies $T_{p,g}(f^*)<\infty$ if and only if there exists a density function $\bar{f}\in\mathcal{D}(A)$ with $Q_p(g,A\bar{f})<\infty$.*

Proof. Set $c := \inf_{f\in\mathcal{D}(A)} T_{p,g}(f) < \infty$. Note that $c < \infty$ due to existence of $\bar{f}$ and thus the trivial case of $c = \infty$ is excluded. Next, take a sequence $(f_k)_{k\in\mathbb{N}} \in \mathcal{D}(A)$ with $T_{p,g}(f_k) \to c$. Then

$$\Omega(f_k) \leq \frac{1}{p} T_{p,g}(f_k) \leq \frac{1}{p}(c+1) \tag{3.4}$$

for sufficiently large k and by the compactness of the sublevel sets of Ω there is a subsequence $(f_{k_l})_{l\in\mathbb{N}}$ converging to some $\tilde{f} \in \mathcal{D}(A)$. The continuity of A implies $Af_{k_l} \rightharpoonup A\tilde{f}$ and the lower semi-continuity of Q_p and Ω gives

$$T_{p,g}(\tilde{f}) \leq \liminf_{l\to\infty} T_{p,g}(f_{k_l}) = c, \tag{3.5}$$

that is, $\tilde{f}$ is a minimizer of $T_{p,g}$. □

3.2. Algorithm

Now we go back to introducing the algorithm that would search for unknown p and $f(x)$. The first result that we need is the following technical Lemma.

Lemma 3.3. *For any pdf $\widetilde{f}$ and any real number $\widetilde{p} \in (0,1)$,*

$$\begin{aligned} &\ell(\widetilde{p},\widetilde{f}) - \ell(p,f) \\ &\leq -\int g(x)\left[(1-w(x))\log\left(\tfrac{1-\widetilde{p}}{1-p}\right) + w(x)\log\left(\tfrac{\widetilde{p}\mathcal{N}\widetilde{f}(x)}{p\mathcal{N}f(x)}\right)\right]dx, \end{aligned} \tag{3.6}$$

where $w(x) = \frac{p\mathcal{N}f(x)}{(1-p)f_0(x)+p\mathcal{N}f(x)}$.

Proof of Lemma 3.3. The result follows by the following straightforward calculations:

$$\begin{aligned} \ell(\widetilde{p},\widetilde{f}) - \ell(p,f) &= -\int g(x)\log\left(\frac{(1-\widetilde{p})f_0(x)+\widetilde{p}\mathcal{N}\widetilde{f}(x)}{(1-p)f_0(x)+p\mathcal{N}f(x)}\right)dx \\ &= -\int g(x)\log\left((1-w(x))\frac{1-\widetilde{p}}{1-p} + w(x)\frac{\widetilde{p}\mathcal{N}\widetilde{f}(x)}{p\mathcal{N}f(x)}\right)dx \\ &\leq -\int g(x)\left[(1-w(x))\log\left(\frac{1-\widetilde{p}}{1-p}\right) + w(x)\log\left(\frac{\widetilde{p}\mathcal{N}\widetilde{f}(x)}{p\mathcal{N}f(x)}\right)\right], \end{aligned}$$

where the last inequality follows by convexity of the negative logarithm function. □

Suppose at iteration t, we get the updated pdf f^t and the updated mixing proportion p^t. Let $w^t(x) = \frac{p^t\mathcal{N}f^t(x)}{(1-p^t)f_0(x)+p^t\mathcal{N}f^t(x)}$, and define

$$\begin{aligned} p^{t+1} &= \int g(x)w^t(x)dx, \\ f^{t+1}(x) &= \alpha^{t+1}\int K_h(x-u)g(u)w^t(u)du, \end{aligned}$$

where α^{t+1} is a normalizing constant needed to ensure that f^{t+1} integrates to one. Then the following result holds.

Theorem 3.4. *For any $t \geq 0$, $\ell(p^{t+1}, f^{t+1}) \leq \ell(p^t, f^t)$.*

Proof of Theorem 3.4. By Lemma 3.3, for an arbitrary density function $\widetilde{f}$ and an arbitrary number $0 < \widetilde{p} < 1$,

$$\begin{aligned} &\ell(\widetilde{p}, \widetilde{f}) - \ell(p^t, f^t) \qquad (3.7)\\ &\leq -\int g(x) \left[(1 - w^t(x)) \log\left(\tfrac{1-\widetilde{p}}{1-p^t}\right) + w^t(x) \log\left(\tfrac{\widetilde{p}\mathcal{N}\widetilde{f}(x)}{p^t \mathcal{N} f^t(x)}\right)\right] dx. \end{aligned}$$

Let $(\widehat{p}, \widehat{f})$ be the minimizer of the right hand side of (3.7) with respect to $\widetilde{p}$ and $\widetilde{f}$. Note that the right-hand side becomes zero when $\widetilde{p} = p^t$ and $\widetilde{f} = f^t$; therefore, the minimum value of the functional on the right hand side must be less then or equal to 0. Therefore, it is clear that $\ell(\widehat{p}, \widehat{f}) \leq \ell(p^t, f^t)$. To verify that the statement of the Theorem 3.4 is true, it remains only to show that $(\widehat{p}, \widehat{f}) = (p^{t+1}, f^{t+1})$.

Note that the right hand side of (3.7) can be rewritten as

$$\begin{aligned} &-\int g(x)[(1 - w^t(x)) \log(1 - \widetilde{p}) + w^t(x) \log \widetilde{p}]dx\\ &-\int g(x) w^t(x) \log \mathcal{N}\widetilde{f}(x)dx + T, \end{aligned}$$

where the term T only depends on (p^t, f^t). The first integral in the above only depends on $\widetilde{p}$ but not on $\widetilde{f}$. It is easy to see that the minimizer of this first integral with respect to $\widetilde{p}$ is $\widehat{p} = \int g(x) w^t(x) dx$. The second integral, on the contrary, depends only on $\widetilde{f}$ but not on $\widetilde{p}$. It can be rewritten as

$$\begin{aligned} &-\int g(x) w^t(x) \log \mathcal{N}\widetilde{f}(x)dx\\ &= -\int\int g(x) w_t(x) K_h(x - u) \log \widetilde{f}(u) du dx\\ &= -\int \left(\int K_h(u - x) g(x) w^t(x) dx \right) \log \widetilde{f}(u) du\\ &= -\frac{1}{\alpha^{t+1}} \int f^{t+1}(u) \log \widetilde{f}(u) du\\ &= \frac{1}{\alpha^{t+1}} \int f^{t+1}(u) \log \frac{f^{t+1}(u)}{\widetilde{f}(u)} du - \frac{1}{\alpha^{t+1}} \int f^{t+1}(u) \log f^{t+1}(u) du. \end{aligned}$$

The first term in the above is the Kullback-Leibler divergence between f^{t+1} and $\widetilde{f}$ scaled by α^{t+1}, which is minimized at f^{t+1}, i.e., for $\widehat{f} = f^{t+1}$. Since the second term does not depend on $\widetilde{f}$ at all, we arrive at the needed conclusion. □

The above suggests that the following algorithm can be used to estimate the parameters of the model (1.1). First, we start with initial values p_0, f^0 at the step $t = 0$. Then, for any $t = 1, 2, \ldots$

- Define the weight

$$w^t(x) = \frac{p^t \mathcal{N} f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N} f^t(x)}. \tag{3.8}$$

