Convergence rate for predictive recursion estimation of finite mixtures

Ryan Martin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
rgmartin@math.uic.edu

October 14, 2011

Abstract

Predictive recursion (PR) is a fast stochastic algorithm for nonparametric estimation of mixing distributions in mixture models. It is known that the PR estimates of both the mixing and mixture densities are consistent under fairly mild conditions, but currently very little is known about the rate of convergence. Here I first investigate asymptotic convergence properties of the PR estimate under model misspecification in the special case of finite mixtures with known support. Tools from stochastic approximation theory are used to prove that the PR estimates converge, to the best Kullback–Leibler approximation, at a nearly root-n rate. When the support is unknown, PR can be used to construct an objective function which, when optimized, yields an estimate the support. I apply the known-support results to derive a rate of convergence for this modified PR estimate in the unknown support case, which compares favorably to known optimal rates.

Keywords and phrases: Density estimation; Kullback–Leibler divergence; Lyapunov function; mixture model; stochastic approximation.

1 Introduction

Nonparametric estimation of mixing distributions is an important and challenging problem in statistics. Recent progress along these lines has been made with the fast stochastic *predictive recursion* (PR) algorithm due to Newton et al. (1998) and Newton (2002). PR is fundamentally different from existing algorithms, such as EM, in a number of ways. Most importantly, PR is not a hill-climbing algorithm. Instead, it learns sequentially like stochastic approximation (Kushner and Yin 2003; Robbins and Monro 1951). In addition, PR is able to estimate a mixing density with respect to any user-defined dominating measure. That is, unlike the nonparmetric maximum likelihood estimate, which is almost surely discrete (Lindsay 1995), the PR estimate can be discrete, continuous, or both, depending on the user's choice of dominating measure.

Theoretically, it has been shown that the PR estimates of both the mixing and mixture densities are consistent under certain conditions; see Section 2 for more details. The goal of this note is to investigate the rate of convergence, about which very little is known. For this, we shall explore further the connection between PR and stochastic approximation developed in Martin and Ghosh (2008). To the author's knowledge, results on the rate of convergence for general stochastic approximations are only fully developed in the finite-dimensional context. Therefore, we shall confine ourselves here to an analysis of PR when the possibly misspecified model assumes that the data-generating distribution is a finite mixture with known support. In this case, we prove that the PR estimate of the mixing distribution converges almost surely at a nearly parametric root-n rate, where the limit is characterized by the mixture model closest to the true data-generating distribution based on the Kullback–Leibler divergence. This result also sheds light on how one should choose PR's tuning parameter in practical applications.

The PR algorithm itself is not naturally suited for the case when the support of the finite mixture model is unknown. But, by applying the general principle in Martin and Tokdar (2011b), I show that PR yields a sort of objective function which can be optimized to estimate the unknown support. I apply the paper's known-support results to establish rates of convergence for this new PR-based unknown-support procedure. Two numerical examples are given to illustrate the method; for more examples and the full computational details, the reader is referred to Martin (2011).

2 Predictive recursion

Suppose independent data $Y_1, \ldots, Y_n$ are available from a distribution with unknown density $m(y)$, which we model as a nonparametric mixture:

$$m_f(y) = \int_{\mathscr{U}} p(y \mid u) f(u) \, d\mu(u), \quad y \in \mathscr{Y}, \tag{1}$$

where $(y, u) \mapsto p(y \mid u)$ is a known kernel on $\mathscr{Y} \times \mathscr{U}$ and $f \in \mathbb{F}$ is unknown and to be estimated. Here $\mathbb{F} = \mathbb{F}(\mathscr{U}, \mu)$ is the set of all densities with respect to a given σ-finite Borel measure μ on $\mathscr{U}$. Newton (2002) presents the following algorithm for nonparametric estimation of f and m_f based on $Y_1, \ldots, Y_n$.

PR algorithm. Choose a density $f_0 \in \mathbb{F}$ and a sequence of weights $\{w_i : i \geq 1\} \subset (0, 1)$. Then, for $i = 1, \ldots, n$, compute $m_{i-1}(y) = m_{f_{i-1}}(y)$ and

$$f_i(u) = (1 - w_i) f_{i-1}(u) + w_i p(Y_i \mid u) f_{i-1}(u) \, / \, m_{i-1}(Y_i). \tag{2}$$

Return $f_n(u)$ and $m_n(y) = m_{f_n}(y)$ as estimates of $f(u)$ and $m_f(y)$, respectively.

