Electronic Journal of Statistics
Vol. 17 (2023) 1007–1042
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2121

# Revisiting consistency of a recursive estimator of mixing distributions*

**Vaidehi Dixit and Ryan Martin**

*Department of Statistics,
North Carolina State University,
2311 Stinson Dr.,
Raleigh, NC 27695, USA*
*e-mail:* vdixit@ncsu.edu; rgmarti3@ncsu.edu

**Abstract:** Estimation of the mixing distribution under a general mixture model is a very difficult problem, especially when the mixing distribution is assumed to have a density. *Predictive recursion* (PR) is a fast, recursive algorithm for nonparametric estimation of a mixing distribution/density in general mixture models. However, the existing PR consistency results make rather strong assumptions, some of which fail for practically relevant mixture models. In this paper, we first develop new consistency results for PR under weaker conditions and then we apply this theory in the important case of mixtures of scaled uniform kernels.

**MSC2020 subject classifications:** Primary 62G20, 62G07.
**Keywords and phrases:** Deconvolution, mixture model, monotone density estimation, predictive recursion, robustness.

Received October 2021.

## 1. Introduction

Mixture models are widely used in statistics and machine learning, often for density estimation and clustering. Here we will be considering a general version of the mixture model, where the mixture density is given by

$$m_P(x) = \int_{\mathbb{U}} k(x \mid u)\, P(du), \tag{1}$$

where $k$ is a known kernel, i.e., where $x \mapsto k(x \mid u)$ is a density for each $u \in \mathbb{U}$, and $P$ is the unknown mixing distribution on (the Borel $\sigma$-algebra of) $\mathbb{U}$. An advantage to this general form is its flexibility: depending on the kernel, the mixture density $m_P$ can take virtually any shape (e.g., DasGupta, 2008, p. 572), making such mixtures a powerful modeling tool for robust, nonparametric density estimation. Here we will assume that we have independent and identically distributed observations from a density $m$—which may or may not have the form (1)—and our primary goal is to estimate the mixing distribution $P$; this, in turn, will also give an estimate of the density $m$.

arXiv: 2110.02465

[*]This work was supported by the U.S. National Science Foundation, grant DMS–1737929.

An alternative perspective on the mixture model formulation considers a hierarchical formulation, where the first layer has iid $\mathbb{U}$-valued random variables, $U_1, \ldots, U_n$, from $P$, and then the second layer has

$$(X_i \mid U_i) \sim k(x \mid U_i), \quad \text{independent}, \; i = 1, \ldots, n.$$

The idea is that the $U_i$'s are latent/unobservable variables and the $X_i$'s are the observable data. It is easy to check that, marginally, the $X_i$'s are iid with density $m_P$ as in (1). The classical deconvolution problem (e.g., Fan, 1991; Stefanski and Carroll, 1990) is a special case where $k(x \mid u)$ is such that the second layer above could be described as "$X_i = U_i + \text{noise}$." This hierarchical formulation sheds light on the difficulties of the problem we are considering; that is, our goal is to estimate the distribution $P$ of the latent variables $U_1, \ldots, U_n$ based only on the corrupted observations $X_1, \ldots, X_n$.

For fitting the general mixture model (1), a number of different strategies are available. A natural approach is to use the nonparametric maximum likelihood estimator (MLE) of $P$ (Chen, 2017; Eggermont and LaRiccia, 1995; Lindsay, 1995) and the corresponding plug-in estimate of the mixture density $m_P$. An important feature of the nonparametric MLE of $P$ is that it is almost surely a discrete distribution (e.g., Lindsay, 1995). Another approach is to assume discreteness of $P$ with a fixed number of components and the component parameters are estimated via EM (Dempster, Laird and Rubin, 1977; McLachlan and Peel, 2000; Teel, Park and Sampson, 2015). Bayesian approaches have also been explored in this context; either by having a prior on $P$ like in Van Dyk and Meng (2001), or a prior on the number of components of $P$ like in Richardson and Green (1997). Nonparametric Bayes methods is an option too (e.g., Nguyen, 2013) but, the posterior for $P$ is supported on discrete distributions, so these methods also are not suited for estimating a smooth mixing density.

An alternative to the likelihood-based frameworks mentioned above, Newton, Quintana and Zhang (1998) proposed a recursive algorithm for nonparametric estimation of $P$, originally designed to serve as an approximation of the posterior mean under the Dirichlet process mixture formulation; see, also, Newton and Zhang (1999). The so-called *predictive recursion* (PR) algorithm estimates the mixing distribution recursively, starting with an initial guess $P_0$ and applies a simple update $(P_{i-1}, X_i) \mapsto P_i$, for each $i = 1, \ldots, n$, resulting in an estimate $P_n$ of $P$ and a corresponding estimate $m_n = m_{P_n}$ of $m$. One advantage of the PR estimator is its computational simplicity and speed. The other is that, unlike the likelihood-based methods above whose estimators are effectively discrete, PR is able to estimate a mixing distribution that has a smooth density with respect to any user-specified dominating measure. To our knowledge, PR is the only general method for estimating mixing distributions that has this property. Further details about the PR algorithm and its properties are discussed in Section 2.

Not being likelihood-based has its advantages when it comes to the smoothness of the estimates, but it also creates challenges when it comes to PR's theoretical justification. It was not until Newton (2002) that a first theoretical convergence analysis of PR was presented, establishing the asymptotic consistency of the PR estimator $P_n$ as $n \to \infty$. These first results, along with those in

Ghosh and Tokdar (2006) and Martin and Ghosh (2008), focus primarily on the case where $\mathbb{U}$ is a known finite set. Tokdar, Martin and Ghosh (2009) extended the results to compact $\mathbb{U}$, which was further extended by Martin and Tokdar (2009) who covered the case of model misspecification, where the true density $m$ need not have exactly the form (1), and bounded the rate of convergence.

Even the latter PR consistency results are based on conditions that can be too restrictive in applications. For example, Williamson (1956) showed that monotone densities are characterized as mixtures of the form (1) where $\mathbb{U} = [0,\infty)$ and $k(x \mid u) = u^{-1}1_{[0,u]}(x)$ is the uniform kernel, $\mathsf{Unif}(x \mid 0, u)$, with $1_A(x)$ being the indicator function of a set $A$. But for this particular kernel, it is not possible to check the sufficient conditions required in, e.g., Theorem 4.5 of Martin and Tokdar (2009). Similar issues would arise in other mixture model applications. Motivated by this deficiency in the state of the art, the focus of the present paper is to establish new asymptotic consistency properties for the PR estimator under weaker and more easily verified conditions.

Following a brief review of the existing theory for PR in Section 2, we establish convergence properties of the PR estimator—of both the mixture and the mixing distribution—under weaker conditions in Section 3. We then apply these new results in Section 4 to our motivating example, namely, monotone density estimation via mixtures of uniform kernels. There we first give a characterization of the best mixing distribution and mixture density within a special class of uniform mixtures. This characterization suggests a particular formulation of the PR algorithm and we use the general results presented in Section 3 to prove that PR consistently estimates this best mixture. Our choice to focus on a special class of uniform mixtures generally introduces some model misspecification bias, but we show that this bias is a vanishing function of two user-specified parameters. Therefore, the bias has no practical impact on PR's performance, as our numerical examples confirm. Finally, some concluding remarks are given in Section 5. Technical details and proofs are presented in the Appendix.

## 2. Background on PR

As mentioned briefly above, PR is a recursive algorithm designed for fast, nonparametric estimation of mixing distributions. The algorithm's inputs include the kernel $k$, an initial guess $P_0$ of the mixing distribution, supported on $\mathbb{U}$, a rule for defining a sequence of weights $i \mapsto w_i \in (0,1)$, and a sequence of data points $X_1, X_2, \ldots$. Then the recursive updates first presented in Newton, Quintana and Zhang (1998) define the PR algorithm:

$$P_i(du) = (1 - w_i)\, P_{i-1}(du) + w_i \frac{k(X_i \mid u)\, P_{i-1}(du)}{\int_{\mathbb{U}} k(X_i \mid v)\, P_{i-1}(dv)}, \quad u \in \mathbb{U}, \quad i \geq 1. \tag{2}$$

After $n$ data points have been observed, the mixing distribution estimator is $P_n$, and the corresponding mixture density estimator is $m_n = m_{P_n}$ defined according to (1). To understand the motivation behind PR, observe that the $i^{\text{th}}$ PR update is just a weighted average of $P_{i-1}(du)$ and the posterior for $U$ with prior $P_{i-1}(du)$ and likelihood $k(X_i \mid u)$. The initial guess $P_0$ is chosen such that it covers $\mathbb{U}$ in an uninformative manner, for example, a uniform distribution over $\mathbb{U}$. The weights, $w_i$, need to be decreasing in $i$ but not too quickly, specifically,

$$w_i \in (0,1), \quad \sum_{i=1}^{\infty} w_i = \infty, \quad \text{and} \quad \sum_{i=1}^{\infty} w_i^2 < \infty. \tag{3}$$

These conditions on the weights are common in the literature on stochastic approximation (e.g., Kushner and Yin, 2003; Lai, 2003; Martin and Ghosh, 2008), the precursor to the modern developments in stochastic gradient descent. A standard class of weights that satisfies (3) is $w_i = a(i+1)^{-b}$ for $b \in (0.5, 1]$ and $a < 2^b$. Some recent and novel applications of PR can be found in Scott et al. (2015), Tansey et al. (2018), and Woody, Padilla and Scott (2022).

The general algorithm above deals with general probability measures and is probably too abstract for practical applications. Typically, the user would have a specific dominating measure $\mu$ on $\mathbb{U}$ in mind, and then he/she can incorporate that information into the algorithm. In that case, the updates in (2) can be expressed in terms of the density or Radon–Nikodym derivative $p_i = dP_i/d\mu$ as

$$p_i(u) = (1-w_i)\, p_{i-1}(u) + w_i \frac{k(X_i \mid u)\, p_{i-1}(u)}{\int_{\mathbb{U}} k(X_i \mid v)\, p_{i-1}(v)\, \mu(dv)}, \quad u \in \mathbb{U}, \quad i \geq 1,$$

where $p_0 = dP_0/d\mu$ is the initial guess. Therefore, PR can be used to estimate a *mixing density*, compared to the nonparametric MLE which is almost surely discrete. Moreover, when the densities are evaluated on a fixed grid in $\mathbb{U}$, and the normalizing constant in the denominator is evaluated using quadrature, computation of the PR estimate, $P_n$, is fast and simple—done in $O(n)$ operations—compared to the nonparametric MLE or a Bayesian estimate based on Markov chain Monte Carlo (MCMC).

The above algorithm is described for the case when data points are arriving one at a time, but, of course, the same procedure can be carried out when the data $X_1, \ldots, X_n$ comes in a batch. When data are both batched and iid, as we consider here, one might be troubled by the fact that $P_n$ depends on the order in which the data are processed. In particular, while there are some potential advantages to PR's order-dependence (Dixit and Martin, 2019), it implies that $P_n$ is not a function of a minimal sufficient statistic. To overcome this, Newton (2002) suggested that one could evaluate the estimator $P_n$ separately on several random permutations of the data sequence and then take averages over permutations. This can be seen as a Monte Carlo estimate of the Rao–Blackwellized estimator, the average over all permutations. It has been shown empirically (e.g., Martin and Tokdar, 2012) that it only takes a few random permutations to remove the order-dependence, so, with the inherent computational efficiency of PR, the permutation-averaged version is still much faster than, say, MCMC.

Not being likelihood-based, it is not immediately obvious that the PR estimates would have any desirable statistical properties. It has, however, been shown that, under certain conditions, both $P_n$ and $m_n$ are consistent estimators.

Before stating these sufficient conditions for consistency, we need to describe *what* the PR estimates are estimating in general.