- Define the updated probability

$$p^{t+1} = \int g(x) w^t(x) dx. \tag{3.9}$$

- Define

$$f^{t+1}(u) = \alpha^{t+1} \int K_h(u-x) g(x) w^t(x) dx. \tag{3.10}$$

Remark 4. *Note that the proposed algorithm is an MM (majorization minimization) algorithm. MM algorithms represent a generalization of the classical EM framework and are commonly used whenever optimization of a difficult objective function is best avoided and a series of simpler objective functions is optimized instead. The concept underlying MM algorithms was stated originally in Ortega and Reinboldt [25] in the context of line search methods. A general introduction to MM algorithms from the statistical viewpoint is available in, for example, Hunter and Lange [19]. As a first step, let* (p^t, f^t) *denote the current parameter values in our iterative algorithm. The main goal is to obtain a new functional* $b^t(p, f)$ *such that, when shifted by a constant, it majorizes* $\ell(p, f)$*. In other words, there must exist a constant* C^t *such that, for any* (p, f) $b^t(p, f) + C^t \geq \ell(p, f)$ *with equality when* $(p, f) = (p^t, f^t)$*. The use of* t *as a superscript in this context indicates that the definition of the new functional* $b^t(p, f)$ *depends on the parameter values* (p^t, f^t)*; these change from one iteration to the other.*

In our case, we define a functional

$$\begin{aligned} b^t(\tilde{p}, \tilde{f}) = &- \int g(x)[(1-\omega^t(x)) \log(1-\tilde{p}) + \omega^t(x) \log \tilde{p}]\, dx \\ &- \int g(x) \omega^t(x) \log \mathcal{N} \tilde{f}(x)\, dx. \end{aligned} \tag{3.11}$$

Note that the dependence on f^t *is through weights* ω^t*. From the proof of the Theorem 3.4, it follows that, for any argument* $(\tilde{p}, \tilde{f})$ *we have*

$$\ell(\tilde{p}, \tilde{f}) - \ell(p^t, f^t) \leq b^t(\tilde{p}, \tilde{f}) - b^t(p^t, f^t).$$

This means, that $b^t(\tilde{p}, \tilde{f})$ *is a majorizing functional; indeed, it is enough to select the constant* C^t *such that* $C^t = \ell(p^t, f^t) - b^t(p^t, f^t)$*. In the proof of the Theorem 3.4 it is the series of functionals* $b^t(\tilde{p}, \tilde{f})$ *(note that they are different at each step of iteration) that is being minimized with respect to* $(\tilde{p}, \tilde{f})$*, and not the original functional* $\ell(\tilde{p}, \tilde{f})$*. This, indeed, establishes that our algorithm is an MM algorithm.*

The following lemma shows that the sequence $\xi_t = \ell(p^t, f^t)$, defined by our algorithm, also has a non-negative limit (which is not necessarily a global minimum of $\ell(p, f)$).

Lemma 3.5. *There exists a finite limit of the sequence* $\xi_t = \ell(p^t, f^t)$ *as* $t \to \infty$*:*

$$L := \lim_{t\to\infty} \xi_t$$

for some $L \geq 0$*.*

Proof of Lemma 3.5. First, note that ξ_t is a non-increasing sequence for any integer t due to the Theorem 3.4. Thus, if we can show that it is bounded from below by zero, the proof will be finished. Then, the functional $\ell(p^t, f^t)$ can be represented as

$$\begin{aligned}
\ell(p^t, f^t) &= KL(g(x), (1-p^t)f_0(x) + p^t \mathcal{N} f^t(x)) + \int g(x)\,dx \\
&\quad - \int [(1-p^t)f_0(x) + p^t \mathcal{N} f^t(x)]\,dx \\
&= KL(g(x), (1-p^t)f_0(x) + p^t \mathcal{N} f^t(x)) + 1 \\
&\quad -(1-p^t) - p^t \int \mathcal{N} f^t(x)\,dx \\
&= KL(g(x), (1-p^t)f_0(x) + p^t \mathcal{N} f^t(x)) + p^t \left[1 - \int \mathcal{N} f^t(x)\,dx\right].
\end{aligned}$$

Now, since K is a proper density function, by Jensen's inequality,

$$\begin{aligned}
\mathcal{N} f^t(x) &= \exp\left\{\int K_h(x-u) \log f^t(u)\,du\right\} \\
&\leq \int K_h(x-u) f_t(u)\,du \equiv \mathcal{S} f^t(x).
\end{aligned}$$

Moreover, using Fubini's theorem, one can easily show that $\int \mathcal{S} f^t(x)\,dx = 1$ since f^t is a proper density function. Therefore, one concludes easily that $\int \mathcal{N} f^t(x)\,dx \leq \int \mathcal{S} f^t(x)\,dx = 1$. Thus, $\ell(p^t, f^t) \geq 0$ is non-negative due to non-negativity of the Kullback-Leibler distance. □

It is, of course, not clear directly from the Lemma 3.5 if the sequence (p^t, f^t), generated by this algorithm, also converges. Being able to answer this question requires establishing a lower semicontinuity property of the functional $\ell(p, f)$. Some additional requirements have to be imposed on the kernel function K in order to obtain the needed result that is given below. We denote Δ the domain of the kernel function K.

Theorem 3.6. *Let the kernel* $K : \Delta \to \mathbb{R}$ *be bounded from below and Lipschitz continuous with the Lipschitz constant* C_K*. Then, the minimizing sequence* (p^t, f^t) *converges to* (p_h^*, f_h^*) *that depends on the bandwidth* h *such that* $L = l(p_h^*, f_h^*)$*.*

Proof. We prove this result in two parts. First, let us introduce a subset of functions $B = \{\mathcal{S}\phi : 0 \leq \phi \in L_1^+(\Delta), \int \phi = 1\}$ where $L_1^+(\Delta)$ denotes a subset of all non-negative functions from $L_1(\Delta)$. Such a subset represents all densities on a closed compact interval that can be represented as linearly smoothed

integrable functions. Every function f_t generated in our algorithm except, perhaps, the initial one, can clearly be represented in this form. This is because, at every step of iteration, $f^{t+1}(x) = \alpha^{t+1} \int K_h(x-u) g(u) w^t(u)\, du = \int K_h(x-u)\phi(u)\, du$ where $\phi(u) = \alpha^{t+1} g(u) w^t(u)$. Moreover, we observe that $\int \phi(u)\, du = \alpha^{t+1} \int g(u) w^t(u)\, du = \alpha^{t+1} p^{t+1}$. Next, one concludes, by using Fubini theorem that, for any $t = 1, 2, \ldots$

$$\int f^{t+1}(x)\, dx = \alpha^{t+1} \int g(u) w^t(u) \left[\int K_h(x-u)\, dx \right] du = 1.$$

Since the iteration step t in the above is arbitrary, we established that $\alpha^t p^t = 1$ and, therefore, $\int \phi(u)\, du = 1$.