PR has some interesting connections to the nonparametric Bayes estimate in the case where the unknown mixing distribution is modeled as a random draw from the Dirichlet process distribution. Martin and Tokdar (2011b) take advantage of this connection to motivate a PR-based semiparametric mixture model analysis where an additional unknown structural parameter is estimated by maximizing a PR-induced approximate marginal likelihood. Martin and Tokdar (2011a) use this general strategy to develop a PR-based methodology for large-scale nonparametric empirical Bayes multiple testing. In Section 4 I apply this method to mixtures with unknown support.

Asymptotic convergence properties of the PR estimates f_n and m_n have only recently become available. Let $\mathbb{M}$ denote the set of mixture densities m_f as f ranges over $\mathbb{F}$. Tokdar et al. (2009) build on the work of Ghosh and Tokdar (2006) to show that when the mixture model is correctly specified (i.e., $m \in \mathbb{M}$), then both f_n and m_n converge almost surely to f and m_f in their respective topologies. Martin and Tokdar (2009) go one step further, showing that if $m \notin \mathbb{M}$, then m_n converges to the closest mixture density $m_{f^\star} \in \mathbb{M}$ as measured by the Kullback–Leibler divergence. As a corollary, if f is identifiable in the postulated mixture model, then f_n converges almost surely to $f^\star$ in the weak topology. They also establish a bound on the rate of convergence for m_n in terms of the PR weight sequence $\{w_n\}$. For weights of the form $w_i = (i+1)^{-\gamma}$, for suitable γ, Martin and Tokdar (2009) obtain a $n^{-1/6}$ bound on the Hellinger convergence rate of m_n to $m_{f^\star}$ for a wide class of kernels $p(y \mid u)$. While this rate is comparable to the rate obtained in Genovese and Wasserman (2000), it leaves a lot to be desired. In fact, simulations in Martin and Tokdar (2009) suggest that the upper bound corresponds to a "worst case scenario" rate of convergence, i.e., when $f^\star$ sits on the boundary of $\mathbb{F}$. I expect that a nearly parametric root-n rate for m_n, like that obtained by Ghosal and van der Vaart (2001), can be achieved by PR, at least in some cases. In Section 3 we show that this conjecture holds in the special known finite support case.

3 Asymptotics for PR with known support

Assume that the true density m is modeled as a finite mixture. That is, $\mathscr{U}$ is a finite set of size s and μ is counting measure. In this case, $\mathbb{F}$ denotes the $(s-1)$-dimensional probability simplex, and I write $f = \{f(u) : u \in \mathscr{U}\}$. Then $m_f(y) = \sum_{u \in \mathscr{U}} p(y \mid u) f(u)$. Throughout, all s-dimensional vectors x will be indexed by $\mathscr{U}$, i.e., $x = \{x(u) : u \in \mathscr{U}\}$. Also, $\langle \cdot, \cdot \rangle$ denotes the usual inner-product and $\| \cdot \|$ the corresponding norm.

We begin by listing two basic assumptions about the mixture model.

Assumption 1. $u \mapsto p(y \mid u)$ is continuous for each $y \in \mathscr{Y}$.

Assumption 2. f is identifiable in model (1), i.e., $f \mapsto m_f$ is one-to-one.

For any density m' on $\mathscr{Y}$, define the Kullback–Leibler divergence of m' from m as $K(m, m') = \int \log\{m(y)/m'(y)\} m(y)\, dy$. Henceforth, I shall silently assume that $K(m, m') < \infty$ for all $m' \in \mathbb{M}$. Then the infimum

$$K^\star = \inf\{K(m, m_f) : f \in \mathbb{F}\},$$

is finite. It follows from Assumption 1 that there exists an $f^\star$ in the closure of $\mathbb{F}$ such that $K(m, m_{f^\star}) = K^\star$; see Lemma 3.1 of Martin and Tokdar (2009). Assumption 2 ensures that $f^\star$ is unique. Allowing the model to be misspecified is particularly important here, given that the assumption of known finite support is rather strong. For example, even if the support $\mathscr{U}$ is unknown, the results that follow show that PR does as well asymptotically as could be hoped for if we simply guess at what $\mathscr{U}$ should be.