Suppose the true density of the iid data $X_1, \ldots, X_n$ is $m^\star$. Of course, there is generally no way to know if $m^\star$ can be expressed as a mixture model of the form (1) for a particular kernel, $k$. When the mixture model is incorrectly specified, there is no "$P^\star$" for the PR estimator $P_n$ to converge to, and we cannot expect $m_n$ to be a consistent estimator of $m^\star$. Instead, there may be a mixture density, $m^\dagger(x) = \int k(x \mid u)\, P^\dagger(du)$, that is "closest" to $m^\star$, and that $P_n$ and $m_n$ would converge to $P^\dagger$ and $m^\dagger$, respectively. Proximity here is measured in terms of the Kullback–Leibler divergence,

$$K(m^\star, m) = \int \log\{m^\star(x)/m(x)\}\, m^\star(x)\, dx.$$

More precisely, let $\mathscr{P}$ denote (a possibly proper subset of) the collection of probability distributions $P$ on $\mathbb{U}$, and define the corresponding set of mixtures of the form (1) for a given kernel $k$,

$$\mathscr{M} = \mathscr{M}(k, \mathscr{P}) = \{m_P : P \in \overline{\mathscr{P}}\},$$

where $\overline{\mathscr{P}}$ is the closure of $\mathscr{P}$ with respect to the weak topology, i.e., $\mathscr{P}$ plus all possible limits of weakly convergent sequences in $\mathscr{P}$. To avoid vaccuous cases, we will assume that $K(m^\star, m)$ is finite for at least one $m \in \mathscr{M}$. This is not a trivial assumption, however; see Section 4. In this case, the "best approximation" of $m^\star$ in $\mathscr{M}$ is the *Kullback–Leibler minimizer*, $m^\dagger$, that satisfies

$$K(m^\star, m^\dagger) = \inf\{K(m^\star, m) : m \in \mathscr{M}\} \tag{4}$$

A relevant question is whether such a minimizer exists and if it is unique. Assuming that $K(m^\star, m)$ is finite for at least one $m \in \mathscr{M}$ and given that it is a convex function, we can expect that a minimizer $m^\dagger$ exists and is unique. Existence of a $P^\dagger$ corresponding to $m^\dagger$ is guaranteed by assuming certain conditions on $k$ and $\mathbb{U}$; see Conditions A1 and A2 in Martin and Tokdar (2009) and, more generally, Liese and Vajda (1987, Ch. 8). However, uniqueness of $P^\dagger$ requires identifiability of the mixture model (1) in $P$.

In Tokdar, Martin and Ghosh (2009), consistency of the PR estimators was established in the case where the mixture model was correctly specified, i.e., when $m^\star \in \mathscr{M}$, so that there exists a true $P^\star \in \mathscr{P}$. That is, under certain conditions, they showed $K(m^\star, m_n) \to 0$ almost surely and that $P_n \to P^\star$ weakly almost surely. Martin and Tokdar (2009) extended these consistency results to the case where the mixture model is not necessarily correctly specified, i.e., where possibly $m^\star \notin \mathscr{M}$. This extension is a practically important one, as it provides a theoretical basis for the PR-based marginal likelihood estimation framework developed in Martin and Tokdar (2011) and later applied in, e.g., Martin and Han (2016), Dixit and Martin (2022). Under conditions slightly stronger than those given in Tokdar, Martin and Ghosh (2009) for the correctly specified case, they showed that $K(m^\star, m_n) \to K(m^\star, m^\dagger)$ and $P_n \to P^\dagger$ weakly, both almost surely. This implies, for example, that the PR estimates do the best they could, asymptotically, relative to the specified model. Moreover, it means that the PR estimator can be understood as (asymptotically) trying to minimize the function $P \mapsto \int \log\{m_P(x)\}\, m^\star(x)\, dx$, similar to what the nonparametric MLE aims to achieve. It turns out, however, that the sufficient conditions for consistency stated in Martin and Tokdar (2009), very similar to those in Tokdar, Martin and Ghosh (2009), are rather restrictive. The most problematic of those assumptions is the following:

$$\sup_{u_1, u_2 \in \mathbb{U}} \int \Bigl\{ \frac{k(x \mid u_1)}{k(x \mid u_2)} \Bigr\}^2 m^\star(x)\, dx < \infty. \tag{5}$$

For nice kernels like $k(x \mid u) = \mathsf{N}(x \mid u, \sigma^2)$ for a fixed $\sigma^2 > 0$, if $\mathbb{U}$ is compact and $m^\star$ has Gaussian-like tails, then (5) can be satisfied. However, if $m^\star$ is heavier-tailed, then (5) could easily fail. More concerning is if we are considering a not-so-nice kernel, such as uniform: $k(x \mid u) = \mathsf{Unif}(x \mid 0, u)$, for $x > 0$ and $u > 0$. The $u$-dependent support implies that the ratio in the above display is infinite on an open interval and, hence, (5) obviously fails. The difficulty in verifying condition (5) in several practical applications is what motivated our present investigation into potentially weaker sufficient conditions.

## 3. New consistency results

### *3.1. Conditions*

The goal is to develop a new set of sufficient conditions for PR consistency that are weak enough that they can be checked in the applications we mentioned above, in particular, the case of uniform kernels for monotone density estimation. First we make clear the setup/conditions, and then we present the main results.

*Condition* 1. The PR algorithm's weights satisfy $w_i = a(i+1)^{-1}$, for $a < \frac{2}{9}$.

*Condition* 2. The mixing distribution support, $\mathbb{U}$, is compact.

*Condition* 3. The kernel, the initial guess $P_0$, with corresponding $m_0 = m_{P_0}$, and the true $m^\star$ satisfy the following integrability property:

$$\sup_{u \in \mathbb{U}} \int \Bigl\{ \frac{k(x \mid u)}{m_0(x)} \Bigr\}^2 m^\star(x)\, dx < \infty. \tag{6}$$

Of course, the specific weights in Condition 1—which are of the same form as the weights used in Hahn, Martin and Walker (2018)—satisfy the basic conditions (3) on the $w_i$'s, but others do too. The reason we adopt this specific choice is that it allows us to replace (5) with the weaker bound (6) discussed more below. And since the choice of weights is entirely in the hands of the user, while the choice of kernel may be determined by the context of the problem and $m^\star$ is a choice made by "Nature" and hidden from the user, it is best to sacrifice on generality in directions the user can control. In our experience, the empirical performance is not sensitive to the choice of $a > 0$ in Condition 1; and since we do not claim that this condition is necessary for PR consistency, in practice we take $w_i = (i+1)^{-1}$. Weight sequences that vanish more slowly than these have some appeal, the theoretical results that we are aware of in the stochastic approximation literature (e.g., Dvoretzky, 1956; Mokkadem and Pelletier, 2007; Sacks, 1958) suggest that $w_i = O(i^{-1})$ are "optimal" in a certain sense.

Condition 2 assumes that the mixing distribution support is compact, but this is typically not a severe practical restriction in practice. Compactness of $\mathbb{U}$ is not strictly needed for the results presented below, but (a) some more complicated notion of compactness is needed, as we briefly discuss in the paragraph leading up to Corollary 2, and (b) Condition 3 might be difficult to check without $\mathbb{U}$ being compact. For these reasons, we opt for the simpler but slightly more restrictive compactness condition listed above. Compactness is also a standard assumption in the literature on the analysis of likelihood-based mixing distribution estimation, e.g., Chen (2017) and Nguyen (2013).

Finally, the most complicated is Condition 3. Intuitively, one cannot expect that an arbitrary pair of inputs $(k, P_0)$ can produce a consistent estimate of any true density $m^\star$, and Condition 3 describes a connection between these pieces that is sufficient for consistent estimation. Of course, like any assumptions about the true/unknown $m^\star$, Condition 3 cannot be "checked". This condition determines which $m^\star$'s the PR algorithm, initialized with $(k, P_0)$, can consistently estimate. The user can identify assumptions about $m^\star$ that they are willing to make and, from there, attempt to check Condition 3. For example, under mild assumptions on $m^\star$, we show in Section 4 below that (6) can be checked for uniform kernels while the condition (5) in Martin and Tokdar (2009) cannot.

To better understand the actual assumption being made in Condition 3, it may help to re-express the integrand in (6) as

$$\frac{k(x \mid u)}{m_0(x)} \cdot \frac{m^\star(x)}{m_0(x)} \cdot k(x \mid u).$$

First, if the PR prior guess $P_0$ is not too tightly concentrated, then the mixture $m_0$ would be heavier-tailed than any individual kernel $k(\cdot \mid u)$. In that case, the first ratio in the above display would be bounded, or at least would not increase too rapidly. Second, we cannot expect PR, or any mixture model-based method for that matter, to be able to do a good job of estimating $m^\star$ if a mixture with a relatively diffuse mixing distribution cannot adequately cover the support of $m^\star$. So the heart of Condition 3 is an assumption that the posited mixture model *can* adequately cover the support of $m^\star$, in the sense that the second ratio in the above display is not blowing up too rapidly. Finally, if the two ratios are well controlled, then the integral with respect to $k(\cdot \mid u)$ should be bounded uniformly in $u$.

### *3.2. Main results*

Our primary goal is to show, under roughly the conditions stated above, that the PR estimator $P_n$ of the mixing distribution is consistent. A direct proof of this result is currently out of reach, but it can be established indirectly by considering the mixture density estimator. In particular, we show below that the PR estimator, $m_n = m_{P_n}$, of the density $m^\star$ is consistent in the sense that $K(m^\star, m_n)$ converges almost surely to $\inf_{m \in \mathscr{M}} K(m^\star, m)$, the minimum Kullback–Leibler divergence over the posited mixture model class $\mathscr{M}$. In the special case where $m^\star \in \mathscr{M}$, this implies consistency in the usual sense: $K(m^\star, m_n) \to 0$ almost surely. In either case, it says that the PR estimator, $m_n$, is close to the best possible mixture approximation of $m^\star$, at least asymptotically. From this we will deduce consistency of the mixing distribution $P_n$; see Corollary 2 below.

**Theorem 1.** *Under Conditions 1–3, the PR estimator, $m_n$, of the density $m^\star$ satisfies $K(m^\star, m_n) \to \inf_{m \in \mathscr{M}} K(m^\star, m)$ almost surely. In particular, if $m^\star \in \mathscr{M}$, then $K(m^\star, m_n) \to 0$ almost surely.*

*Proof.* See Appendix A.1. □

Here we give a very rough sketch of the proof strategy. Start by writing $K_n = K(m^\star, m_n) - \inf_{m \in \mathscr{M}} K(m^\star, m)$, and let $\mathcal{A}_i$ denote the $\sigma$-algebra generated by the observations $X_1, \ldots, X_i$, for $i = 1, 2, \ldots$. We show in the proof that

$$\mathsf{E}(K_n \mid \mathcal{A}_{n-1}) = K_{n-1} - w_n T(P_{n-1}) + w_n^2 \mathsf{E}(Z_n \mid \mathcal{A}_{n-1}), \quad n \geq 1,$$

where

$$T(P) = \int_{\mathbb{U}} \Big\{ \int_{\mathbb{X}} \frac{m(x)}{m_P(x)} \, k(x \mid u) \, dx \Big\}^2 P(du) - 1, \tag{7}$$

and $Z_n$ is a "remainder" term defined in the appendix. It follows from Jensen's inequality that $T(P) \geq 0$, with equality if and only if $P = P^\dagger$, the Kullback–Leibler minimizer. If we could ignore the remainder term, then $K_n$ would be a non-negative supermartingale and, therefore, would converge almost surely to some $K_\infty$. Of course, the remainder term cannot be ignored, so we will use the "almost supermartingale" results in Robbins and Siegmund (1971) to accommodate this. Moreover, to show that $K_\infty$ is 0 almost surely, we will use some new and useful properties of the function $T$ in (7) which were overlooked in the analysis presented in Martin and Tokdar (2009).

When the mixture model is correctly specified, so that $m^\dagger = m^\star$, it follows from Theorem 1 and the familiar properties of Kullback–Leibler divergence that $m_n \to m^\star$ almost surely in Hellinger or total variation distance, i.e., that $\int (m_n^{1/2} - m^{\star 1/2})^2 \, dx$ and $\int |m_n - m^\star| \, dx$ both go to 0 almost surely. In the general case where the mixture model is misspecified, Theorem 1 still strongly suggests that $m_n \to m^\dagger$, but some effort is required to connect the Kullback–Leibler difference to a distance between $m_n$ and $m^\dagger$. Towards this, define the *Hellinger contrast* $\rho(m_1, m_2) = \rho_{m^\star}(m_1, m_2)$, which is given by

$$\rho^2(m_1, m_2) = \int (m_1^{1/2} - m_2^{1/2})^2 (m^\star / m^\dagger) \, dx.$$

This is just a weighted version of the ordinary Hellinger distance—with weight function $m^\star / m^\dagger$—so it is a proper metric. Clearly, if the mixture model is correctly specified, so that $m^\dagger = m^\star$, then $\rho$ is exactly the Hellinger distance. See

Patilea (2001) and Kleijn and van der Vaart (2006) for further details on the Hellinger contrast. The following result establishes that $\rho(m^\dagger, m_n) \to 0$ almost surely, which implies that the limit $m_\infty$ of $m_n$ satisfies $m_\infty = m^\star$ almost everywhere with respect to the measure with Lebesgue density $m^\star$. Under some additional conditions, namely, that $m^\dagger$ is suitably close to $m^\star$, the PR estimator $m_n$ is shown to converge to $m^\dagger$ in total variation distance, which implies the limit is equal to $m^\dagger$ almost everywhere with respect to Lebesgue measure.