By definition of set B, it is clear that, as long as the kernel function is a proper density function (and so is non-negative), any $f \in B$ is non-negative and so every function in the set B is bounded from below. If the kernel function is Lipschitz continuous on Δ it is clearly bounded from above by some positive constant $M : \sup_{x \in \Delta} K(x) < M$. Thus, every function $f \in B$ satisfies $f(x) \leq M < \infty$. This implies that the set B is uniformly bounded. Also, by definition of set B, for any two points $x, y \in \Delta$ we have

$$|f(x) - f(y)| \leq \int |K_h(x-u) - K_h(y-u)| \phi(u)\, du \leq C_K |x-y|,$$

where the constant C_K depends on the choice of kernel K but not on the function f. This establishes the equicontinuity of the set B. Therefore, by Arzela-Ascoli theorem the set of functions B is a compact subset of $C(\Delta)$ with a sup metric.

Since for every $t = 2, 3, \ldots$ $f^t \in B$, by Arzela-Ascoli theorem we have a subsequence $f^{t_k} \to f_h^*$ as $k \to \infty$ uniformly over Ω. Since for every $t = 1, 2, \ldots$ p^t is bounded between 0 and 1, there exists, by Bolzano-Weierstrass theorem, a subsequence $p^{t_k} \to p_h^*$ as $k \to \infty$ in the usual Euclidean metric. Consider a Cartesian product space $\{(p, f)\}$ where every $p \in [0, 1]$ and $f \in C(\Delta)$. To define a metric on such a space we introduce an m-product of individual metrics for some non-negative m. This means that, if the first component space has a metric d_1 and the second d_2, the metric on the Cartesian product is $(|d_1|^m + |d_2|^m)^{1/m}$ for some non-negative m. For example, the specific case $m = 0$ corresponds to $|d_1| + |d_2|$ and $m = \infty$ corresponds to $\max(d_1, d_2)$. For such an m-product metric, clearly, we have a subsequence $(p^{t_k}, f^{t_k}) \to (p_h^*, f_h^*)$ that converges to (p_h^*, f_h^*) in the m-product metric. Without loss of generality, assume that the subsequence coincides with the whole sequence (p^t, f^t). Of course, such a sequence $(p^t, f^t) \in [0, 1] \times C(\Delta)$ for any t.

Now, that we know that there is always a converging sequence (p^t, f^t), we can proceed further. Since each f^t is bounded away from zero and from above, then so is the limit function $f_h^*(x)$ in the limit (p_h^*, f_h^*). This implies that $(p^t, \log f^t) \to (p_h^*, \log f_h^*)$ uniformly in the m-product topology as well and the same is true also for $(p^t, \mathcal{S} \log f^t)$. Analogously, the uniform convergence follows also in $(p^t, \mathcal{N} f^t) \to (p_h^*, \mathcal{N} f_h^*)$; moreover, $(1 - p^t) f_0 + p^t \mathcal{N} f^t \to (1 - p_h^*) f_0 + p_h^* \mathcal{N} f_h^*$ uniformly in the m-product topology. Since the function

$\psi(t) = -\log t + t - 1 \geq 0$, Fatou Lemma implies that

$$\int g(x)\psi((1-p_h^*)f_0(x) + p_h^*\mathcal{N}f_h^*(x))\,dx$$
$$\leq \liminf \int g(x)\psi((1-p^t)f_0(x) + p^t\mathcal{N}f^t(x))\,dx.$$

The lower semicontinuity of the functional $\ell(p, f)$ follows immediately and with it the conclusion of the Theorem 3.6. □

The above result can also be proved in the case where the densities involved have their support on the entire real line. To do so, it is necessary to impose constraints on the tails of these densities. The following result from the functional analysis forms the cornerstone of this analysis.

Lemma 3.7. *(Fréchet-Kolmogorov theorem) Let B be a bounded set in $L_p(\mathbb{R})$ with $p \in [1, \infty)$. The subset B is relatively compact if and only if the following properties hold for any function $f \in B$:*

1. $\lim_{r\to\infty} \int_{|x|>r} |f|^p = 0$ *uniformly on B,*
2. $\lim_{a\to 0} ||\tau_a f - f||_p = 0$ *uniformly on B.*

where $\tau_a f$ denotes the translation of f by a, that is, $\tau_a f(x) = f(x-a)$.

A very nice proof of this result can be found in e.g. an expository paper Hanche-Olsen and Holden [17]. Now we can formulate the following result.

Corollary 3.7.1. *Let all of the conditions of Theorem 3.6 be true but assume that the unknown density $f(x)$ and the known density $f_0(x)$ are now defined on the entire real line $\mathbb{R}$. Also, assume that $f_0(x)$ is bounded everywhere from above. Then, the convergence result of Theorem 3.6 remains correct.*

Proof. The only part of the proof of Theorem 3.6 that needs updating is that of establishing compactness of the subset B. Now, we need to establish its compactness as a subset of $L^1(\mathbb{R})$. To get this done, we will use Lemma 3.7. Fist, recall that our algorithm updates the density estimate at each step as $f^{t+1}(x) = \alpha^{t+1} \int K_h(x-u)g(u)w^t(u)\,du = \int K_h(x-u)\phi(u)\,du$ where $\phi(u) = \alpha^{t+1}g(u)w^t(u)$ is a density function belonging to $L_1^+(\mathbb{R})$, and so $\int \phi(u)\,du = 1$. Earlier, we showed that there exists a subsequence $p^{t_k} \to p_h^*$ by Bolzano-Weierstrass theorem. In order to use Lemma 3.7, we first show that at any step of iteration p_{t+1} is bounded away from zero. Indeed, from our algorithm we can see that $p^{t+1} = \int g(x)w^t(x)\,dx$; thus, if we show that the weight $w^t(x)$ is always bounded away from zero for any t, the probability p^{t+1} is bounded away from zero as well. Since the kernel function K is bounded from below, we can easily claim that for any $f \in B$ $f = \int K_h(x-u)\phi(u)\,du \geq \inf_{x\in\Omega} K_h(x-u) \int \phi(u)\,du = K_* > 0$. Next, by definition of the smoothing operator $\mathcal{N}f^t(x)$, and since $f^t \in B$ for any step of iteration t, we have $\mathcal{N}f^t(x) = \exp\{\int K_h(x-u)\log f_t(u)\,du\} \geq \exp\{\log K^* \int K_h(x-u)\,du\} = K^* > 0$ since the kernel function is a proper density function. Now, recall that at each step t the

weight is $w^t(x) = \frac{p^t \mathcal{N} f^t(x)}{(1-p^t) f_0(x) + p^t \mathcal{N} f^t(x)}$. If we assume that $f_0(x)$ is bounded from above, and since $\mathcal{N} f^t(x)$ is always bounded from above by M, we can conclude that the denominator of the integrand in the definition of $w^t(x)$ is bounded from above and, therefore, $w^t(x)$ is always bounded from below as long as p^t is bounded from below. Using the above argument, it is easy to see that as long as we start with $p^0 > 0$, p^t will stay bounded away from zero at every step of iteration. Thus, we can claim that the limit $p_h^* > 0$. Since $a_k p_k = 1$ for all k, there must be $a^{t_k} \to a_h^*$ and a^{t_k} is bounded from above by some $M_a > 0$.