Following Martin and Ghosh (2008), express the PR update $f_{n-1} \mapsto f_n$, $n \geq 1$, as follows:

$$f_n(u) = f_{n-1}(u) + w_n \Phi(Y_n, f_{n-1})(u), \quad u \in \mathscr{U}, \tag{3}$$

where, for generic $y \in \mathscr{Y}$ and $f \in \mathbb{F}$, the mapping $\Phi(y,f)$ is defined as

$$\Phi(y,f)(u) = f(u)\Big\{\frac{p(y \mid u)}{m_f(y)} - 1\Big\}.$$

Equation (3) shows that PR is a special case of a general Robbins–Monro type of stochastic approximation algorithm designed to find roots of the mapping

$$\varphi(f)(u) = f(u)\Big\{\int \frac{p(y \mid u)}{m_f(y)} m(y)\,dy - 1\Big\}, \quad f \in \mathbb{F}, \quad u \in \mathscr{U}. \tag{4}$$

This $\varphi(f)$ is nothing but the conditional expectation of $\Phi(Y_n, f_{n-1})$, under the true density m, given f_{n-1} equals f. The following result is an immediate consequence of the definitions and construction above.

Lemma 1. *The sequence $Z_n(u)$, for $u \in \mathscr{U}$, given by*

$$Z_n(u) = \Phi(Y_n, f_{n-1})(u) - \varphi(f_{n-1})(u), \tag{5}$$

is a martingale difference sequence with respect to the σ-algebra $\mathscr{A}_n$ generated by $Y_1, \ldots, Y_n$. Moreover, $\|Z_n\|^2$ is bounded for all $n \geq 1$.

According to stochastic approximation theory (e.g., Kushner and Yin 2003), convergence properties of f_n, as $n \to \infty$, can be found by investigating the asymptotic behavior of solutions of an appropriate ordinary differential equation (ODE). Specifically, let $\{f^t : t \geq 0\}$ denote a generic trajectory in $\mathbb{F}$. Then the limiting behavior of solutions f^t of the ODE $df^t/dt = \varphi(f^t)$, as $t \to \infty$, can be used to study the limiting behavior of the PR sequence f_n, as $n \to \infty$. For this purpose, I will need some basic definitions and results from the theory of ODEs.

Lemma 2. *The mixing distribution $f^\star$ is an equilibrium point of the ODE $df^t/dt = \varphi(f^t)$; in other words, $\varphi(f^\star)(u) = 0$ for all u.*

Proof. Plugging $f^\star$ into the expression in (4) gives

$$\varphi(f^\star)(u) = f^\star(u)\Big\{\int \frac{p(y \mid u)}{m_{f^\star}(y)} m(y)\,dy - 1\Big\}.$$

By the fact that $f^\star$ minimizes $K(m, m_f)$, it follows from Lemma 3.3 of Martin and Tokdar (2009) that $\varphi(f^\star)(u) \leq 0$ for each u. But since $\sum_u \varphi(f^\star)(u)$ vanishes, it must be that $\varphi(f^\star)(u) = 0$ for each u, proving the claim. □

The goal is to show that $f^\star$ is a stable equilibrium in the sense that any solution to the ODE converges to $f^\star$, regardless of the initial condition. For this, a *Lyapunov function* will be useful.

Definition 1. A function $\ell : \mathbb{F} \to \mathbb{R}$ is a Lyapunov function at $f^\star$ for the ODE $df^t/dt = \varphi(f^t)$ if (i) $\ell(f)$ is continuously differentiable in a neighborhood of $f^\star$, (ii) $\ell(f) \geq 0$ with equality if and only if $f = f^\star$, and (iii) $\dot{\ell}(f) = \langle \nabla \ell(f), \varphi(f) \rangle \leq 0$.

Lyapunov's theory, described beautifully in LaSalle and Lefschetz (1961), states that if a Lyapunov function $\ell(f)$ exists at $f = f^\star$, then $f^\star$ is a stable equilibrium point. Next I show that a slight variation of the Kullback–Leibler divergence is a Lyapunov function in the present context.