**Corollary 1.** *Under the conditions of Theorem 1, $\rho(m^\dagger, m_n) \to 0$ almost surely. Moreover, if $m^\dagger/m^\star \in L_\infty(m^\star)$, then $m_n \to m^\dagger$ almost surely in total variation.*

*Proof.* See the proof of Corollary 4.10 in Martin and Tokdar (2009). □

Finally, what can be said about the convergence of the mixing distribution estimator, $P_n$? Again, Theorem 1 strongly suggests that $P_n$ is converging to $P^\dagger$ in some sense, but we cannot make that leap immediately. In particular, without additional assumptions, there is no guarantee that $P^\dagger$ is unique or even that $P_n$ converges at all. For this, we will need *identifiability* of the mixture model (1) and tightness of $(P_n)$. Under Condition 2, as we assume here, tightness of $P_n$ follows from Prokhorov's theorem. If compactness of $\mathbb{U}$ is not a feasible assumption, then one can instead verify the more general sufficient condition, namely, Condition A6 in Martin and Tokdar (2009), for tightness of $P_n$.

We will also require the following condition on the kernel density $k$, expressed in terms of a general sequence of mixing distributions $(Q_t)$ on $\mathbb{U}$:

$$Q_t \to Q_\infty \text{ weakly implies } m_{Q_t}(x) \to m_{Q_\infty}(x) \text{ for almost all } x. \tag{8}$$

In words, (8) states that the kernel is such that weak convergence of mixing distributions implies almost everywhere pointwise convergence of mixture densities. This is a key assumption in the available consistency results for the nonparametric MLE; see Condition (KW2) in Chen (2017). Condition (8) holds if $u \mapsto k(x \mid u)$ is bounded and continuous for almost all $x$, as was assumed in Martin and Tokdar (2009) and elsewhere. However, in some examples, like in Section 4, continuity of the kernel fails, but condition (8) can be verified.

**Corollary 2.** *In addition to the conditions of Theorem 1, assume that*

- *the mixture model (1) is identifiable, i.e., $m_P = m_{P'}$ almost everywhere implies $P = P'$,*
- *the kernel is such that (8) holds,*
- *and $m^\dagger/m^\star \in L_\infty(m^\star)$.*

*Then the Kullback–Leibler minimizer $P^\dagger$ is unique and $P_n \to P^\dagger$ weakly almost surely.*

*Proof.* Since $P_n$ is tight, there exists a subsequence $P_{n(t)}$ such that $P_{n(t)} \to P_\infty$ weakly, for some $P_\infty$. By (8), we have pointwise convergence of the mixture densities, i.e., $m_{n(t)}(x) \to m_\infty(x)$ for almost all $x$, and then $m_{n(t)} \to m_\infty$ in total variation distance thanks to Scheffé's theorem. But Corollary 1 already gives us $m_n \to m^\dagger$ almost surely in total variation distance on the full/original sequence.

Therefore, it must be that $m_\infty = m^\dagger$ almost surely and, by identifiability, that $P_\infty = P^\dagger$. Since any such convergent subsequence of $P_n$ would have the same almost weak limit, $P^\dagger$, it must be that $P_n$ itself converges weakly almost surely to $P^\dagger$, as claimed. □

Identifiability of the mixture model $m_P$ in $P$ is non-trivial. Additively-closed one-parameter families of distributions were proved to be identifiable in Teicher (1961). Identifiability of finite mixtures of gamma and of Gaussian distributions was proved in Teicher (1963). Scale mixtures of uniform distributions, like we discuss in Section 4 below, were shown to be identifiable in Williamson (1956). More generally, identifiability of mixture models needs to be checked on a case-by-case basis. The boundedness assumption on $m^\dagger/m^\star$, as in Corollary 1, is needed simply to convert convergence of $m_n$ to $m^\dagger$ in the Hellinger contrast to convergence in total variation. This conversion would not be necessary if (8) implied convergence in Hellinger contrast, but showing this would require some knowledge of the relationship between $m^\star$ and $m^\dagger$; whether there is condition weaker than "$m^\dagger/m^\star \in L_\infty(m^\star)$" remains an open question. The reason this condition is needed at all is because we allow for the possibility that the mixture model is misspecified—the condition completely disappears when the model is correctly specified, since $m^\dagger = m^\star$. If no such conditions appear in other literature on mixing distribution estimation (e.g., Chen, 2017), then it is because they are assuming the model is correctly specified.

## 4. Application: Mixtures of beta kernels

### *4.1. Background*

A density $m$ supported on $\mathbb{X} = [0, \infty)$ is called *monotone* (or *1-monotone*) if it is non-increasing, *2-monotone* if it is non-increasing and convex, and *s-monotone* ($s \geq 3$) if $(-1)^t m^{(t)}$ is non-negative, non-increasing, and convex for $t = 0, 1, \ldots, s-2$. Models that involve $s$-monotone densities are common in, say, the time-to-event literature; for details about this and other shape-constrained problems, see Groeneboom and Jongbloed (2014). What makes these type of models relevant to our investigation here is the beautiful mixture representation of Williamson (1956), which states that for any $s$-monotone density $m$ on $\mathbb{X}$, there exists a mixing distribution $P$, supported on $\mathbb{U} = [0, \infty)$, such that,

$$m(x) = \int_0^\infty k_s(x \mid u)\, P(du), \tag{9}$$

where the kernel $k_s$ is a scaled beta density, i.e.,

$$k_s(x \mid u) = s u^{-s} (u - x)^{s-1}\, 1_{[0,u]}(x), \quad x \in \mathbb{X}, \quad u \in \mathbb{U}.$$

So, in applications where the underlying density on $\mathbb{X}$ is believed to have certain monotonicity properties, a mixture model would be a completely natural way to move forward. Then questions about estimation of the underlying mixing distribution might be relevant, and the PR algorithm would be an obvious candidate for answering those questions.

Our primary focus in what follows is the case where $s = 1$, so that the kernel $k_1(x \mid u) = u^{-1} 1_{[0,u]}(x)$ is the $\mathsf{Unif}(0, u)$ density. This corresponds to a model that simply assumes the true density is non-increasing on $\mathbb{X}$. Most of the key results below hold for general $s$. There are, however, some important practical details that we can provide in the $s = 1$ case that are still out of reach in the general-$s$ case, so it makes sense to focus our attention on the former. We explain the specifics in Remark 1 below.

Let $X_1, \ldots, X_n$ be iid from a monotone ($s = 1$) density $m^\star$. One approach to estimating $m^\star$ is to calculate the nonparametric MLE, also known as the *Grenander estimator* (Grenander, 1956), which is the left derivative of the least concave majorant of the empirical distribution function. It is known that Grenander's is a consistent estimator of $m^\star$, with consistency results obtained in Rao (1969) and Groeneboom (1985). However, as shown in, e.g., Woodroofe and Sun (1993), the Grenander estimator tends to over-estimate near the origin and, in particular, is inconsistent at the origin. The same authors proposed a penalized likelihood estimator that penalizes the Grenander estimator at the origin and is also consistent overall. Extensions of these results to the $s = 2$ and $s > 2$ case have been made in, e.g., Groeneboom, Jongbloed and Wellner (2001) and Balabdaoui and Wellner (2007), respectively.

Another approach is Bayesian, whereby a prior distribution on $m$ is imposed by using the mixture characterization in (9) along with a suitable prior on the mixing distribution $P$. A natural choice is a Dirichlet process prior on $P$, leading to a Dirichlet process mixture of uniforms model, in the $s = 1$ case, for the density $m$; see Bornkamp and Ickstadt (2009). Although this approach seems straightforward, obtaining asymptotic consistency results for the posterior distribution is made difficult by the uniform kernel's varying support. In particular, if the support for the mixing distribution is not suitably chosen, then the Kullback–Leibler divergence of a posited mixture model from the true density would be infinite, which creates problems for verifying the so-called "Kullback–Leibler property" (Schwartz, 1965; Wu and Ghosal, 2008) in the classical Bayesian consistency theory. Some strategies have been suggested in, e.g., Salomond (2014), who showed that the Bayesian posterior distribution under the Dirichlet process mixture prior has a near optimal concentration rate in total variation. More recently, Martin (2019) proposed the use of an empirical, or data-driven prior for which the prior support conditions required for asymptotic consistency are automatically satisfied, and showed that the corresponding empirical Bayes posterior distribution concentrates around the true monotone density at nearly optimal minimax rate. But the fully Bayesian solutions are computationally non-trivial and somewhat time consuming; moreover, the estimates tend to be relatively rough. The PR algorithm, which is computationally fast and tends to produce smooth estimates, is a natural alternative to the aforementioned likelihood-based methods.

### 4.2. PR for uniform (and beta) mixtures

Suppose that the true density $m^\star$ is a monotone density supported on $[0,\infty)$, with $s = 1$; extensions to the general $s$ case will be discussed in Remark 1 below. We know that $m^\star$ can be written as a mixture in (9), so there exists a mixing distribution $P^\star$, which is also supported on $[0,\infty)$. This point is relevant because of the following unique feature of beta mixtures: if $m_P$ is a mixture model as in (9) with $P$ supported on $[0,L)$, then $m_P(x) = 0$ for all $x > L$ and, hence, if $L < \infty$, then $K(m^\star, m_P) \equiv \infty$. Therefore, the upper bound of $\mathbb{U}$ being $\infty$ creates some serious challenges. For practical implementation of the PR algorithm, and for the theory as discussed above, a compact mixing distribution support is needed. This calls for a different approach.

For a fixed $L \in (0,\infty)$, define a new target, $m^{\star L}$, which is simply $m^\star$ restricted and renormalized to $[0,L)$. That is, if $M^\star$ denotes the distribution function corresponding to the density $m^\star$, then

$$m^{\star L}(x) = \frac{m^\star(x)\,1_{[0,L]}(x)}{M^\star(L)}.$$

Alternatively, $m^{\star L}$ can be viewed as the conditional density of $X$, given $X \leq L$; see below. The point of this adjustment is that $m^{\star L}$ has a known and bounded support, so a mixture model with mixing distribution supported on (a large subset of) $[0,L)$ can be fit with the PR algorithm to efficiently and accurately estimate this new target $m^{\star L}$. Note that $m^{\star L}$ can be made arbitrarily close to $m^\star$ by choosing $L$ sufficiently large (see below), so this modification has no practical consequences.

For technical and practical reasons, we cannot use the PR algorithm when the support of the mixing distribution contains $u = 0$, so we introduce a new lower bound $\ell \in (0,L)$, which can be arbitrarily small. Then the proposed mixture model to be fit by PR is

$$m_P(x) = \int_{\mathbb{U}} k_1(x \mid u)\,P(du), \quad x \in [0,L], \quad \mathbb{U} = [\ell, L]. \tag{10}$$

While both $m_P$ above and the adjusted target $m^{\star L}$ are supported on $[0,L]$, the model in (10) is still *slightly misspecified* through the introduction of the lower bound $\ell > 0$ of the mixing distribution support. In particular, note that $m_P(x)$ is constant for $x \in [0,\ell]$. But the fact that $\ell$ can be taken arbitrarily small means that there are no practical consequences to this misspecification. It does complicate the convergence analysis, but, fortunately, the theory presented in Section 3 above is general enough to handle this.

Given that the mixture model (10) is slightly misspecified, it is important to know what we can expect the PR algorithm to do. Theorem 1 states that, roughly, the PR estimator $m_n$ will converge to the Kullback–Leibler minimizer $m^\dagger$. Since the supports of $m^{\star L}$ and the model densities $m_P$ in (10) are the same, we avoid the "$K(m^{\star L}, m_P) \equiv \infty$" problem so minimizing the Kullback–Leibler divergence is well-defined. To understand the bias coming from model misspecification, it will be important to understand what $m^\dagger$ looks like. Incidentally, Williamson (1956) established that beta mixtures are identifiable, so there is a unique mixing distribution, $P^\dagger$, supported on $\mathbb{U}$, at which the Kullback–Leibler divergence is attained. The following lemma gives the details.

**Lemma 1.** *For the targeted monotone density $m^{\star L}$ supported on $[0, L]$, if the proposed mixture model is as in (10), then the unique minimizer, $P^\dagger = P^{\dagger \ell, L}$, of the Kullback–Leibler divergence $P \mapsto K(m^{\star L}, m_P)$ is given by*

$$P^\dagger = a_\ell \, \delta_{\{\ell\}} + a_{\mathbb{U}} \, P^\star|_{\mathbb{U}} + a_L \, \delta_{\{L\}}, \tag{11}$$

*where $\delta_{\{t\}}$ is the Dirac point-mass at $t$, $P^\star|_{\mathbb{U}}$ is $P^\star$ restricted to $\mathbb{U} = [\ell, L]$, and the coefficients are given by*

$$a_\ell = \frac{P^\star([0,\ell])}{M^\star(L)}, \quad a_{\mathbb{U}} = \frac{P^\star([0,L])}{M^\star(L)}, \quad a_L = \frac{Lm^\star(L)}{M^\star(L)},$$

*with $M^\star$ the distribution function corresponding to $m^\star$. Then the best approximation of $m^{\star L}$ under model (10) is $m^\dagger = m_{P^\dagger}$, given by*

$$m^\dagger(x) = a_\ell \, k_1(x \mid \ell) + a_{\mathbb{U}} \int_{\mathbb{U}} k_1(x \mid u) \, P^\star(du) + a_L \, k_1(x \mid L). \tag{12}$$

*Proof.* See Appendix A.2. □

The characterization result in Lemma 1 is intuitive. There is a true $P^\star$ that characterizes the true monotone mixture density $m^\star$, both generally supported on $[0, \infty)$. Our proposed model restricts the mixing distribution's support to $[\ell, L]$, so it makes sense that the best approximation would agree with $P^\star$ on $[\ell, L]$ and then suitably allocate the remaining mass to the endpoints $\ell$ and $L$.