Now, we can check the first condition in Lemma 3.7 for functions that belong to the set B. For any fixed mixture density $g(x)$, $\lim_{r\to\infty} \int_{|x|>r} g(x)\,dx = 0$; therefore, for any $\epsilon_g > 0$, there exists $r' > 0$ such that $\int_{|x|>r'} g(x)\,dx < \epsilon_g$. Since the kernel function K is a proper density function defined on a finite interval support Δ, for any $\epsilon_K > 0$ there exists $r > r'$ such that $\int_{|x|>r} K_h(x-u)\,dx < \epsilon_K$ for any $|u| \le r'$. This implies that

$$
\begin{aligned}
\int_{|x|>r} |f^{t_k}|\,dx &= \int_{|x|>r} \alpha^{t_k}\,dx \int_{-\infty}^{\infty} K_h(x-u) g(u) w^{t_k-1}(u)\,du \\
&\le \alpha^{t_k} \int_{|x|>r} dx \int_{-\infty}^{\infty} K_h(x-u) g(u)\,du \\
&= \alpha^{t_k} \int_{|x|>r} dx \left(\int_{|u|\le r'} K_h(x-u) g(u)\,du + \int_{|u|>r'} K_h(x-u) g(u)\,du \right) \\
&= \alpha^{t_k} \int_{|u|\le r'} g(u)\,du \int_{|x|>r} K_h(x-u)\,dx \\
&\quad + \alpha^{t_k} \int_{|u|>r'} g(u)\,du \int_{|x|>r} K_h(x-u)\,dx \\
&\le \alpha^{t_k} \left(\int_{|u|\le r'} \epsilon_K\, g(u)\,du + \int_{|u|>r'} g(u)\,du \right) \le M_a\,(\epsilon_K + \epsilon_g)\,,
\end{aligned}
$$

and so the first condition of the Lemma 3.7 has been verified. To verify the second condition we note first that, due to Lipschitz continuity of the kernel function and the fact that it is defined on a finite interval, we have

$$
\begin{aligned}
&\int_{-\infty}^{\infty} |f^{t_k}(x-a) - f^{t_k}(x)|\,dx \\
&\quad = \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} (K_h(x-a-u) - K_h(x-u)\phi(u))\,du \right| dx \\
&\quad \le \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} |(K_h(x-a-u) - K_h(x-u)|\,\phi(u))\,du \\
&\quad = \int_{-\infty}^{\infty} \phi(u)\,du \int_{-\infty}^{\infty} |(K_h(x-a-u) - K_h(x-u)|\,dx \\
&\quad \le \int_{-\infty}^{\infty} |a| C_K |\Delta| \phi(u)\,du = |a| C_K |\Delta|,
\end{aligned}
$$

and so for any $|a| < \rho$ we have the integral bounded from above by $C_K \rho |\Delta|$ that does not depend on the function $f \in B$. □

4. An empirical version of our algorithm

In practice, the number of observations n sampled from the target density function g is finite. This necessitates the development of the empirical version of our algorithm that can be implemented in practice. Many proof details here are similar to proofs of properties of the algorithm we introduced in the previous chapter. Therefore, we will be brief in our explanations. Denote the empirical cdf of the observations X_i, $i = 1, \ldots, n$ $G_n(x)$ where $G_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}$. Then, we define a functional

$$\begin{aligned} l_n(f,p) &= -\int \log((1-p)f_0(x) + p\mathcal{N}f(x))\, dG_n(x) \\ &\equiv -\sum_{i=1}^{n} \log((1-p)f_0(X_i) + p\mathcal{N}f(X_i)). \end{aligned}$$

The following analogue of the Lemma 3.3 can be easily established.

Lemma 4.1. *For any pdf* $\widetilde{f}$ *and* $\widetilde{p} \in (0,1)$,

$$\begin{aligned} &l_n(\widetilde{f},\widetilde{p}) - l_n(f,p) \\ &\le -\int \left[(1-w(x)) \log\left(\frac{1-\widetilde{p}}{1-p}\right) + w(x) \log\left(\frac{\widetilde{p}\mathcal{N}\widetilde{f}(x)}{p\mathcal{N}f(x)}\right) \right] dG_n(x), \end{aligned}$$

where the weight $w(x) = \frac{p\mathcal{N}f(x)}{(1-p)f_0(x)+p\mathcal{N}f(x)}$.

The proof is omitted since it is almost exactly the same as the proof of the Lemma 3.3.

Now we can define the empirical version of our algorithm. Denote (p_n^t, f_n^t) values of the density f and probability p at the iteration step t. Define the weights as $w_n^t(x) = \frac{p_n^t \mathcal{N} f_n^t(x)}{(1-p_n^t)f_0(x)+p_n^t \mathcal{N} f_n^t(x)}$. We use the subscript n everywhere intentionally to stress that these quantities depend on the sample size n. For the next step, define (p_n^{t+1}, f_n^{t+1}) as

$$\begin{aligned} p_n^{t+1} &= \int w_n^t(x) dG_n(x) = \frac{1}{n}\sum_{i=1}^{n} w_n^t(X_i) \\ f_n^{t+1}(x) &= \alpha_n^{t+1} \int K_h(x-u) w_n^t(u) dG_n(u) \\ &= \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} K_h(x-X_i) w_n^t(X_i), \end{aligned}$$

where α_n^{t+1} is a normalizing constant such that f_n^{t+1} is a valid pdf. Since $\int K_h(X_i - u)du = 1$ for $i = 1, \ldots, n$, we get

$$1 = \int f_n^{t+1}(u)du = \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} w_n^t(X_i),$$

and hence,

$$\alpha_n^{t+1} = \frac{n}{\sum_{i=1}^{n} w_n^t(X_i)}.$$

The following result establishes the descent property of the empirical version of our algorithm.

Theorem 4.2. *For any $t \geq 0$, $\ell_n(p_n^{t+1}, f_n^{t+1}) \leq \ell_n(p_n^t, f_n^t)$.*

The proof of this result follows very closely the proof of the Theorem 3.4 and is also omitted for brevity.

Remark 5. *As before, the empirical version of the proposed algorithm is an MM (majorization - minimization) algorithm that represents a generalization of the classical EM setting. More specifically, we can show that there exists another functional $b_n^t(p, f)$ such that, when shifted by a constant, it majorizes $l_n(p, f)$. It is easy to check that such a functional is*

$$\begin{aligned} b_n^t(\tilde{p}, \tilde{f}) &= -\int [(1 - \omega_n^t(x)) \log(1 - \tilde{p}) + \omega_n^t(x) \log \tilde{p}]\, dG_n(x) \\ &\quad - \int \omega_n^t(x) \log \mathcal{N}\tilde{f}(x)\, dG_n(x). \end{aligned} \tag{4.1}$$

Note that in the proof of the Theorem 4.2 it is the series of functionals $b_n^t(\tilde{p}, \tilde{f})$ that is being minimized with respect to $(\tilde{p}, \tilde{f})$, and not the original functional $l_n(\tilde{p}, \tilde{f})$.