Lemma 3. *The mapping $\ell : \mathbb{F} \to [0, \infty)$ given by*

$$\ell(f) = K(m, m_f) - K^\star + \textstyle\sum_u f(u) - 1 \tag{6}$$

is a Lyapunov function for the ODE $df^t/dt = \varphi(f^t)$.

Proof. Properties (i) and (ii) in Definition 1 are obvious. For property (iii), simple calculus reveals that $\varphi(f)(u) = -f(u)\{\nabla\ell(f)\}(u)$, from which it follows that $\dot{\ell}(f) = -\sum_u f(u)\{\nabla\ell(f)\}(u)^2 \leq 0$. That equality is obtained if and only if $f = f^\star$ follows from the fact that $f^\star$ is the unique minimizer of $K(m, m_f)$ and, hence, the only point at which $\nabla\ell(f)$ vanishes. □

The function $\ell(f)$ in (6) can be viewed as a Lagrange multiplier version of the Kullback–Leibler divergence with the trivial constraint $\sum_u f(u) = 1$. This is consistent with the interpretation of PR as an algorithm that asymptotically minimizes $K(m, m_f)$ over $\mathbb{F}$ (Martin and Tokdar 2009). Another important observation, used in Lemma 5 below, is that $\ell(f)$ is convex.

Next I state an extension of the PR convergence theorem in Martin and Ghosh (2008) for the case where the true data-generating density m need not belong to the class $\mathbb{M}$ of mixture models (1). For this we need

Assumption 3. $\sum_n w_n = \infty$ and $\sum_n w_n^{1+\varepsilon} < \infty$ for some $\varepsilon \in (0, 1]$.

In practice, it is common to take $w_n = (n+1)^{-\gamma}$ for $\gamma \in (1/2, 1]$. Then Assumption 3 holds with $\varepsilon > \gamma^{-1} - 1$.

Theorem 1. *Under Assumptions 1–3, $f_n \to f^\star$ almost surely, where $f^\star$ is the unique minimizer of $K(m, m_f)$ over $\mathbb{F}$.*

Proof. In light of Lemmas 1–3, the claim follows from Theorem 5.2.3 of Kushner and Yin (2003) and the continuity of $\varphi(f)$; see Martin and Ghosh (2008). □

The main result on a rate of convergence for PR will make use of a general theorem on convergence rates of stochastic approximation (Chen 2002, Theorem 3.1.1); see Appendix A. But two preliminary result are needed first.

Lemma 4. *The sequence Z_n in (5) satisfies $\sum_{n=1}^\infty w_n^{1-\delta} Z_n < \infty$ almost surely for $\delta \in (0, (1-\varepsilon)/2]$, where ε is as in Assumption 3.*

Proof. Let $X_N = \sum_{n=1}^N w_n^{1-\delta} Z_n$. By Lemma 1, $\{X_N : N \geq 1\}$ is a martingale sequence and, since $\{Z_n\}$ is bounded,

$$\mathsf{E}\|X_N\|^2 = \sum_{n=1}^N w_n^{2(1-\delta)} \mathsf{E}\|Z_n\|^2 \leq \text{const} \cdot \sum_{n=1}^\infty w_n^{2(1-\delta)}.$$

Taking $\delta \leq (1-\varepsilon)/2$, it follows from Assumption 3 that $\mathsf{E}\|X_N\|^2$ is uniformly bounded in N. Then the martingale convergence theorem (Breiman 1992, Theorem 5.14) implies that X_N converges almost surely, completing the proof. □

An additional assumption about the weights is required. For weights given by $w_n = (n+1)^{-\gamma}$, this assumption holds as long as $\gamma < 1$.

Assumption 4. $\{w_n\}$ satisfies $w_{n+1}^{-1} - w_n^{-1} \to 0$.