From Section 2, recall that the implementation of the PR algorithm begins with an initial guess $P_0$, and that this effectively determines the dominating measure with respect to which $P_n$ has a density. PR's ability to choose the underlying dominating measure comes in handy in cases like this where we know that the target mixing distribution, $P^\dagger$, has an "unusual" dominating measure. From Lemma 1, we know that the best mixing distribution for fitting mixture model (10) to $m^{\star L}$ puts point masses at the endpoints, $\ell$ and $L$, of $\mathbb{U}$, and has a density with respect to Lebesgue measure on the interior of $\mathbb{U}$. So, naturally, we can initialize the PR algorithm with a starting guess $P_0$ that has a density with respect to the dominating measure $\delta_{\{\ell\}} + \lambda_{\mathbb{U}} + \delta_{\{L\}}$, where $\lambda_{\mathbb{U}}$ denotes Lebesgue measure on $\mathbb{U}$. Specifically, our proposal is to initialize the PR algorithm at

$$P_0 = p_{0,\ell} \, \delta_{\{\ell\}} + (1 - p_{0,\ell} - p_{0,L}) \, P_{0,\mathbb{U}} + p_{0,L} \, \delta_{\{L\}},$$

where $p_{0,\ell}$ and $p_{0,L}$ are positive with sum strictly less than 1, and $P_{0,\mathbb{U}}$ has a density with respect to Lebesgue measure, e.g., $P_{0,\mathbb{U}}$ could just be a uniform distribution on $\mathbb{U}$. Then the estimate, $P_n$, after the $n^{\text{th}}$ iteration will have the same form

$$P_n = p_{n,\ell} \, \delta_{\{\ell\}} + (1 - p_{n,\ell} - p_{n,L}) \, P_{n,\mathbb{U}} + p_{n,L} \, \delta_{\{L\}},$$

and the corresponding mixture density estimate, $m_n$, is obtained as usual by integrating the kernel with respect to the mixing distribution $P_n$.

### 4.3. Theoretical results

Now that we know what PR ought to converge to, we are ready to state our main result of this section. First, a word about the notation/terminology that follows. In our previous results, when we wrote "almost surely," this was referring to the law that corresponds to iid sampling from $m^\star$. In the results below, $m^{\star L}$ is the target, so we will write "$m^{\star L}$-almost surely" to be clear that it is with respect to the law corresponding to iid sampling from $m^{\star L}$. Recall that $m^{\star L}$ is the conditional density of $X$, given $X \leq L$, so this modified law can be interpreted as iid sampling from $m^\star$, but throwing away any data points that exceed $L$. Again, since $L$ can be taken arbitrarily large, there are no practical consequences of this restriction. In fact, a bound on the bias induced by both the $L$- and $\ell$-restrictions is given in Proposition 1 below.

**Theorem 2.** *Consider the mixture model $m_P$ in (10) with compact mixing distribution support $\mathbb{U} = [\ell, L]$, and let $m^{\star L}$ denote the true $m^\star$ restricted and renormalized to $[0, L]$. If the PR algorithm is initialized at a $P_0$ that includes a point mass at $L$, then the PR estimator $m_n$ satisfies*

$$K(m^{\star L}, m_n) \to K(m^{\star L}, m^\dagger), \quad m^{\star L}\text{-almost surely}$$

*where $m^\dagger$ is as given in Lemma 1. Moreover, $m_n$ converges $m^{\star L}$-almost surely to $m^\dagger$ in total variation distance and the mixing distribution estimates $P_n$ converges weakly $m^{\star L}$-almost surely to $P^\dagger$ in (11).*

*Proof.* See Appendix A.3. □

*Remark* 1. The result in Theorem 2 is specifically for the ordinary monotone case, $s = 1$, where the kernel $k_1$ is a scaled uniform. For the general $s$ case, however, with a scaled beta kernel $k_s$, some things can be said. Indeed, the conditions of Theorem 1 can be established for general $s$ (see Appendix A.6), so consistency of PR's mixture density estimator follows. What is currently out of reach is a general-$s$ characterization of the Kullback–Leibler minimizer $P^\dagger = P^{\dagger \ell, L}$ like we have for $s = 1$ in Lemma 1. We expect that a similar characterization can be given in the general-$s$ case—our conjecture is that $P^\dagger$ has the same form as above, i.e., agreeing with $P^\star$ in the interior of $[\ell, L]$ but with point masses at the endpoints. If this conjecture is true, then consistency of $P_n$ as stated in Theorem 2 above and the follow-up results in Propositions 1–2 below would also hold for the general-$s$ case.

Our choice to restrict the mixing distribution's support to $\mathbb{U} = [\ell, L]$ introduces some bias. That is, the limit $m^\dagger$ of the sequence of PR estimators is the Kullback–Leibler minimizer over all mixtures supported on $\mathbb{U} = [\ell, L]$, but it is different from $m^{\star L}$, which is different from $m^\star$. Intuitively, if $\ell \approx 0$ and $L \approx \infty$, then the bias ought to be negligible. The next result confirms this intuition by bounding the bias as a function of $(\ell, L)$.

**Proposition 1.** *The $L_1$ distance between the monotone $m^\star$ and the best approximation $m^\dagger$ in (12) under the restricted model (10) is bounded as*

$$\int |m^\dagger(x) - m^\star(x)|\,dx \leq 2\bigl\{1 - M^\star(L) + M^\star(L)^{-1}P^\star([0,\ell])\bigr\}. \tag{13}$$

*Proof.* See Appendix A.4. □

To make the bound in (13) more concrete, we consider a specific case. A common choice in the literature (e.g., Martin, 2019; Salomond, 2014) is to assume $m^\star$ has tails that vanish exponentially fast, so that $m^\star(x) \leq \exp(-bx^r)$, for all large $x$ and some positive constants $b$ and $r$; the case $r = \infty$ corresponds to $m^\star$ having a bounded support. From this, and standard asymptotic bounds on the incomplete gamma function, it follows that $1 - M^\star(L) \lesssim L^{-r}\exp(-bL^r)$, for large $L$. Furthermore, if, e.g., $P^\star$ has a bounded density at 0, then we have $P^\star([0,\ell]) \lesssim \ell$. Combining these two, we arrive at the following, more explicit bound on the $L_1$ bias as a function of $(\ell, L)$:

$$\int |m^\star(x) - m^\dagger(x)|\,dx \lesssim L^{-r}e^{-bL^r} + \ell.$$

Clearly, by taking $\ell$ small and $L$ even just moderately large, the overall bias as a result of restricting to $\mathbb{U} = [\ell, L]$ can be made negligibly small.

As a final technical detail in this section, we consider the problem of estimating $m^\star(0)$, the density at its mode, the origin. This is an interesting and challenging problem, with a variety of applications; see, e.g., Vardi (1989). In particular, Woodroofe and Sun (1993) highlight examples such as time between breakdowns of a system and distribution of galaxies that require the estimation of this modal $m^\star(0)$. The PR algorithm gives an obvious estimator of $m^\star(0)$, in particular, $m_n(0)$. The following result gives a theoretical basis for using this estimate and simulations in Section 4.4 show that the proposed estimate at 0 performs well when compared to existing methods.

**Proposition 2.** *Under the assumptions of Theorem 2, $m_n(0) \to m^\dagger(0)$ $m^{\star L}$-almost surely. Furthermore, the bias is bounded as*

$$m^\dagger(0) - m^\star(0) \lesssim 1 - M^\star(L) \to 0, \quad \textit{as } L \to \infty.$$

*Proof.* See Appendix A.5. □

To be clear, no claim is being made that that the PR estimator $m_n(0)$ is a consistent estimator of $m^\star(0)$. Proposition 2 is simply saying that $m_n(0) \to m^\dagger(0)$ and that the difference between $m^\dagger(0)$ and $m^\star(0)$ can be made arbitrarily small by picking $L$ sufficiently large. There are technical challenges that arise when considering a sort of sieve with support $\mathbb{U}_n = [\ell_n, L_n]$ that is expanding as $n \to \infty$ in conjunction with the recursive estimator. Some comments on this are made in Section 5 below.

## 4.4. *Numerical illustrations*

### 4.4.1. *Monotone density estimation*

In this section we compare different methods for monotone density estimation to our PR-based method. The four methods we consider are the Grenander estimate, a Bayesian approach using a Dirichlet process, Bayesian approach using an empirical prior, and the method based on optimization of the penalized likelihood. The Grenander estimate is based on the nonparametric MLE and can be calculated easily using the R package `fdrtool` (Klaus and Strimmer, 2015). Settings for the Dirichlet process mixture and the empirical Bayes were based on those suggested in Martin (2019) and we used his R codes.[1] The penalized likelihood maximization was based on Woodroofe and Sun (1993) and we used one of the values recommended by those authors for their penalization parameter, i.e., $\alpha = n^{-1} \log n$. For PR, we take the mixing distribution support to be $\mathbb{U} = [\ell, L]$, with $\ell = 10^{-5}$ and $L = \max(X)$. The initial guess $P_0$ is taken to be uniform on $\mathbb{U}$. To reduce the dependence of the PR estimator on the data order, we average the PR estimates over 25 random permutations of the data. As for the weights, we (mostly) follow Condition 1 and take $w_i = (1+i)^{-1}$.

First, we consider data coming from a study of suicide risks reported in Silverman (1986), which consists of lengths of psychiatric treatment for $n = 86$ patients used as control. As per the detailed study of suicide risks in Copas and Fryer (1980), there is a higher risk for suicide in the early stages of treatment, so modeling these data with a monotone density is appropriate. Figure 1 shows a comparison of the four monotone density estimation methods discussed above with PR over a histogram of the data. PR gives a smooth estimate of the monotone density in a very short amount of time, much faster than the Bayes and empirical Bayes estimates that require Markov chain Monte Carlo. The take-away message is that, PR's misspecification bias—due to the choice of $\ell$ and $L$—can be easily controlled and that it gives a high-quality estimate compared to the other four methods. In fact, the PR estimate in this case is smoother than that of the other four methods, a desirable feature in applied data analysis.

Second, we consider two true monotone densities $m^\star$, namely, the half standard normal and a uniform mixture of a $\mathsf{Beta}(u \mid 2, 2)$ mixing density. We carry out the simulation study over sample sizes of $n = 50, 100, 200$. For each $n$, we generate 200 data sets of size $n$ and produce the five different estimates on each data set. As our metric of comparison, we use the total variation (or $L_1$) distance between the true density and the estimate. Additionally since inconsistency of the Grenander estimate at the origin is a well-known complication we also look at the ratio $\hat{m}(0)/m^\star(0)$ for each method. Boxplots summarizing both the $L_1$ distance and the at-the-origin ratio for the two simulations are shown in Figures 2 and 3. Notably, performance of PR is better than the Grenander estimator over all sample sizes. It is also faster and with slightly better performance when compared to the two Bayesian estimates, and is comparable to the

[1] `https://www4.stat.ncsu.edu/~rmartin/Codes/ebmono.R`

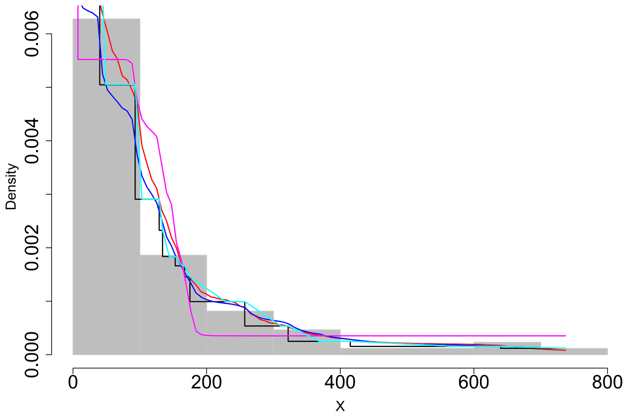
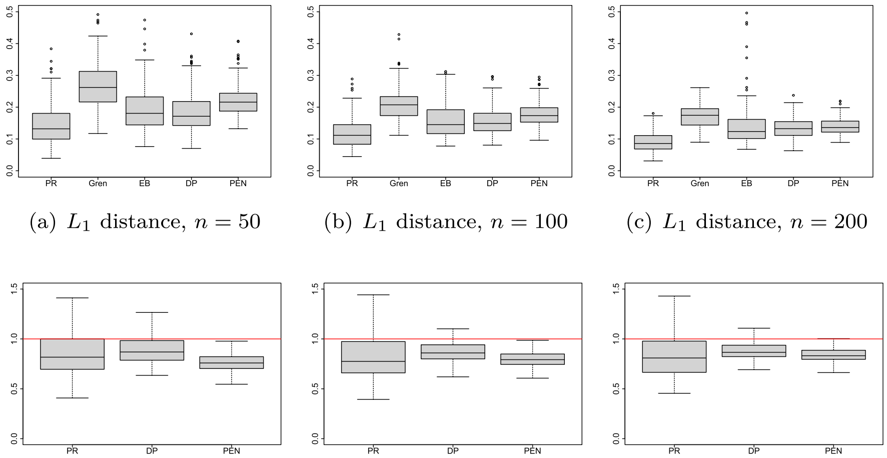


Fig 1. *A monotone density is fit to the suicide risk data from Silverman (1986) with the four different methods: PR (red), Grenander (black), empirical Bayes (blue), Bayes (magenta), and penalized likelihood (cyan).*

penalized likelihood estimate. For estimating the density at 0, we compare PR with only the best-performing methods, namely the penalized nonparametric MLE near 0 and the DP mixture Bayes estimator. Even though PR is not tailored specifically for estimation of $m^\star(0)$, as the penalized nonparametric MLE is, its performance is very competitive with these other methods.