Remark 6. *Note also that this algorithm can be easily generalized to the multivariate case. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the unknown density function and $f_0 : \mathbb{R}^d \to \mathbb{R}$ be the unknown one. We assume that the target density $g : \mathbb{R}^d \to \mathbb{R}$ is a two-component mixture of the unknown component f and the known component f_0 with the weight $0 < p < 1$. Our data consist of sample $\vec{X}_1, \ldots, \vec{X}_n$ generated by g. Since Lemma 4.1 and Theorem 4.2 of our manuscript only depend on some fairly basic tools, such as Jensen's inequality and convexity of the negative logarithm function, both of them remain true in the multivariate case and the following algorithm can be defined.*

Denote (p_n^t, f_n^t) values of the density f and probability p at the iteration step t. We use the subscript n everywhere intentionally to stress that these quantities depend on the sample size n. For the next step, define (p_n^{t+1}, f_n^{t+1}) as

$$\begin{aligned} p_n^{t+1} &= \frac{1}{n} \sum_{i=1}^{n} w_n^t(\mathbf{X}_i) \\ f_n^{t+1}(\mathbf{x}) &= \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i) w_n^t(\mathbf{X}_i), \end{aligned}$$

where $w_n^t(\mathbf{x}) = \frac{p_n^t \mathcal{N} f_n^t(\mathbf{x})}{(1-p_n^t) f_0(\mathbf{x}) + p_n^t \mathcal{N} f_n^t(\mathbf{x})}$ *is the weight (probability) that an observation* $\mathbf{x}$ *has been generated by an unknown component density, and* α_n^{t+1} *is a normalizing constant such that* f_n^{t+1} *is a valid density function. Since* $\int K_h(\mathbf{X}_i - \mathbf{u}) d\mathbf{u} = 1$ *for* $i = 1, \ldots, n$*, and we assume that* K *is a symmetric density function, we find that*

$$1 = \int f_n^{t+1}(\mathbf{u}) d\mathbf{u} = \frac{\alpha_n^{t+1}}{n} \sum_{i=1}^{n} w_n^t(X_i), \tag{4.2}$$

and hence,

$$\alpha_n^{t+1} = \frac{n}{\sum_{i=1}^{n} w_n^t(\mathbf{X}_i)}. \tag{4.3}$$

It can be verified immediately that the resulting algorithm possesses the descent property and is an MM algorithm, as before.

As before, we can also show that the sequence $\ell_n(p_n^t, f_n^t)$ generated by our algorithm does not only possess the descent property but is also bounded from below.

Lemma 4.3. *There exists a finite limit of the sequence* $\xi_n^t = \ell_n(p_n^t, f_n^t)$ *as* $t \to \infty$*:*

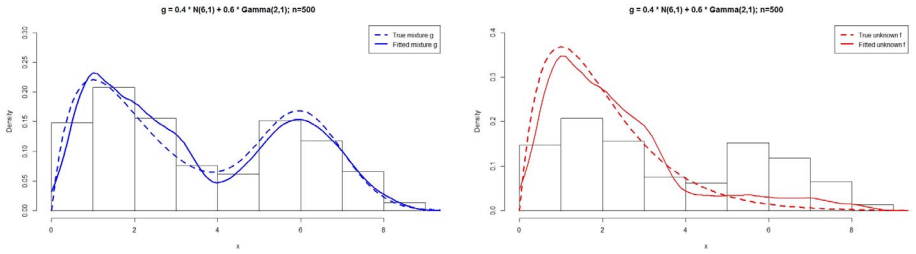
$$L_n = \lim_{t \to \infty} \xi_n^t$$

for some $L_n \geq 0$*.*

The proof is almost exactly the same as the proof of the Lemma 3.5 and is omitted in the interest of brevity. Finally, one can also show that the sequence (p_n^t, f_n^t) generated by our algorithm converges to (p_n^*, f_n^*) such that $L_n = l_n(p_n^*, f_n^*)$. The proof is almost the same as that of the Theorem 3.6 and is omitted for conciseness.

5. Simulations and comparison

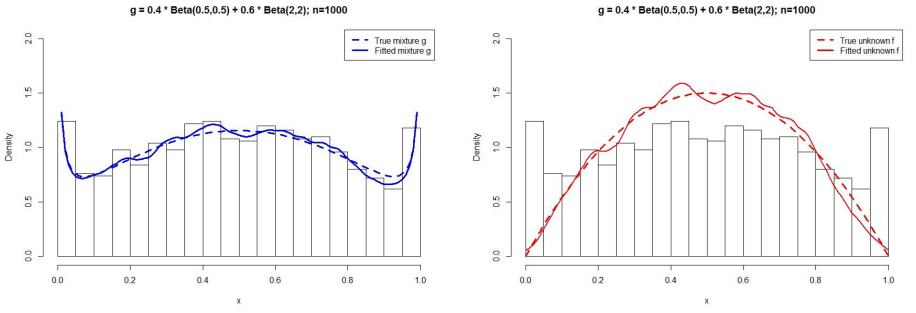
In this section, we will use the notation $I_{[x>0]}$ for the indicator function of the positive half of the real line and $\phi(x)$ for the standard Gaussian distribution. For our first experiment, we generate n independent and identically distributed observations from a two component normal gamma mixture with the density $g(x)$ defined as $g(x) = (1-p) f_0(x) + p f(x)$. Thus, the known component is $f_0(x) = \frac{2}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) I_{[x>0]}$ while the unknown component is $Gamma(\alpha, \beta)$, i.e., $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} I_{[x>0]}$. Note that we truncate the normal distribution so that it stays on the positive half of the real line. We choose the sample size $n = 500$, the probability $p = 0.6$, $\mu = 6$, $\sigma = 1$, $\alpha = 2$ and $\beta = 1$. The initial weight is $p_0 = 0.2$ and the initial assumption for the unknown component distribution is $Gamma(4, 2)$. The rescaled triangular function $K_h(x) = \frac{1}{h}\left(1 - \frac{|x|}{h}\right) I(|x| \leq h)$ is used as the kernel function. We use a



(a) Fitted mixture density (b) Fitted unknown component density

FIG 1. *Mixture of Gaussian (6,1) and Gamma (2,1)*

fixed bandwidth throughout the sequence of iterations and this fixed bandwidth is selected according to the classical Silverman's rule of thumb that we describe here briefly for completeness; for more details, see Silverman [28]. Let SD and IQR be the standard deviation and interquartile range of the data, respectively. Then, the bandwidth is determined as $h = 0.9 \min\left\{SD, \frac{IQR}{1.34}\right\} n^{-1/5}$. We use the absolute difference $|p_n^{t+1} - p_n^t|$ as a stopping criterion; at every iteration step, we check if this difference is below a small threshold value d that depends on required precision. If it is, the algorithm is stopped. The analogous rule has been described for classical parametric mixtures in McLachlan and Peel [24]. In our setting, we use the value $d = 10^{-5}$. The computation ends after 259 iterations, with an estimate $\hat{p} = 0.6661$; the Figure 1(a) shows the true and estimated mixture density function $g(x)$ while the Figure 1(b) shows both true and estimated second component density f. Both figures show a histogram of the observed target distribution $g(x)$ in the background. Both the fitted mixture density $\hat{g}(x)$ and the fitted unknown component density function $\hat{f}(x)$ are quite close to their corresponding true density functions everywhere.