Lemma 5. *Let $J = D\varphi(f^\star)$ denote the derivative of φ evaluated at $f = f^\star$. If $f^\star$ is in the interior of $\mathbb{F}$, then all eigenvalues of J are negative.*

Proof. Simple calculus reveals that $J = D\varphi(f^\star)$ is of the form

$$J(u,v) = -f^\star(u) \int \frac{p(y \mid u) p(y \mid v)}{m_{f^\star}(y)^2} m(y)\, dy, \quad u, v \in \mathscr{U}.$$

In matrix notation, write $J = -\text{diag}(f^\star) \cdot \nabla^2 \ell(f^\star)$, where $\text{diag}(f^\star)$ is a diagonal matrix with the elements of $f^\star$ as its diagonal entries, and $\nabla^2 \ell(f^\star)$ is the second derivative matrix of $\ell(f)$ evaluated at $f = f^\star$. Since $f^\star$ is in the interior of $\mathbb{F}$, all entries are positive and, hence, $\text{diag}(f^\star)$ is positive definite. Since $\ell(f)$ is convex on $\mathbb{F}$, $\nabla^2 \ell(f^\star)$ is also positive definite. The claim follows from the fact that the product of these two positive definite matrices, which is $-J$, must have positive eigenvalues. □

An interesting observation is that the matrix $P = -J^\top$, the negative transpose of the Jacobian J in Lemma 5, is a transition probability matrix for an irreducible, aperiodic Markov chain on $\mathscr{U}$. This chain is also reversible and has $f^\star$ as its stationary distribution. But how this observation might be useful in studying the asymptotic convergence of PR remains unclear.

In light of Assumptions 1–4, Lemmas 4 and 5, and the existence of a Lyapunov function proved in Lemma 3, the main result on the convergence rate of PR is a consequence of Chen's theorem in Appendix A.

Theorem 2. *Assume that $f^\star$ lies in the interior of $\mathbb{F}$. Then under Assumptions 1–4, $\|f_n - f^\star\| = o(w_n^\delta)$ almost surely for δ in Lemma 4.*

When the weights are given by $w_n = (n+1)^{-\gamma}$, for $\gamma \in (1/2, 1)$, it follows from Theorem 2 and the previous discussion that $\|f_n - f^\star\| = o(n^{-(1-1/2\gamma)})$ almost surely. Since γ can be chosen arbitrarily close to 1, it follows that the convergence rate can be made arbitrarily close to $n^{-1/2}$ almost surely.

A slightly stronger version of Theorem 2 could be obtained if weight sequences were allowed to satisfy $w_{n+1}^{-1} - w_n^{-1} \to \alpha$, with $\alpha > 0$. For example, if $w_n = (n+1)^{-1}$, then $\alpha = 1$. This extension would make the root-n rate possible, but it would require all eigenvalues of J in Lemma 5 to be less than $-1/2$. At this point it is unclear whether this claim is true; standard bounds for eigenvalues, such as those in Gershgorin's theorem or Proposition 2 in Diaconis and Stroock (1991), are not helpful in this case.

Almost sure rates of convergence for the mixture density m_n to $m_{f^\star}$ are available as consequences of Theorem 2. The L_1 rate follows immediately from its definition, while the rate for the Kullback–Leibler contrast, $K(m, m_n) - K^\star$, requires a simple second-order Taylor approximation of $\ell(f)$ at $f = f^\star$.

Corollary 1. *Under the conditions on Theorem 2, $\int |m_n - m_{f^\star}|\, dy = o(w_n^\delta)$ almost surely for δ in Lemma 4. Likewise, $K(m, m_n) - K^\star = o(w_n^{2\delta})$.*

Martin and Tokdar (2009) derive a bound of $o(W_n^{-1})$ for $K(m, m_n) - K^\star$ in the general compact $\mathscr{U}$ case, where $W_n = \sum_{i=1}^n w_i$. When $w_n = (n+1)^{-\gamma}$, the bound for $K(m, m_n) - K^\star$ in Martin and Tokdar (2009) becomes $o(n^{-(1-\gamma)})$, which can be no faster than $n^{-1/3}$ under their conditions. Compare this to the rate of $o(n^{-(2-1/\gamma)})$ obtained from Corollary 1, which is considerably faster than $n^{-1/3}$ for $\gamma \approx 1$, albeit for the special known finite support case. So, regarding the PR weights $\{w_i : i \geq 1\}$, the message here, contrary to that in Martin and Tokdar (2009), is that the faster the weights vanish the faster the overall convergence.

4 PR with unknown support

The PR convergence theory in the previous section assumes the finite support is known and only the mixing distribution is unknown. In practice, however, both the support and mixing distribution are unknown and to be estimated. To close this gap, I propose here a new PR-based approach for handling the unknown support case. The asymptotic results in Section 3 will be used to prove consistency of this new procedure. Two simple examples are also given for illustration, but the computational details, simulations, and extensions will be presented elsewhere (Martin 2011).