Since the primary motivation for PR is estimation of a mixing distribution, next we compare the estimates of $P$ based on PR and the nonparametric MLE in the example with $P^\star = \mathsf{Beta}(2,2)$. To show a comparison of PR versus a standard nonparametric MLE, we give the estimates of the distribution function corresponding to $P^\star$ for both methods in Figure 4. Here PR's advantage of providing a smooth estimate is clear. The nonparametric MLE has no density, but we can easily get PR's mixing density estimate. Figure 5 shows the PR mixing density estimates compared to the true density for $P^\star$. While the estimates can be wiggly for a given data set—a result of there being limited information about the distribution of the latent $U_i$'s in the data—the PR estimator is clearly very accurate on average.

Finally, since the other methods, e.g., the nonparametric MLE, are tailored more towards cases where the underlying mixing distribution is discrete, a reviewer asked how the PR estimator—which typically would not be aware of the discreteness—might fare in such a case. To investigate this, we consider the case where $P^\star$ is a $\mathsf{Bin}(5, \frac{1}{2})$ distribution, shifted to be supported on $\mathbb{U} = \{1, \ldots, 6\}$. Then the true density $m^\star$ is the corresponding $P^\star$-mixture of uniform kernels, which is piecewise constant. Figure 6 shows a plot of the PR and Grenander mixture density estimates based on a sample of size $n = 200$ from this model. The PR estimate is relatively smooth whereas the Grenander estimator is, of course, piecewise constant, similar to the true $m^\star$. Surprisingly, having the piecewise constant structure correctly specified does not obviously lead to a better



(a) $L_1$ distance, $n = 50$ (b) $L_1$ distance, $n = 100$ (c) $L_1$ distance, $n = 200$

(d) $\hat{m}(0)/m^\star(0)$, $n = 50$ (e) $\hat{m}(0)/m^\star(0)$, $n = 100$ (f) $\hat{m}(0)/m^\star(0)$, $n = 200$

FIG 2. *True monotone density is a* $\mathsf{Beta}(u \mid 2, 2)$ *mixture of uniform kernels.*

estimator—Grenander seems to be over-selecting the number of constant intervals, failing to capture the piecewise constant structure in $m^\star$. This observation is not specific to this particular simulated data set: we repeated the above experiment 200 times and found that the average Kullback–Leibler divergence from $m^\star$ was 0.015 for PR and 0.046 for Grenander.

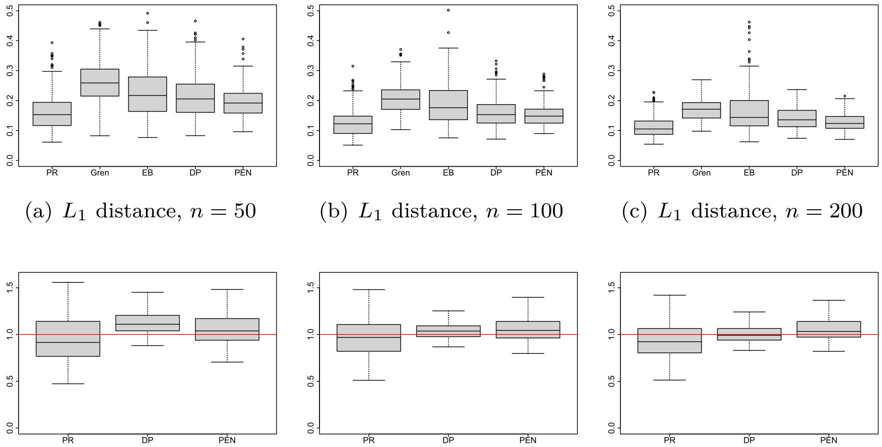
### 4.4.2. *Multiple testing problem with p-values*

Consider the large-scale significance testing problem like those that first attracted attention in Efron (2004, 2008). These are common in genomics applications where the goal is to determine, among a large collection of $n$ many genes, which ones are "interesting" in the sense of, say, having non-negligible association with a particular observable phenotype. Since the number of genes is very large, and it is believed that a relatively small number of those genes are actually "interesting," it is of practical importance to quickly screen the data to identify a relatively small set of genes that deserve more careful investigation. This determines a large-scale significance testing problem, where the goal is to test the sequence of null hypotheses,

$$H_{0i} : \text{gene } i \text{ is uninteresting} \quad i = 1, \ldots, n$$

Let $X_i$ denote the p-value associated with an individual test of $H_{0i}$. We can treat the collection of p-values $X_1, \ldots, X_n$ as the data available for the large-scale significance test. Following Genovese and Wasserman (2002) and others, it makes sense to model the p-values as a *two-groups* mixture

$$m(x) = \pi\, k_1(x \mid 1) + (1 - \pi) \int k_1(x \mid u)\, P_{\text{non-null}}(du).$$



(a) $L_1$ distance, $n = 50$ (b) $L_1$ distance, $n = 100$ (c) $L_1$ distance, $n = 200$

(d) $\hat{m}(0)/m^{\star}(0)$, $n = 50$ (e) $\hat{m}(0)/m^{\star}(0)$, $n = 100$ (f) $\hat{m}(0)/m^{\star}(0)$, $n = 200$

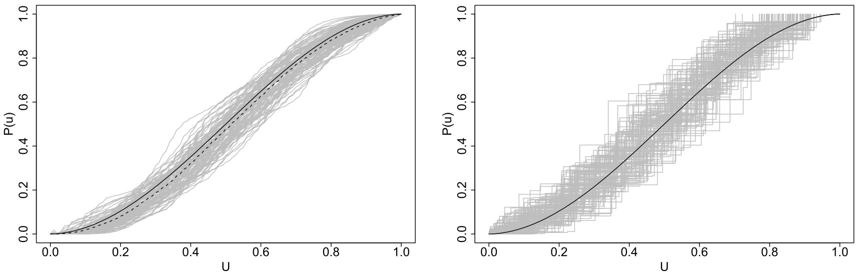
FIG 3. *True monotone density is half standard normal.*

The intuition is that the uninteresting, null genes have p-values distributed as $\mathsf{Unif}(0,1)$, whereas the interesting, non-null genes will have p-values stochastically smaller than $\mathsf{Unif}(0,1)$ with a non-increasing density.

Fitting this mixture model can proceed in one of several different ways, but this is non-trivial. The reason is that one does not have access to the only-non-null data that carries direct information about the non-null component. With PR, this is straightforward to address because we can simply choose the underlying dominating measure for the mixture to be Lebesgue measure on $[0,1]$ plus a point mass at the endpoint 1. Then the right-hand side of the above display is just a version of (1) with mixing distribution $P$ being absolutely continuous with respect to $\text{Leb}_{[0,1]} + \delta_{\{1\}}$. The PR solution can easily handle this and it produces an estimate $\pi_n$ of the null proportion, the density $p_{\text{non-null}}(u) = dP_{\text{non-null}}/du$, and of course the overall mixture density $m_n(x)$.

For an illustration of this idea, consider the famous hereditary breast cancer study by Hedenfalk et al. (2001). A histogram of the p-values associated with the $n = 3226$ genes under investigation is shown in Figure 7, where the two-groups mixture model structure is apparent. We fit the above mixture model to these data using PR—with $P_0$ initialized as a mass $\pi_0 = 0.8$ at 1 plus 0.2 times $\mathsf{Unif}(0,1)$ distribution—and it returns the estimates as overlaid in Figure 7. In particular, PR estimates the null weight as $\pi_n = 0.920$ and it is clear that the mixture is able to capture the spike in p-values at the origin corresponding to the non-null cases. For a PR-driven rule to select interesting genes, we can apply the *local false discovery rate* thresholding procedure advocated for by Efron and others. Indeed, the PR estimate of the local false discovery rate is

$$\widehat{\text{lfdr}}(x) = \frac{\pi_n}{m_n(x)}.$$



(a) PR

(b) Nonparametric MLE

Fig 4. *Plots of the mixing distribution function estimates (gray) using PR and the nonparametric MLE over 100 data sets of size* $n = 200$*, where the true mixing distribution function (black) corresponds to* $P^\star = \mathsf{Beta}(2,2)$.

The genes that are "interesting" ought to have small local false discovery rate values, so a decision rule would announce gene $i$ as interesting if $\widehat{\text{lfdr}}(x_i) \leq r$ for a suitable threshold $r$. A plot of the estimated local false discovery rate curve is shown on the negative axis in Figure 7. If we choose a threshold of $r = 0.15$, then we flag 222 genes as interesting, which includes 26 out of the 27 cases identified in Lee et al. (2003) based on known biological connections to breast cancer mutations.

## 5. Conclusion

Estimation of mixing distributions in mixture models is a challenging problem, one for which there are very few satisfactory methods available. To our knowledge, the PR algorithm is the one general method available that is both fast and capable of nonparametrically estimating a mixing distribution having a density with respect to any user-specified dominating measure. Despite the simple and fast implementation of the PR algorithm, and the strong empirical performance observed in numerous applications, its theoretical analysis and justification is non-trivial because of the recursive structure. Previous work has established consistency of the PR estimates under relatively strong conditions. Most concerning is that there are known examples, such as monotone density estimation using uniform mixtures, for which the sufficient conditions in previous work do not hold. The main focus of the present paper was to weaken those overly-strong conditions in order to broaden the range of problems in which PR can be applied. In particular, the new sufficient conditions can be checked for mixtures of uniform kernels, which puts PR in a position to solve the non-trivial problem of monotone density estimation on $[0, \infty)$.