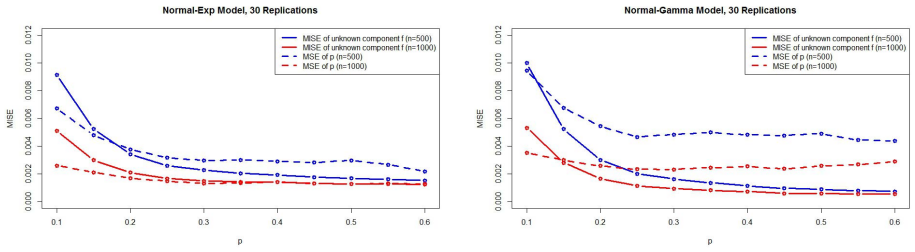


(a) Fitted mixture density (b) Fitted unknown component density

FIG 2. *Mixture of Beta (0.5,0.5) and Beta (2,2)*

In the second experiment, we generate n i.i.d. observations from a two component beta - beta mixture with the density $g(x)$ defined as $g(x) = (1-p)f_0(x) + pf(x)$. The known component is $f_0(x) = Beta(0.5, 0.5)$, while the unknown component is $f(x) = Beta(2, 2)$ with both having a $[0, 1]$ support. The sample size is set as $n = 500$ and the probability weight $p = 0.6$. We assume that the starting value of the probability weight is $p_0 = 0.3$ and the initial assumption for the unknown component distribution is $Beta(4, 4)$. The rescaled triangular kernel $K_h(x) = \frac{1}{h}\left(1 - \frac{|x|}{h}\right) I(|x| \leq h)$ is used with a fixed bandwidth $h = ...$ The algorithm is stopped when the absolute difference $|p_n^{t+1} - p_n^t| < 10^{-5}$. The computation ends after around 80 iterations, with an estimate $\hat{p} = 0.601$; the Figure 2(a) shows the true and estimated mixture density function $g(x)$ while the Figure 2(b) shows both true and estimated second component density f. Both figures show a histogram of the observed target distribution $g(x)$ in the background. Note that both the fitted mixture density $\hat{g}(x)$ and the fitted unknown component density function $\hat{f}(x)$ are quite close to corresponding true density functions.

We also analyze performance of our algorithm in terms of the mean squared error (MSE) of estimated weight $\hat{p}$ and the mean integrated squared error (MISE) of $\hat{f}$. We will use two models for this purpose. The first model is the normal exponential model where the (known) normal component is the same as before while the second (unknown) component is an exponential density function $f(x) = \lambda e^{-\lambda x} I_{[x>0]}$ with $\lambda = 0.5$; the value of p used is $p = 0.6$. The second model is the same normal-gamma model as before. For each of the two models, we plot MSE of $\hat{p}$ and MISE of $\hat{f}$ against the true p for sample sizes $n = 500$ and $n = 1000$. Here, we use 30 replications. The algorithm appears to show rather good performance even for the sample size $n = 500$. Note that MISE of the unknown component f seems to decrease with the increase in p. Possible reason for this is the fact that, the larger p is, the more likely it is that we are sampling from the unknown component and so the number of observations that are actually generated by f grows; this seems to explain better precision in estimation of f when p is large.



(a) Normal-Exponential mixture (b) Normal-Gamma mixture

FIG 3. *MISE of $\hat{f}$ and MSE of $\hat{p}$*

Another important issue in practice is, of course, the bandwidth selection. Earlier, we simply used a fixed bandwidth selected using the classical Silverman's rule of thumb. In general, when the unknown density is not likely to be normal, the use of Silverman's rule may be a somewhat rough approach. Moreover, in an iterative algorithm, every successive step of iteration brings a refined estimate of the unknown density component; therefore, it seems a good idea to put this knowledge to use. Such an idea was suggested earlier in Chauveau et al. [6]. Here we suggest using a version of the K-fold cross validation method specifically adopted for use in an iterative algorithm. First, let us suppose we have a sample $X_1, \ldots, X_n$ of size n; we begin with randomly partitioning it into K approximately equal subsamples. For ease of notation, we denote each of these subsamples X^k, $k = 1, \ldots, K$. Randomly selecting one of the K subsamples, it is possible to treat the remaining $K-1$ subsamples as a training dataset and the selected subsample as the validation dataset. We also need to select a grid of possible bandwidths. To do so, we compute the preliminary bandwidth h_s first according to the Silverman's rule of thumb; the bandwidth grid is defined as lying in an interval $[h_s - l, h_s + l]$ where $2*l$ is the range of bandwidths we plan to consider. Within this interval, each element of the bandwidth grid is computed as $h_i = h_s \pm \frac{i}{M} l$, $i = 0, 1, \ldots, M$ for some positive integer M. At this point, we have to decide whether a fully iterative bandwidth selection procedure is necessary. It is worth noting that a fully iterative bandwidth selection algorithm leads to the situation where the bandwidth changes at each step of iteration. This, in turn, implies that the monotonicity property of our algorithm derived in Theorem 4.2 is no longer true. To reconcile these two paradigms, we implement the following scheme. As in earlier simulations, we use the triangular smoothing kernel. First, we iterate a certain number of times T to obtain a reasonably stable estimate of the unknown f; if we do it using the full range of the data, we denote the resulting estimate $\hat{f}_{nh}^T(x) = \frac{\alpha_n^T}{n} \sum_{i=1}^n K_h(x - X_i) w_n^{T-1}(X_i)$. Integrating the resulting expression, we can obtain the squared L_2-norm of $\hat{f}_{nh}^T(x)$ as

$$||\hat{f}_{nh}^T||_2^2 = \int \left[\frac{\alpha_n^T}{n} \sum_{i=1}^n w_n^{T-1}(X_i) K_h(x - X_i) \right]^2 dx.$$

For each of K subsamples of the original sample, we can also define a "leave-kth subsample out" estimator of the unknown component f as $\hat{f}_{nh,-X_k}^T(x)$, $k = 1, \ldots, K$ obtained after T steps of iteration. At this point, we can define the CV optimization criterion as (see, for example, Eggermont et al. [12])as

$$CV(h) = ||\hat{f}_{nh}^T||_2^2 - \frac{2}{n} \sum_{k=1}^K \sum_{x_i \in X_k} \hat{f}_{nh,-X_k}^T(x_i).$$

Finally, we select

$$h^* = argmin\, CV(h)$$

as a proper bandwidth. Now, we fix the bandwidth h^* and keep it constant beginning with the iteration step $T+1$ until the convergence criterion is achieved

and the process is stopped. An example of a cross validation curve of $CV(h)$ is given in Figure 4. Here, we took a sample of size 500 from a mixture model with a known component of $N(6,1)$, an unknown component of $Gamma(2,1)$ and a mixing proportion $p = 0.5$; we also chose $K = 50$, $l = 0.4$, $M = 10$, and $T = 5$. We tested the possibility of using larger number of iterations before selecting the optimal bandwidth h; however, already $T = 10$ results in the selection of h^* close to zero. We believe that the likeliest reason for that is the overfitting of the estimate of the unknown component f. The minimum of $CV(h)$ is achieved at around $h = 0.68$. Using this bandwidth and running the algorithm until the stopping criterion is satisfied, gives us the estimated mixing proportion $\hat{p} = 0.497$. As a side remark, in this particular case the Silverman's rule of thumb gives a very similar estimated bandwidth $\hat{h} = 0.72$.