4.1 Setup

Let $\overline{\mathscr{U}}$ be a compact set, large enough that there is a finite mixture supported in $\overline{\mathscr{U}}$ that gives a sufficiently accurate approximation to m. Take U to be a generic finite subset of $\overline{\mathscr{U}}$. By treating U as the fixed support, a run of PR will produce a sequence of estimates $\{(f_{i,U}, m_{i,U}) : i = 1, \ldots, n\}$ of the mixing and mixture distributions, whose dependence on the chosen support set U are now made explicit. In the same vein, write $\mathbb{F}_U$ for the $(|U|-1)$-dimensional probability simplex and define $K^\star(U) = \inf\{K(m, m_{f,U}) : f \in \mathbb{F}_U\}$, the smallest Kullback–Leibler number for mixtures supported on U.

The jumping off point is that the result $K(m, m_{n,U}) - K^\star(U) = o(w_n^{2\delta})$ of Corollary 1 holds "pointwise" for all U; that is, the particular support U plays no role in the analysis of Section 3. Thus, in the present case where the support is unknown, a reasonable strategy is to estimate the support by minimizing, over U, some estimate of $K(m, m_{n,U})$. This is the approach advocated by Martin and Tokdar (2011b). Indeed, by making connections to PR and Dirichlet process mixture models, they argue that, in the present context, the appropriate estimate of $K(m, m_{n,U})$ is

$$K_n(U) = \sum_{i=1}^n \log \frac{m(Y_i)}{m_{i-1,U}(Y_i)}, \quad U \subset \overline{\mathscr{U}}, \quad |U| < \infty. \tag{7}$$

Then the goal is to minimize $K_n(U)$ over U. But since it is not possible to perform this optimization over all finite $U \subset \overline{\mathscr{U}}$, some adjustment must be made. Consider starting with a fixed finite subset $\mathscr{U}$ of $\overline{\mathscr{U}}$ obtained by chopping up $\overline{\mathscr{U}}$ into a sufficiently fine grid, so that $|\mathscr{U}|$ is large. Then the collection of all subsets U of $\mathscr{U}$ is huge—it has $2^{|\mathscr{U}|} - 1$ elements—but finite so it is possible to minimize $K_n(U)$ over $U \subseteq \mathscr{U}$. Martin (2011) uses a simulated annealing strategy to perform this optimization. Once the minimizer $\hat{U}_n$ of

$K_n(U)$ is obtained, PR is run once more to produce $f_{n,\hat{U}_n}$ and $m_{n,\hat{U}_n}$ as estimates of the mixing and mixture distributions, respectively.

4.2 Large-sample theory

For simplicity, I will assume that the true density m is indeed a mixture density of the postulated form with support contained in $\mathscr{U}$; the more general case can be handled similarly, but with an additional technical assumption (Martin and Tokdar 2011b, Assumption 6). Also, assume that $w_n = (n+1)^{-\gamma}$ for some $\gamma \in (0.5, 1)$. To get convergence of the approximation $K_n(U)$ to $K^\star(U)$, I will need one additional assumption, stated next, which holds for many common kernels, including normal and Poisson.

Assumption 5. There exists a finite constant $A > 0$ such that

$$\max_{u_1,u_2,u_3 \in \mathscr{U}} \int \Bigl\{ \frac{p(y \mid u_1)}{p(y \mid u_2)} \Bigr\}^2 p(y \mid u_3)\, dy \leq A.$$

Under Assumptions 1–5, one can follow the proof of Theorem 2 in Martin and Tokdar (2011b) to conclude that, for each fixed $U \subseteq \mathscr{U}$,

$$\lim_{n\to\infty} \Bigl| c_n \bigl\{ K_n(U) - K^\star(U) \bigr\} - \frac{c_n}{n} \sum_{i=1}^n \bigl\{ K(m, m_{i-1,U}) - K^\star(U) \bigr\} \Bigr| = 0, \tag{8}$$

almost surely, for any sequence c_n that satisfies $c_n = O(n^{1/2-\varepsilon})$ for some $\varepsilon > 0$. It follows from Corollary 1 that the summation in (8) is of the order $n^{1/\gamma-1}$. So, if $\varepsilon > \max\{0, \gamma^{-1} - 3/2\}$, the right-most term in the modulus in (8) vanishes and, therefore, so must the left-most term. This proves that, for $\gamma \approx 1$, $K_n(U) \to K^\star(U)$ pointwise in U at a rate just slower than $n^{-1/2}$. But since $2^{\mathscr{U}}$ is finite, the convergence is also uniform. The following theorem summarizes this result.