There are a number of possible extensions and/or open problems that could be considered. First, from a practical or methodological point of view, there is a natural extension of the motivating monotone density estimation application. That is, what can be done if the location of the mode itself is unknown? This

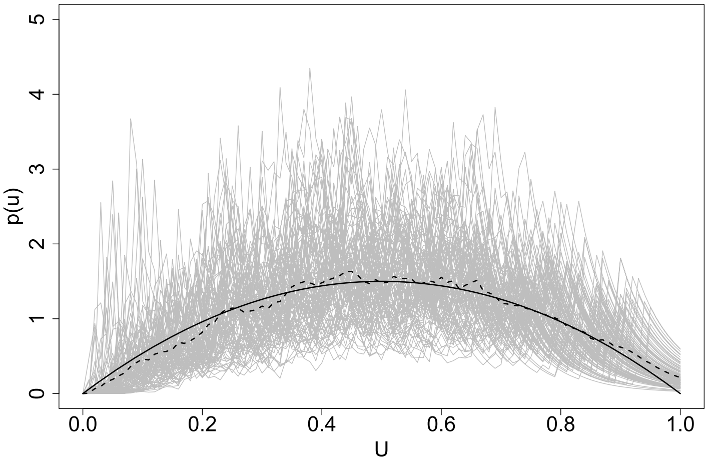


Fig 5. *Plot of the PR mixing density estimates $p_n$ (gray) over 100 datasets, average $p_n$ over the 100 datasets (dashed) against the true density $p$ (black) when $n = 200$ observations are drawn from a uniform mixture of $p$.*

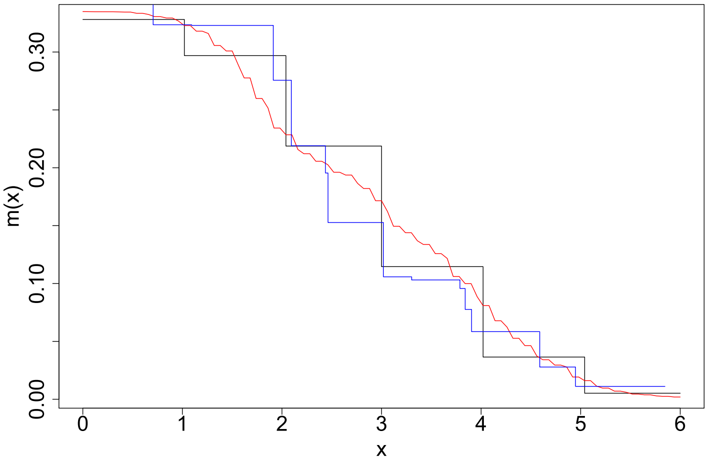


Fig 6. *A piecewise constant mixture density $m^\star$ (black) with the Grenander estimate (blue) and the PR estimate (red).*

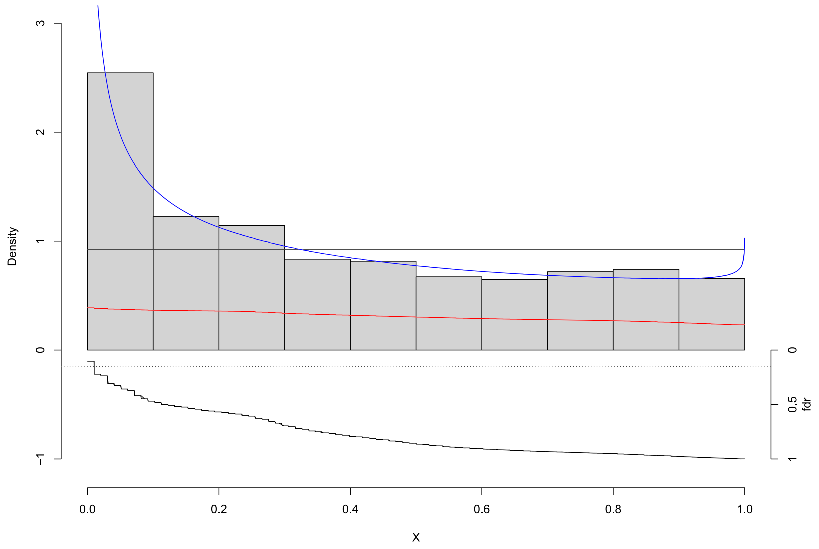


Fig 7. *Histogram of the hereditary breast cancer study dataset with the fitted $m_n$ (blue), estimated $m_0$ (black) and $m_1$ (red). The negative axis shows $\widehat{lfdr}(x)$ with a threshold at $r = 0.15$ identifying the interesting cases.*

is a non-trivial problem and has been investigated by a number of researchers, including Liu and Ghosh (2020). In the PR framework, the natural approach would be to treat the mode as an unknown, non-mixing parameter contained in the kernel, and apply the PR marginal likelihood strategy in Martin and Tokdar (2011) to estimate both the mode and the mode-specific mixing distribution. How this proposal compares to existing methods remains to be investigated.

Second, from a theoretical point of view, it may be undesirable to work with a fixed and compact mixing distribution support $\mathbb{U}$. A natural extension would be to introduce a type of sieve, to allow the support to depend on the sample size, i.e., $\mathbb{U} = \mathbb{U}_n$. The use of a $n$-dependent support $\mathbb{U}_n$, however, is difficult and awkward in the context of PR. First, unlike usual likelihood-based methods that assume all the data to be available at once, PR is technically meant to be used for recursive estimation with online data. In that case, having a sample size dependent support is unnatural since the sample size is not set in advance. But even if we ignore PR's recursive structure and treat it as being applied to batch data, the analysis is based on martingales that do implicitly treat the data points one by one in a sequence, so having any $n$-specific components in the algorithm itself is awkward. Beyond awkwardness, there is a specific technical obstacle. Much of the analysis depends on properties of the functional $T$ defined in (7). This functional depends on $\mathbb{U}$ and so, if $\mathbb{U}$ is made to depend on $n$, then we end up with a sequence, $T_n$, of functionals that are applied to the PR sequence of estimates, $P_n$, so new techniques would be needed in order to analyze a sequence of random variables like $T_n(P_{n-1})$.

One more remark concerning future PR-related research directions is worth making here. It concerns PR's rate of convergence. Certain "worst-case" rate results are available in Martin and Tokdar (2009), and these can surely be extended to the broader context considered here, but key insights are still lacking. In particular, it is not clear how the PR results can incorporate the smoothness assumptions about $P^\star$ or $m^\star$ necessary to improve on the "worst-case" rates available. For example, in the monotone density estimation case, it is well-known that the optimal $L_1$ rate for estimation of $m^\star$ is $n^{-1/3}$, but this assumes that $m^\star$ is differentiable (e.g., Kim and Pollard, 1990, Example 6.5). Without that assumption, the optimal rate is slower, and it is this slower rate that the existing PR rate results can be compared to. To be clear, it is not that PR actually converges at such a slow rate, it is that the currently available proof techniques are lacking.

## Appendix A: Proofs

### *A.1. Proof of Theorem 1*

We start by reviewing some details from the analysis in Martin and Tokdar (2009). From the recursive form of the PR estimate of the mixing distribution, and the linearity of the mixture model, clearly a similar recursive form holds for the mixture. That is,

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

where

$$h_{n,y}(x) = \frac{\int k(x \mid u)\, k(y \mid u)\, P_{n-1}(du)}{m_{n-1}(y)}, \quad x, y \in \mathbb{X}.$$

For later, define the function $H_{n,y}(x)$ as

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

By Taylor's theorem, we can write

$$\log(1 + x) = x - x^2 R(x), \quad x > -1,$$

where the remainder term $R$ satisfies $0 \leq R(x) \leq \max\{1, (1 + x)^{-2}\}$. This remainder bound will be important later.

Let $K_n = K(m^\star, m_n)$. Then from that recursive form of the mixture density updates above, and this Taylor approximation, it can be shown that

$$\begin{aligned} K_n = K_{n-1} - w_n \int H_{n,X_n}(x)\, m^\star(x)\, dx \\ + w_n^2 \int H_{n,X_n}^2(x)\, R(w_n H_{n,X_n}(x))\, m^\star(x)\, dx. \end{aligned}$$

Next, let $\mathcal{A}_r$ denote the $\sigma$-algebra generated by data $X_1, \ldots, X_r$, for $r \geq 1$. Now take conditional expectation of the above display, given $\mathcal{A}_{n-1}$, to get

$$\mathsf{E}(K_n \mid \mathcal{A}_{n-1}) = K_{n-1} - w_n T(P_{n-1}) + w_n^2 E(Z_n \mid \mathcal{A}_{n-1}), \tag{14}$$

where

$$\begin{aligned} T(\Phi) &= \int_{\mathbb{U}} \left\{ \int_{\mathbb{X}} \frac{m^\star(x)}{m_\Phi(x)} k(x \mid u)\, dx \right\}^2 \Phi(du) - 1 \\ Z_n &= \int_{\mathbb{X}} H_{n,X_n}^2(x)\, R(w_n H_{n,X_n}(x))\, m^\star(x)\, dx. \end{aligned}$$

If we let $K_n^\star = K_n - K(m^\star, m^\dagger)$, then the same relationship as in (14) holds, i.e.,

$$\mathsf{E}(K_n^\star \mid \mathcal{A}_{n-1}) = K_{n-1}^\star - w_n T(P_{n-1}) + w_n^2 \mathsf{E}(Z_n \mid \mathcal{A}_{n-1}). \tag{15}$$

Surprisingly, this form is nice—it is an *almost supermartingale* like those studied by Robbins and Siegmund (1971). Below we restate (a simple version of) Robbins and Siegmund's main theorem for the reader's convenience.

**Robbins–Siegmund Theorem.** *Consider a sequence of non-negative random variables $(M_n, \zeta_n, \xi_n)$, where $(M_n)$ is adapted to a filtration $(\mathcal{A}_n)$. If*

$$\begin{gathered} \mathsf{E}(M_n \mid \mathcal{A}_{n-1}) \leq M_{n-1} - \zeta_{n-1} + \xi_{n-1} \\ \textstyle\sum_n \xi_n < \infty \quad \textit{almost surely,} \end{gathered} \tag{16}$$

*then $M_n$ converges and $\sum_n \zeta_n < \infty$ almost surely.*

The equation in (15) satisfies the criterion in (16), where $\zeta_{n-1} = w_n T(P_{n-1})$ and $\xi_{n-1} = w_n^2 \mathsf{E}(Z_n \mid \mathcal{A}_{n-1})$. We need to check that $\sum_n w_n^2 \mathsf{E}(Z_n \mid \mathcal{A}_{n-1})$ is finite almost surely, which amounts to getting a suitable upper bound on $Z_n$ and its conditional expectation. Here is where our analysis starts to differ from that in Martin and Tokdar (2009).

The most complicated part of $Z_n$ is its dependence on the Taylor approximation remainder described above. Recalling that upper bound, we have

$$R(w_n H_{n,X_n}(x)) \leq \max[1, \{1 + w_n H_{n,X_n}(x)\}^{-2}].$$

But since $h_{n,X_n}$ and $m_{n-1}$ are density functions, their ratio is non-negative, so

$$w_n H_{n,X_n}(x) = w_n \Bigl( \frac{h_{n,X_n}(x)}{m_{n-1}(x)} - 1 \Bigr) \geq -w_n > -w_1.$$

Therefore, $R(w_n H_{n,X_n}(x)) \leq \max\{1, (1 - w_1)^{-2}\}$, a constant, so

$$Z_n \lesssim \int H_{n,X_n}^2(x)\, m^\star(x)\, dx \leq 1 + \int \Bigl( \frac{h_{n,X_n}(x)}{m_{n-1}(x)} \Bigr)^2 m^\star(x)\, dx.$$

Since we only need to get an upper bound up to a multiplicative constant, we will ignore that constant lumped inside of "$\lesssim$" in what follows; we will also ignore the leading "1+" since the bound will ultimately get multiplies by $w_n^2$, which itself is summable by assumption. From this bound, plug in the definition of $h_{n,X_n}$ to get

$$\begin{aligned} Z_n &\leq \int \Bigl\{ \frac{\int k(x \mid u)\, k(X_n \mid u)\, P_{n-1}(du)}{m_{n-1}(x)\, m_{n-1}(X_n)} \Bigr\}^2 m^\star(x)\, dx \\ &\leq \int \frac{\int k^2(x \mid u)\, k^2(X_n \mid u)\, P_{n-1}(du)}{m_{n-1}^2(x)\, m_{n-1}^2(X_n)}\, m^\star(x)\, dx, \end{aligned}$$

where the second inequality is by Cauchy–Schwarz. Next, we focus on one of the terms in the denominator, say, $m_{n-1}(x)$. From that recursive form for the mixture density updates, we immediately see that

$$m_{n-1}(x) \geq (1 - w_{n-1})\, m_{n-2}(x) \geq \cdots \geq m_0(x) \prod_{i=1}^{n-1} (1 - w_i), \quad \text{any } x.$$

Plug in this lower bound for both terms in the denominator of the bound for $Z_n$ to get

$$Z_n \leq \prod_{i=1}^{n-1} (1 - w_i)^{-4} \int \frac{\int k^2(x \mid u)\, k^2(X_n \mid u)\, P_{n-1}(du)}{m_0^2(x)\, m_0^2(X_n)}\, m^\star(x)\, dx.$$

Now take conditional expectation with respect to $\mathcal{A}_{n-1}$ and interchange the order of integration (which is allowed since the integrand is non-negative) to get

$$\mathsf{E}(Z_n \mid \mathcal{A}_{n-1}) \leq \prod_{i=1}^{n-1} (1 - w_i)^{-4} \int \Bigl\{ \int \frac{k^2(x \mid u)}{m_0^2(x)}\, m^\star(x)\, dx \Bigr\}^2 P_{n-1}(du).$$

By Condition 3, we have that the expression inside curly braces above is bounded, uniformly in $u$, by a constant. Therefore,

$$\mathsf{E}(Z_n \mid \mathcal{A}_{n-1}) \lesssim \prod_{i=1}^{n-1} (1 - w_i)^{-4}.$$

Next we used the assumed form of the weight sequence, in Condition 1, to bound the above product. In general, we have

$$\log \prod_{i=1}^{n-1} (1 - w_i)^{-4} = -4 \sum_{i=1}^{n-1} \log(1 - w_i).$$

Using the standard bound, $-\log(1 - w) \geq w(1 - w)^{-1}$, and the fact that the $w_i$'s are decreasing, we have

$$\log \prod_{i=1}^{n-1} (1 - w_i)^{-4} = -4 \sum_{i=1}^{n-1} \log(1 - w_i) \leq \frac{4}{1 - w_1} \sum_{i=1}^{n-1} w_i.$$

According to Condition 1, $w_i = a(i+1)^{-1}$, the summation in the above expression is of the order $\log n$, which implies

$$\prod_{i=1}^{n-1} (1 - w_i)^{-4} \leq n^{8a/(2-a)}.$$

Putting everything together, we get

$$w_n^2 \mathsf{E}(Z_n \mid A_{n-1}) \lesssim n^{-2+8a/(2-a)}.$$

Since $a < \frac{2}{9}$, the exponent is less than $-1$, hence the upper bound is summable almost surely, thus verifying the hypothesis of the Robbins–Siegmund theorem. Consequently, we can conclude that

$$K_n^\star \to K_\infty^\star \quad \text{and} \quad \sum_n w_n T(P_{n-1}) < \infty, \quad \text{almost surely.}$$

It remains to show that the limit, $K_\infty^\star$ is 0 almost surely.