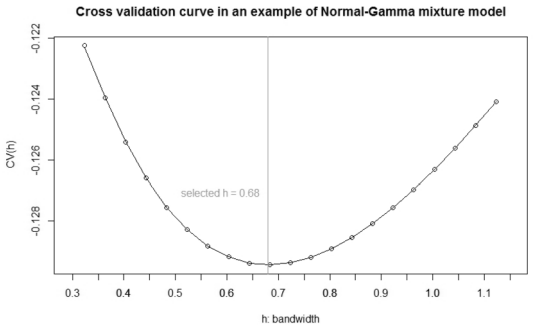


FIG 4. *A plot of $CV(h)$ used for bandwidth selection*

As a last step, we want to compare our method with the symmetrization method of Bordes et al. [3]. To do this, we will use a normal-normal model since the method of Bordes et al. [3] is only applicable when an unknown component belongs to a location family. Although such a model does not satisfy the sufficient criterion of the Lemma 2.1, it satisfies the necessary and sufficient identifiability criterion given in Lemma 4 of Patra and Sen [26] (see also Remark 3 from the Supplement to Patra and Sen [26] for even clearer statement about identifiability for normal-normal models in our context); therefore, we can use it for testing purposes. The known component has Gaussian distribution with mean 0 and standard deviation 1, the unknown has mean 6 and standard deviation 1, and we also consider two possible choices of mixture weight, $p = 0.3$ and $p = 0.5$. The results for two different sample sizes, $n = 500$, and $n = 1000$, and 200 replications, are given below in Tables 1 and 2. Each estimate is accompanied by its standard deviation in parentheses. Note that the proper expectation here is that our method should perform similarly to the method of Bordes et al. [3] but not beat it, for several reasons. First, the mean of the unknown Gaussian distribution is directly estimated as a parameter in the symmetrization method, while it is the nonparametric probability density function that is directly estimated by our method. Thus, in order to calculate the mean of the second component, we

Table 1
Mean(SD) of estimated p/μ obtained by the symmetrization method

$K = 200$	$n = 500$	$n = 1000$
$p = 0.3/\mu = 6$	0.302(0.022)/5.989(0.095)	0.302(0.016)/5.998(0.064)
$p = 0.5/\mu = 6$	0.502(0.024)/5.999(0.067)	0.502(0.017)/6.003(0.050)

Table 2
Mean(SD) of estimated p/μ obtained by our algorithm

$K = 200$	$n = 500$	$n = 1000$
$p = 0.3/\mu = 6$	0.315(0.024)/5.772(0.238)	0.312(0.018)/5.818(0.178)
$p = 0.5/\mu = 6$	0.516(0.026)/5.855(0.155)	0.512(0.018)/5.883(0.117)

have to take an extra step when using our method and employ numerical integration. This is effectively equivalent to estimating a functional of an unknown (and so estimated beforehand) density function; therefore, somewhat lower precision of our method when estimating the mean, compared to symmetrization method, where the mean is just a Euclidean parameter, should be expected. Second, when using symmetrization method, we followed an acceptance/rejection procedure exactly as in [3]. That procedure amounts to dropping certain "bad" samples whereas our method keeps all the samples. Third, the method of [3], when estimating an unknown component, uses the fact that this component belongs to a location family - something that our method, more general in its assumptions, does not do. Keeping all of the above in mind, we can see from Tables 1 and 2 that both methods produce comparable results, especially when the sample size is $n = 1000$. Also, as explained above, it does turn out that our method is practically as good as the method of [3] when it comes to estimating probability p and slightly worse when estimating the mean of the unknown component. However, even when estimating the mean of the unknown component, increase in sample size from 500 to 1000 reduces the difference in performance substantially.

6. A real data example

The acidification of lakes in parts of North America and Europe is a serious concern. In 1983, the US Environmental Protection Agency (EPA) began the EPA National Surface Water Survey (NSWS) to study acidification as well as other characteristics of US lakes. The first stage of NSWS was the Eastern Lake Survey, focusing on particular regions in Midwestern and Eastern US. Variables measured include acid neutralizing capacity (ANC), pH, dissolved organic carbon, and concentrations of various chemicals such as iron and calcium. The sampled lakes were selected systematically from an ordered list of all lakes appearing on $1 : 250,000$ scale US Geological Survey topographic maps. Only surface lakes with the surface area of at least 4 hectares were chosen.

Out of all these variables, ANC is often the one of greatest interest. It describes the capability of the lake to neutralize acid; more specifically, low (negative) values of ANC can lead to a loss of biological resources. We use a dataset

containing, among others, ANC data for a group of 155 lakes in north-central Wisconsin. This dataset has been first published in Crawford et al. [9] in Table 1 and analyzed in the same manuscript. Crawford et al. [9] argue that this dataset is rather heterogeneous due to the presence of lakes that are very different in their ANC within the same sample. In particular, seepage lakes, that have neither inlets nor outlets tend to be very low in ANC whereas drainage lakes that include flow paths into and out of the lake tend to be higher in ANC. Based on this heterogeneity, Crawford et al. [9] suggested using an empirical mixture of two lognormal densities to fit this dataset. Crawford [8] also considered that same dataset; they suggested using a modification of Laplace method to estimate posterior component density functions in the Bayesian analysis of a finite lognormal mixture. Note that Crawford [8] viewed the number of components in the mixture model as a parameter to be estimated; their analysis suggests a mixture of either two or three components.

The sample histogram for the ANC dataset is given on Figure 1 of [8]. The histogram is given for a log transformation of the original data $log(ANC + 50)$. Crawford [8] selected this transformation to avoid numerical problems arising from maximization involving a truncation; the choice of 50 as an additive constant is explained in more detail in [8]. The empirical distribution is clearly bimodal; moreover, it exhibits a heavy upper tail. This is suggestive of a two-component mixture where the first component may be Gaussian while the other is defined on the positive half of the real line and has a heavy upper tail. We estimate a two-component density mixture model for this empirical distribution using two approaches. First, we follow the Bayesian approach of [8] using the prior settings of Table 4 in that manuscript. Switching to our framework next, we assume that the normal component is a known one while the other one is unknown. For the known normal component, we assume the mean $\mu_1 = 4.375$ and $\sigma_1 = 0.416$; these are the estimated values obtained in [8] under the assumption of two component Gaussian mixture for the original (not log transformed) data and given in their Table 4. Next, we apply our algorithm in order to obtain an estimate of the mixture proportion and a non-parametric estimate of the unknown component to compare with respective estimates in [8]. We set the initial value of the mixture proportion as $p^0 = 0.3$ and the initial value of the unknown component as a normal distribution with mean $\mu_2^0 = 8$ and standard deviation $\sigma_2^0 = 1$. The iterations stop when $|p^{t+1} - p^t| < 10^{-4}$. After 171 iterations, the algorithm terminates with an estimate of mixture proportion $\hat{p} = 0.4875$; for comparison purposes, [8] produces an estimate $\hat{p}_{Bayesian} = 1 - 0.533 = 0.4667$. The Figure 5 shows the resulting density mixtures fitted using the method of [8] and our method against the background histogram of the log-transformed data. The Figure 6 illustrates the fitted first component of the mixture according to the method of [8] as well as the second component fitted according to both methods. Once again, the histogram of the log-transformed data is used in the background.