Theorem 3. *Choose weights $w_n = (n+1)^{-\gamma}$ with $\gamma \in (0.5, 1)$ and let $\varepsilon > \max\{0, \gamma^{-1} - 3/2\}$. Then, under Assumptions 1–5, $n^{1/2-\varepsilon}\{K_n(U) - K^\star(U)\} \to 0$ almost surely as $n \to \infty$. Moreover, since U ranges only over a finite set, $n^{1/2-\varepsilon} K_n(\hat{U}_n) \to 0 = K^\star(U^\star)$, where $U^\star \subseteq \mathscr{U}$ is the support of the true mixture distribution.*

If I define a distance d between two sets as the cardinality of their symmetric difference, then Theorem 3 states that $d(\hat{U}_n, U^\star) = o(n^{-1/2+\varepsilon})$. In other words, $\hat{U}_n$ is a nearly root-n d-consistent estimate of $U^\star$. Furthermore, a nearly root-n rate of convergence for $f_{n,\hat{U}_n}$ can be obtained, which I now sketch. With a slight abuse of notation, I can bound the total variation distance between $f_{n,\hat{U}_n}$ and $f^\star$ as follows:

$$\begin{aligned} d_{\text{TV}}(f_{n,\hat{U}_n}, f^\star) &= \sum_{u \in \mathscr{U}} |f_{n,\hat{U}_n}(u) - f^\star(u)| \\ &= \sum_{u \in \hat{U}_n \cap U^{\star c}} f_{n,\hat{U}_n}(u) + \sum_{u \in \hat{U}_n^c \cap U^\star} f^\star(u) + \sum_{u \in \hat{U}_n \cap U^\star} |f_{n,\hat{U}_n}(u) - f^\star(u)| \\ &\leq d(\hat{U}_n, U^\star) + d_{\text{TV}}(f_{n,\hat{U}_n}, f_{n,U^\star}) + d_{\text{TV}}(f_{n,U^\star}, f^\star). \end{aligned}$$

The two outer-most terms on the right-hand side vanish at a nearly root-n rate according to Theorems 3 and 2, respectively. The middle term is more difficult to analyze, but it is clear that the data-dependent PR mapping $U \mapsto f_{n,U}$ is, in some sense, continuous in U. So, the convergence of $d_{\text{TV}}(f_{n,\hat{U}_n}, f_{n,U^\star})$ is also driven by $d(\hat{U}_n, U^\star)$. Therefore, the rate for $d_{\text{TV}}(f_{n,\hat{U}_n}, f^\star)$ must also be nearly $n^{-1/2}$.

Recall that Chen (1995) showed that, for finite mixtures, the optimal rate of convergence is $n^{-1/4}$. In that case, the unknown finite support is allowed to be anything, essentially nonparametric, so the rates are relatively slow. In contrast, by restricting the set of candidate supports to subsets of a large but ultimately finite set $\mathscr{U}$, I am able to achieve a nearly parametric root-n rate of convergence.