The key to proving this last claim is a special property of the $T$ function. For a generic mixing distribution $P$, supported on $\mathbb{U}$, rewrite $T$ as

$$T(P) = \int (g_P - 1)^2 \, dP,$$

where

$$g_P(u) = \int \frac{k(x \mid u)}{m_P(x)} \, m^\star(x) \, dx.$$

For any bounded and continuous function $h : \mathbb{U} \to \mathbb{R}$, it follows from the standard bound $|\int \cdots du| \leq \int |\cdots| \, du$ and Cauchy–Schwartz that

$$\Bigl| \int (g_P - 1) \, h \, dP \Bigr|^2 \leq \Bigl\{ \int |g_P - 1| \, |h| \, dP \Bigr\}^2 \leq T(P) \int h^2 \, dP. \tag{17}$$

This implies the lower bound

$$T(P) \geq \sup_{h : \int h^2 \, dP = 1} \Bigl\{ \int (g_P - 1) \, h \, dP \Bigr\}^2,$$

where the supremum is over all bounded and continuous functions $h$ with $\int h^2 \, dP = 1$. For an alternative look at the integral in the curly braces above, define the operator $\phi$ that maps a probability measure $P$ on $\mathbb{U}$ to a new probability measure, $\phi(P)$, on $\mathbb{U}$ according to the formula

$$\phi(P)(A) = \int_A g_P(u) \, P(du), \quad A \subseteq \mathbb{U}, \text{ measurable.}$$

Then that expression in curly braces is simply

$$\int h \, d\phi(P) - \int h \, dP.$$

A consequence of the Robbins–Siegmund theorem is that $\sum_n w_n T(P_{n-1}) < \infty$ almost surely. Since $w_n$ itself is vanishing too slowly to be summable, it must be that there exists a subsequence $P_{n(t)}$ such that $T(P_{n(t)}) \to 0$ almost surely. Therefore,

$$\sup_{h:\int h^2 \, dP_{n(t)}=1} \Big\{ \int h \, d\phi(P_{n(t)}) - \int h \, dP_{n(t)} \Big\}^2 \to 0, \quad \text{almost surely.}$$

Since the original sequence $P_n$ is tight, there is a sub-subsequence $P_{n(t_s)}$ with a weak limit, and the above result implies that the limit is a fixed point of $\phi$. However, the only fixed points of this mapping are Kullback–Leibler minimizers, say, $P^\dagger$; see, for example, Lemma 3.4 in Shyamalkumar (1996). This implies $K^\star_{n(t_s)}$ is vanishing almost surely. However, by the Robbins–Siegmund theorem, we have that the original sequence $K^\star_n$ converges almost surely to some $K^\star_\infty$. But if the original sequence has a limit and the limit is 0 on a subsequence, then it must be that $K^\star_\infty = 0$ almost surely. Putting everything together, we have shown that $K^\star_n = K(m^\star, m_n) - K(m^\star, m^\dagger) \to 0$ almost surely, which implies $K(m^\star, m_n) \to K(m^\star, m^\dagger)$, and completes the proof.

### *A.2. Proof of Lemma 1*

The proof proceeds in two steps. First we express the modified target $m^{\star L}$ as a uniform mixture and identify the corresponding mixing distribution, denoted by $P^{\star L}$. Then we solve the optimization problem that consists of identifying the mixing distribution, $P^\dagger = P^{\dagger \ell, L}$, supported on $\mathbb{U} = [\ell, L]$, that minimizes $P \mapsto K(m^{\star L}, m_P)$.

First, recall the definition of $m^{\star L}$,

$$m^{\star L}(x) = \frac{m^\star(x) \, 1_{[0,L]}(x)}{M^\star(L)}, \quad x \in [0, \infty),$$

where $M^\star$ is the distribution function corresponding to the density $m^\star$. By direct calculation, for the denominator we have

$$M^\star(L) = P^\star([0,L]) + L m^\star(L).$$

The numerator can also be rewritten as

$$m^\star(x) \, 1_{[0,L]}(x) = \int_0^L k_1(x \mid u) \, P^\star(du) + m^\star(L)$$

After a bit of algebra to simplify the ratio of the sums in the previous two displays, we are able to write $m^{\star L}$ as a mixture

$$m^{\star L}(x) = \int k_1(x \mid u) \, P^{\star L}(du), \tag{18}$$

where

$$P^{\star L} = \pi \, \widetilde{P}^{\star L} + (1 - \pi) \, \delta_{\{L\}}, \tag{19}$$

with $\pi$ and $\widetilde{P}^{\star L}$ defined as

$$\pi = \frac{P^\star([0,L])}{P^\star([0,L]) + Lm^\star(L)} \quad \text{and} \quad \widetilde{P}^{\star L}(du) = \frac{P^\star(du)1_{[0,L]}(u)}{P^\star([0,L])}.$$

That is, $m^{\star L}$ is a uniform mixture, where the mixing distribution $P^{\star L}$ is a convex combination $P^\star$ renormalized to $[0,L]$ and a point mass at $L$.

For step 2, we want to find the minimizer of $P \mapsto \kappa(P) := K(m^{\star L}, m_P)$, over all mixing distributions supported on $\mathbb{U} = [\ell, L]$, where $m^{\star L}$ has the mixture form presented above. Using the above notation, the lemma's claim is that the minimizer is

$$P^\dagger = \omega\, \delta_{\{\ell\}} + P^{\star L}|_{\mathbb{U}},$$

where $P^{\star L}|_{\mathbb{U}}$ is $P^{\star L}$ restricted (but not renormalized) from $[0,L]$ to $\mathbb{U} = [\ell, L]$, and $\omega = P^{\star L}([0,\ell])$. If we can show that the Gateaux derivative of $\kappa$, evaluated at $P^\dagger$, in the direction of any other distribution $H$ on $\mathbb{U}$, is vanishing, then we will have proved the claim. The Gateaux derivative at a generic $P$, in the direction of $H$, is

$$\frac{d}{dt}\kappa((1-t)P + tH)\Big|_{t=0} = \int_0^L \Big\{1 - \frac{m_H(x)}{m_P(x)}\Big\}\, m^{\star L}(x)\, dx.$$

Let $m^\dagger = m_{P^\dagger}$, which has the form

$$m^\dagger(x) = \omega\, k_1(x \mid \ell) + \int_\ell^L k_1(x \mid u)\, P^{\star L}(du).$$

Then the goal is to show that

$$\int_0^L \Big\{1 - \frac{m_H(x)}{m^\dagger(x)}\Big\}\, m^{\star L}(x)\, dx = 0 \quad \text{for all } H \text{ supported on } \mathbb{U},$$

or, equivalently, to show that

$$1 - \int_0^\ell \frac{m_H(x)}{m^\dagger(x)}\, m^{\star L}(x)\, dx - \int_\ell^L \frac{m_H(x)}{m^\dagger(x)}\, m^{\star L}(x)\, dx = 0 \tag{20}$$

On the interval $x \in (\ell, L]$, it is clear that $m^\dagger(x) = m^{\star L}(x)$, so

$$\int_\ell^L \frac{m_H(x)}{m^\dagger(x)}\, m^{\star L}(x)\, dx = \int_\ell^L m_H(x)\, dx. \tag{21}$$

Next, since both $P^\dagger$ and $H$ are supported on $\mathbb{U} = [\ell, L]$, the two mixture densities $m^\dagger$ and $m_H$ are constant on the interval $x \in [0, \ell]$. This implies

$$\int_0^\ell \frac{m_H(x)}{m^\dagger(x)}\, m^{\star L}(x)\, dx - \int_0^\ell m_H(x)\, dx = \frac{m_H(0)}{m^\dagger(0)} \int_0^\ell \{m^{\star L}(x) - m^\dagger(x)\}\, dx.$$

We claim that the integral on the right-hand side is 0. To see this, first integrate $m^\dagger$:

$$\begin{aligned}\int_0^\ell m^\dagger(x)\,dx &= \omega + \int_0^\ell \int_\ell^L k_1(x \mid u)\,P^{\star L}(du)\,dx \\ &= \omega + \ell m^{\star L}(\ell) \\ &= \frac{P^\star([0,L]) + \ell m^\star(\ell)}{M^\star(L)}.\end{aligned}$$

Similarly, integrate $m^{\star L}$:

$$\begin{aligned}\int_0^\ell m^{\star L}(x)dx &= \frac{1}{M^\star(L)} \int_0^\ell \Bigl\{\int_0^L k_1(x \mid u)\,P^\star(du) + m^\star(L)\Bigr\}dx \\ &= \frac{1}{M^\star(L)}\Bigl\{\int_\ell^L (\ell/u)P^\star(du) + \int_0^\ell P^\star(du) + \ell m^\star(L)\Bigr\} \\ &= \frac{1}{M^\star(L)}\bigl\{\ell m^\star(\ell) - \ell m^\star(L) + P^\star([0,\ell]) + \ell m^\star(L)\bigr\} \\ &= \frac{P^\star([0,\ell]) + \ell m^\star(\ell)}{M^\star(L)}.\end{aligned}$$

Clearly the two integrals above are the same, which implies that

$$\int_0^\ell \{m^{\star L}(x) - m^\dagger(x)\}\,dx = 0,$$

and, consequently, that

$$\int_0^\ell \frac{m_H(x)}{m^\dagger(x)}\,m^{\star L}(x)\,dx = \int_0^\ell m_H(x)\,dx. \tag{22}$$

Plugging the relations (21) and (22) into the left-hand side of (20) proves the claim, i.e., that the Gateaux derivative of $\kappa$ at $P^\dagger$ vanishes in all directions $H$, which implies that $P^\dagger$ is the minimizer of the Kullback–Leibler divergence.

### *A.3. Proof of Theorem 2*

To prove $K(m^{\star L}, m_n) \to K(m^{\star L}, m^\dagger)$, we apply Theorem 1. Condition 1 is in the user's control and, hence, is easy to satisfy. Condition 2 requires the support of the mixing distribution to be compact, which is clearly satisfied by $\mathbb{U} = [\ell, L]$. Condition 3 is the only non-trivial condition, and it requires

$$\sup_{u \in [\ell,L]} \int_0^L \Bigl\{\frac{k_1(x \mid u)}{m_0(x)}\Bigr\}^2 m^{\star L}(x)\,dx < \infty,$$

where $m_0$ is the mixture density corresponding to the initial guess, $P_0$, which contains point masses. The key point is, thanks to the point mass at $L$,

$$m_0(x) \geq p_{0,L}\, k_1(x \mid L) = p_{0,L}\, L^{-1}, \quad x \in [0, L].$$

Since the denominator above is uniformly bounded away from 0, and, similarly, the numerator is uniformly bounded by $\ell^{-1}$, Condition 3 clearly holds.

Next, the claim about convergence of $m_n$ to $m^\dagger$ in total variation follows immediately from Corollary 1 and the fact that $m^{\star L}$ is bounded away from 0. Finally, for the claim about weak convergence of $P_n$ to $P^\dagger$, we apply Corollary 2. We have already stated that $m^\dagger / m^{\star L} \in L_\infty$ since $m^{\star L}$ is bounded away from 0. So all that remains is to check that the uniform kernel satisfies the abstract condition (8), which we do next.

Imagine a generic sequence of mixing distributions $Q_t$ supported on $\mathbb{U} = [\ell, L]$ and assume they converge weakly to $Q_\infty$. The condition (8) concerns the behavior of the mixture density $m_{Q_t}(x)$. Note that the uniform kernel is not a continuous function in $u$ for a given $x$, but it is upper-semicontinuous. Recall that the mixture densities are constant for $x \in [0, \ell]$. This means that the value of the mixture density on a set of positive measure is determined by its value at $x = \ell$, so some care will be needed below; in particular, we'll have to deal with the cases $x \in [0, \ell]$ and $x \in (\ell, L]$ separately.