Note that the mixture density curves based on both methods are rather similar in Figure 5. One notable difference is that the method of [8] suggests mixture with a peak at the value of transformed ANC of about 6.4 whereas our

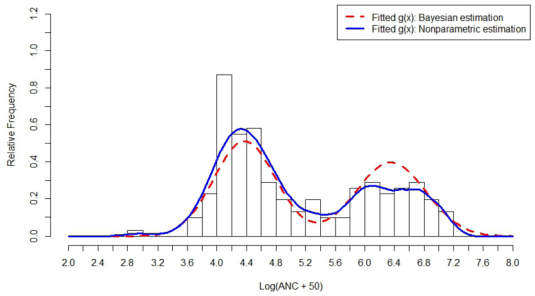


FIG 5. *Fitted mixture densities*

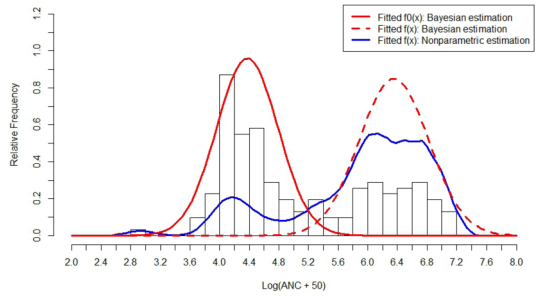


FIG 6. *Fitted component densities*

method produces a curve that seems to be following the histogram more closely in that location. The Figure 6 also seems to show that our method describes the data more faithfully than that of [8]. Indeed, the second parametric component fitted by the method of [8] is unable to reproduce the first peak around 4.2 at all. By doing so, the method of [8] suggests that the first peak is there only due to the first component. Our method, on the contrary, suggests that the first peak is at least partly due to the second component as well. Note that [8] discusses the possibility of a three component mixture for this dataset; results of our analysis suggest a possible presence of the third component as well based on a bimodal pattern of our fitted second component density curve. Finally, note that the method of [8] produces an estimated second component that implies a much higher second peak than the data really suggests whereas our method gives a more realistic estimate.

7. Discussion

The method of estimating two component semiparametric mixtures with a known component introduced in this manuscript relies on the idea of maximizing the smoothed penalized likelihood of the available data. Another possible view of this problem is as the Tikhonov-type regularization problem (some-

times also called *variational regularization scheme* in optimization literature). The resulting algorithm is an MM algorithm that possesses the descent property with respect to the smoothed penalized likelihood functional. Moreover, we also show that the resulting algorithm converges under mild restrictions on the kernel function used to construct a nonlinear smoother $\mathcal{N}$. The algorithm also shows reasonably good numerical properties, both in simulations and when applied to a real dataset. If necessary, a number of acceleration techniques can be considered in case of large datasets; for more details, see e.g. Lange et al. [21].

As opposed to the symmetrization method of Bordes et al. [3], our algorithm is also applicable to situations where the unknown component does not belong to any location family; thus, our method can be viewed as a more universal one of two. Comparing our method directly to that of Bordes et al. [3] and Patra and Sen [26] is a little difficult since our method is based,essentially, on perturbation of observed data the amount of which is controlled by the non-zero bandwidth h; thus, we arrive at what is apparently a solution different from that suggested in both [3] and [26].

There are a number of outstanding questions remaining concerning the model (1.1) that will have to be investigated as a part of our future research. First, the constraint that an unknown density is defined on a compact space is, of course, convenient when proving convergence of the algorithm generated sequence; however, it would be desirable to lift it later. We believe that, at the expense of some additional technical complications, it is possible to prove all of our results when the unknown density function $f(x)$ is defined on the entire real line but has sufficiently thin tails. Second, an area that we have not touched in this manuscript is the convergence of resulting solutions. For example, a solution obtained by running an empirical version of our algorithm (p_n^*, f_n^*) would be expected to converge to a solution (p^*, f^*) obtained by the use of the original algorithm in the integral form. Analogously, as $h \to 0$, it is natural to expect that (p^*, f^*) would converge to (p, f) such that the identity $(1-p)f_0(x) + pf(x) = g(x)$ is satisfied. We expect that some recent results in optimization theory concerning Tikhonov-type regularization methods with non-metric fitting functionals (see, for example, Flemming [13]) will be helpful in this undertaking. Our research in this area is ongoing.

Acknowledgements

The authors would like to acknowledge thoughtful comments from two reviewers that have helped to improve the presentation in this manuscript. The work of Michael Levine has been partially funded by the NSF-DMS grant 1208994.

References

[1] Bar-Lev, S. K. and P. Enis (1986). Reproducibility and natural exponential families with power variance functions. *The Annals of Statistics 14*(4), 1507–1522. MR0868315

[2] Bar-Lev, S. K. and O. Stramer (1987). Characterizations of natural exponential families with power variance functions by zero regression properties. *Probability Theory and Related Fields 76*(4), 509–522. MR0917676

[3] Bordes, L., C. Delmas, and P. Vandekerkhove (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics 33*(4), 733–752. MR2300913

[4] Bordes, L. and P. Vandekerkhove (2010). Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Mathematical Methods of Statistics 19*(1), 22–41. MR2682853

[5] Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics 38*(1), 100–145. MR2589318

[6] Chauveau, D., D. R. Hunter, and M. Levine (2015). Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys 9*, 1–31. MR3310969

[7] Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics 9*(1), 15–28. MR0216626

[8] Crawford, S. L. (1994). An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association 89*(425), 259–267. MR1266298

[9] Crawford, S. L., M. H. DeGroot, J. B. Kadane, and M. J. Small (1992). Modeling lake-chemistry distributions: Approximate bayesian methods for estimating a finite-mixture model. *Technometrics 34*(4), 441–453.

[10] Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*(3), 463–474. MR0254956

[11] Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press, Cambridge, United Kingdom. MR2724758

[12] Eggermont, P. P. B., V. N. LaRiccia, and V. LaRiccia (2001). *Maximum penalized likelihood estimation*, Volume 1. Springer, New York. MR1837879

[13] Flemming, J. (2010). Theory and examples of variational regularization with non-metric fitting functionals. *Journal of Inverse and Ill-Posed Problems 18*(6), 677–699. MR2746688

[14] Flemming, J. (2011). *Generalized Tikhonov regularization: basic theory and comprehensive results on convergence rates*. Ph. D. thesis.

[15] Hall, P., A. Neeman, R. Pakyari, and R. Elmore (2005). Nonparametirc inference in multivariate mixtures. *Biometrika Trust 92*(3), 667–678. MR2202653

[16] Hall, P. and X. Zhou (2003). Nonparametric estimation of component distributions in multivariate mixture. *The Annals of Statistics 31*(1), 201–224. MR1962504

[17] Hanche-Olsen, H. and H. Holden (2010). The kolmogorov-riesz compactness theorem. *Expositiones Mathematicae 28*(4), 385–394. MR2734454

[18] Hofmann, B., B. Kaltenbacher, C. Poeschl, and O. Scherzer (2007). A convergence rates result for tikhonov regularization in banach spaces with non-smooth operators. *Inverse Problems 23*(3), 987–1010. MR2329928

[19] Hunter, D. R. and K. Lange (2004). A tutorial on mm algorithms. *The American Statistician 58*(1), 30–37. MR2055509

[20] Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(3), 461–493. MR2420411

[21] Lange, K., D. R. Hunter, and I. Yang (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics 9*(1), 1–20. MR1819865

[22] Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics 11*(1), 86–94. MR0684866

[23] Lindsay, B. G. and P. Basak (1993). Multivariate normal mixtures: a fast consistent method of moments. *Journal of the American Statistical Association 88*(422), 468–476. MR1224371

[24] McLachlan, G. and D. Peel (2004). *Finite mixture models*. Wiley, Hoboken, New Jersey. MR1789474

[25] Ortega, J. and W. Reinboldt (1970). Iterative solution of nonlinear equations with multiple variables. MR0273810

[26] Patra, R. K. and B. Sen (2015). Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(4), 869–893. MR3534354

[27] Robin, S., A. Bar-Hen, J.-J. Daudin, and L. Pierre (2007). A semiparametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics and Data Analysis 51*(12), 5483–5493. MR2407654

[28] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press, Boca Raton, Florida. MR0848134