Start with the case $x \in (\ell, L]$. The kernel $u \mapsto k_1(x \mid u)$ is bounded and continuous except for the jump discontinuity at $u = x$. It is possible that the limit $Q_\infty$ of the sequence $Q_t$ of mixing distributions puts positive mass at $u = x$, i.e., that $x$ is a discontinuity point of $Q_\infty$. In such cases, $m_{Q_t}(x)$ may not converge or, even if it does converge, the limit may not equal $m_{Q_\infty}(x)$. However, $Q_\infty$'s set of discontinuity points has Lebesgue measure 0. For any $x \in (\ell, L]$ that is not a discontinuity point of $Q_\infty$, the kernel is effectively bounded and continuous, so $Q_t \to Q_\infty$ weakly implies $m_{Q_t}(x) \to m_{Q_\infty}(x)$. This verifies (8) for the range $x \in (\ell, L]$.

For the case $x \in [0, \ell]$, again, we know that the mixture density is constant in $x$. Therefore, if there is an issue with convergence of the mixture density at $x = \ell$, then that implies an issue on a set of positive Lebesgue measure, hence (8) fails. However, while the kernel is only upper-semicontinuous in general, $u \mapsto k_1(\ell \mid u)$ is bounded and continuous on the support of the $Q_t$ sequence, so we get $m_{Q_t}(\ell) \to m_{Q_\infty}(\ell)$ automatically from the definition of weak convergence. This implies the same for all $x \in [0, \ell]$, so (8) holds there too.

### *A.4. Proof of Proposition 1*

By the triangle inequality, we have

$$\int |m^\dagger - m^\star|\, dx \leq \int |m^\dagger - m^{\star L}|\, dx + \int |m^{\star L} - m^\star|\, dx. \tag{23}$$

Now we consider each term in the upper bound (23) separately. Start with the second term, splitting up the range of integration, we immediately get

$$\int |m^{\star L} - m^\star|\, dx = \int_0^L \Bigl| \frac{m^\star}{M^\star(L)} - m^\star \Bigr|\, dx + 1 - M^\star(L)$$

$$
\begin{aligned}
&= \frac{1}{M^\star(L)} |1 - M^\star(L)| \int_0^L m^\star \, dx + 1 - M^\star(L) \\
&= 2\{1 - M^\star(L)\}.
\end{aligned}
$$

For the first term in (23), we borrow the calculations in the proof of Lemma 1 above. In particular, on the interval $x \in [\ell, L]$, the two densities are the same, but on the interval $x \in [0, \ell)$, the absolute difference between densities is bounded by

$$|m^{\star L}(x) - m^\dagger(x)| \leq \omega k_1(x \mid \ell) + \int_0^\ell k_1(x \mid u) \, P^{\star L}(du), \quad x \in [0, \ell).$$

Now integrate to get

$$
\begin{aligned}
\int |m^\dagger - m^{\star L}| \, dx &= \int_0^\ell |m^\dagger - m^{\star L}| \, dx \\
&\leq \omega + P^{\star L}([0, \ell]) \\
&= 2 \cdot \frac{P^\star([0, \ell])}{M^\star(L)}.
\end{aligned}
$$

Combining the two bounds proves the claim.

### *A.5. Proof of Proposition 2*

As shown in the proof of Theorem 2, $m_n(\ell) \to m^\dagger(\ell)$ almost surely with respect to $m^{\star L}$. Since $m_n(0) = m_n(\ell)$ and $m^\dagger(0) = m^\dagger(\ell)$ by Equation (10), the proof of the first claim is complete. To bound the bias, i.e., the difference between the quantity being estimated, $m^\dagger(0)$, and and the true density at the origin, $m^\star(0)$, we proceed as follows.

$$
\begin{aligned}
m^\dagger(0) - m^\star(0) &= a_\ell \ell^{-1} + a_{\mathbb{U}} \int_{\mathbb{U}} u^{-1} \, P^\star(du) + a_L L^{-1} - \int_0^\infty u^{-1} P^\star(du) \\
&= \Big\{ a_\ell \ell^{-1} - \int_0^\ell u^{-1} P^\star(du) \Big\} + \Big\{ (a_{\mathbb{U}} - 1) \int_{\mathbb{U}} u^{-1} \, P^\star(du) \Big\} \\
&\quad + \Big\{ a_L L^{-1} - \int_L^\infty u^{-1} P^\star(du) \Big\}.
\end{aligned}
$$

Using the definitions of $a_\ell$, $a_{\mathbb{U}}$, and $a_L$, the bound $P^\star([0, \ell]) \lesssim \ell$, and the fact that $\int_{\mathbb{U}} u^{-1} \, P^\star(du) = O(1)$ as a function of $(\ell, L)$, it is easy to check that each of the three terms on the right-hand side above can be bounded by $1 - M^\star(L)$. That is,

$$
\begin{aligned}
&a_\ell \ell^{-1} - \int_0^\ell u^{-1} \, P^\star(du) \lesssim M^\star(L)^{-1} - 1 \lesssim 1 - M^\star(L) \\
&(a_{\mathbb{U}} - 1) \int_{\mathbb{U}} u^{-1} \, P^\star(du) \lesssim 1 - M^\star(L)
\end{aligned}
$$

$$a_L L^{-1} - \int_L^\infty u^{-1}\, P^\star(du) \lesssim 1 - M^\star(L),$$

which completes the proof of the claim.

### A.6. Support for the claim in Remark 1

The claim is that the conditions of Theorem 1 can be checked for the scaled beta kernel

$$k_s(x \mid u) = s u^{-s} (u-x)^{s-1}\, 1_{[0,u]}(x), \quad x \in \mathbb{X}, \quad u \in \mathbb{U}.$$

The weights are in the user's control and the support $\mathbb{U} = [\ell, L]$ is compact, so we only need to verify Condition 3. The PR is assumed to be initialized at $P_0$ having a point mass at $L$, i.e.,

$$P_0 = (1 - p_{0,L})\, P_{0,\text{cont}} + p_{0,L}\, \delta_{\{L\}},$$

where $p_{0,L} \in (0,1)$ is fixed by the user. Then it is easy to see that

$$m_0(x) = \int k_s(x \mid u)\, P_0(du) \geq p_{0,L}\, k_s(x \mid L).$$

Condition 3 concerns the ratio $k_s(x \mid u)/m_0(x)$ which, in the present case, for $x \in [0, L]$ and $u \in [\ell, L]$, can be upper-bounded as

$$\frac{k_s(x \mid u)}{m_0(x)} \leq \frac{k_s(x \mid u)}{p_{0,L}\, k_s(x \mid L)} \leq \frac{L^s}{\ell^s}.$$

The above ratio is uniformly bounded by a constant, so we immediately get

$$\sup_{u \in [\ell, L]} \int_0^L \Big\{ \frac{k_s(x \mid u)}{m_0(x)} \Big\}^2 m^{\star L}(x)\, dx < \infty.$$

Since all the conditions of Theorem 1 have been verified, it follows that

$$K(m^{\star L}, m_n) - K(m^{\star L}, m^\dagger) \to 0, \quad m^{\star L}\text{-almost surely.}$$

## Acknowledgments

The authors are grateful to anonymous reviewers for their valuable feedback on a previous version of the manuscript.

## References

Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a $k$-monotone density: limit distribution theory and the spline connection. *Ann. Statist.* **35** 2536–2564. MR2382657

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics* **65** 198–205. MR2665861

Chen, J. (2017). Consistency of the MLE under mixture models. *Statist. Sci.* **32** 47–63. MR3634306

Copas, J. and Fryer, M. (1980). Density estimation and suicide risks in psychiatric treatment. *Journal of the Royal Statistical Society: Series A (General)* **143** 167–176.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability. Springer Texts in Statistics*. Springer, New York. MR2664452

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 1–22.

Dixit, V. and Martin, R. (2019). Permutation-based uncertainty quantification about a mixing distribution. *arXiv:1906.05349*.

Dixit, V. and Martin, R. (2022). Estimating a mixing distribution on the sphere using predictive recursion. *Sankhya B* **84** 596–626. MR4502743

Dvoretzky, A. (1956). On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I* 39–55. University of California Press, Berkeley-Los Angeles, Calif. MR0084911

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99** 96–104.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model.

Eggermont, P. and LaRiccia, V. (1995). Maximum smoothed likelihood density estimation for inverse problems. *The Annals of Statistics* **23** 199–220.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics* **19** 1257–1272. MR1126324

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517. MR1924303

Ghosh, J. K. and Tokdar, S. T. (2006). Convergence and consistency of Newton's algorithm for estimating mixing distribution. In *Frontiers in Statistics* 429–443. World Scientific.

Grenander, U. (1956). On the theory of mortality measurement: part II. *Scandinavian Actuarial Journal* **1956** 125–153.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser.* 539–555. Wadsworth, Belmont, CA. MR0822052

Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742

Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints. Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge University Press, New York. MR3445293

Hahn, P. R., Martin, R. and Walker, S. G. (2018). On recursive Bayesian predictive distributions. *Journal of the American Statistical Association* **113** 1085–1093.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M. et al. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344** 539–548.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics* **18** 191–219. MR1041391

Klaus, B. and Strimmer, K. (2015). fdrtool: Estimation of (Local) False Discovery Rates and Higher Criticism R package version 1.2.15.

Kleijn, B. J. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics* **34** 837–877.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, Second ed. Springer-Verlag, New York. MR1993642

Lai, T. L. (2003). Stochastic approximation. *Ann. Statist.* **31** 391–406. MR1983535

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M. and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19** 90–97.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

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

Liu, B. and Ghosh, S. K. (2020). On empirical estimation of mode based on weakly dependent samples. *Computational Statistics & Data Analysis* **152** 107046.

Martin, R. (2019). Empirical priors and posterior concentration rates for a monotone density. *Sankhya A* **81** 493–509. MR4043484

Martin, R. and Ghosh, J. K. (2008). Stochastic approximation and Newton's estimate of a mixing distribution. *Statistical Science* **23** 365–382.

Martin, R. and Han, Z. (2016). A semiparametric scale-mixture regression model and predictive recursion maximum likelihood. *Computational Statistics and Data Analysis* **94** 75–85.

Martin, R. and Tokdar, S. T. (2009). Asymptotic properties of predictive recursion: robustness and rate of convergence. *Electronic Journal of Statistics* **3** 1455–1472.

Martin, R. and Tokdar, S. T. (2011). Semiparametric inference in mixture models with predictive recursion marginal likelihood. *Biometrika* **98** 567–582.

Martin, R. and Tokdar, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics* **13** 427–439.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley-Interscience, New York. MR1789474

Mokkadem, A. and Pelletier, M. (2007). A companion for the Kiefer-Wolfowitz-Blum stochastic approximation algorithm. *The Annals of Statistics*

**35** 1749–1772. MR2351104

Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhya A* **64** 306–322.

Newton, M. A., Quintana, F. A. and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical Nonparametric and Semiparametric Bayesian Statistics* 45–61. Springer.

Newton, M. A. and Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86** 15–26.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422

Patilea, V. (2001). Convex models, MLE and misspecification. *The Annals of Statistics* **29** 94–123. MR1833960

Rao, B. P. (1969). Estimation of a unimodal density. *Sankhyā A* **31** 23–36.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 731–792.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* 233–257. Elsevier.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Annals of Mathematical Statistics* **29** 373–405. MR0098427

Salomond, J.-B. (2014). Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Stat.* **8** 1380–1404. MR3263126

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **4** 10–26. MR0184378

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association* **110** 459–471. MR3367240

Shyamalkumar, N. (1996). Cyclic $I_0$ projections and its applications in statistics Technical Report, Technical Report 96-24, Dept. Statistics, Purdue Univ., West Lafayette, IN.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London. MR848134

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. MR1054861

Tansey, W., Oluwasanmi, K., Poldrack, R. A. and Scott, J. G. (2018). False discovery rate smoothing. *Journal of the American Statistical Association* **113** 1156–1171.

Teel, C., Park, T. and Sampson, A. R. (2015). EM estimation for finite mixture models with known mixture component size. *Communications in Statistics-Simulation and Computation* **44** 1545–1556.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics* **32** 244–248.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathe-*

*matical Statistics* **34** 1265–1269.

Tokdar, S. T., Martin, R. and Ghosh, J. K. (2009). Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics* **37** 2502–2522. MR2543700

Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* **10** 1–50.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **76** 751–761.

Williamson, R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Mathematical Journal* **23** 189–207. MR0077581

Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when $f$ is non-increasing. *Statistica Sinica* **3** 501–515.

Woody, S., Padilla, O. H. M. and Scott, J. G. (2022). Optimal post-selection inference for sparse signals: a nonparametric empirical Bayes approach. *Biometrika* **109** 1–16. MR4374637

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics* **2** 298–331. MR2399197