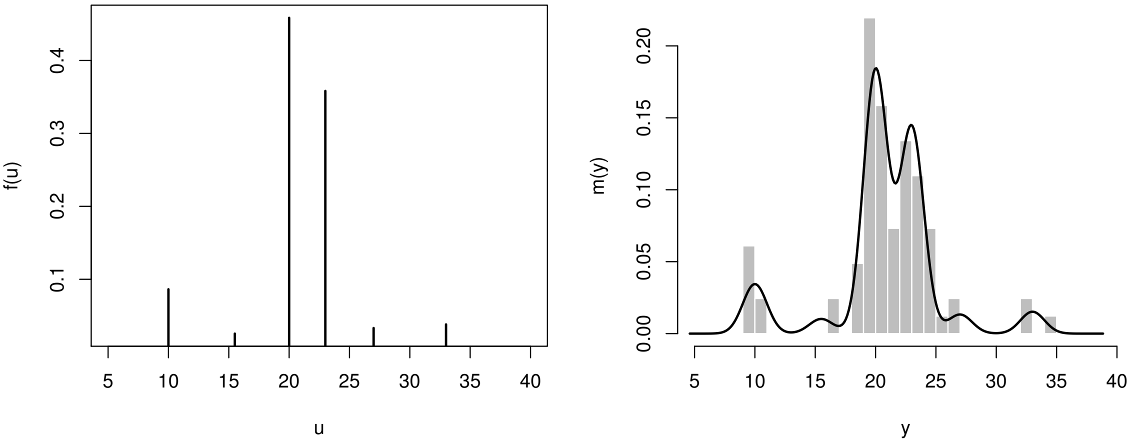
4.3 Examples

Here I give two relatively simple real-data examples—a Gaussian location mixture and a Poisson mixture—to illustrate the potential of the proposed method.

Example 1. Under the Big Bang model, galaxies should form clusters and the relative velocities of the galaxies should be similar within clusters. Roeder (1990) considers velocity data for $n = 82$ galaxies. She models this data as a finite Gaussian mixture, with the number and location of the mixture components unknown. The assumption is that each galactic cluster is a single component of the Gaussian mixture. The presence of multiple mixture components is consistent with the hypothesis of galaxy clustering.

We apply the methodology outlined above to estimate the mixing distribution f. We will consider a simple Gaussian mixture model in which each component has variance $\sigma^2 = 1$, based on the *a priori* considerations of Escobar and West (1995). From the observed velocities, it is apparent that the mixture components should be centered somewhere in the interval $\overline{\mathscr{U}} = [5, 40]$, so we choose a grid of candidate support points $\mathscr{U} = \{5.0, 5.5, 6.0, \ldots, 39.5, 40.0\}$. Figure 1 shows the corresponding estimates of the mixing and mixture distribution. The PR method identifies six galaxy clusters, and the estimates of U and f closely match those of Ishwaran et al. (2001) and others.

Example 2. Karlis and Xekalaki (2001, Table 1) present data on the number of defaulted installments in a Spanish financial institution. This data has a high number of zero counts, as well as substantial overdispersion. This suggests a Poisson mixture model, and here we compare the PR-based estimates to others presented in the literature. The first three rows of Table 1 show the estimates of (f, U) for three methods in an zero-inflated Poisson mixture model. These include an estimate based on the AIC penalty, the SCAD-based penalized likelihood approach of Chen and Khalili (2008), and a minimum Hellinger distance method for count data (Woo and Sriram 2007). I start by bounding the support by $\overline{\mathscr{U}} = [0, 30]$ and taking $\mathscr{U}$ to be a set of 100 equispaced points in $\overline{\mathscr{U}}$. All but the Woo–Sriram estimates have five support points, including zero. Besides this, we find that the corresponding estimates are quite similar. An attractive feature of this method is that no special adjustments are needed for zero-inflation. That is, zero-inflation can be achieved by simply including zero in the grid $\mathscr{U}$ and letting the data decide if a mass at zero is appropriate. Fitted values were obtained for each of the four methods (not shown) and I find that, for small y-values, where the observed



(a) Mixing distribution

(b) Mixture distribution

Figure 1: Plots of the PR estimates for the galactic velocity data in Example 1.

Estimates	$(u_1, f(u_1))$	$(u_2, f(u_2))$	$(u_3, f(u_3))$	$(u_4, f(u_4))$	$(u_5, f(u_5))$
AIC–BIC	(0, .314)	(.298, .435)	(4.37, .200)	(10.99, .048)	(26.51, .002)
MSCAD	(0, .328)	(.302, .417)	(4.19, .193)	(9.78, .055)	(20.01, .007)
WS	(0, .373)	(.36, .385)	(4.52, .199)	(11.26, .043)	
SASA	(0, .328)	(.303, .418)	(4.24, .201)	(10.91, .051)	(27.27, .002)

Table 1: Estimates of (f, U) for the financial data Poisson mixture in Example 2. The first three rows are taken from Chen and Khalili (2008, Table 10).

counts are relatively large, the PR-based estimate appears to provide a better overall fit compared to the others.

Acknowledgments

The author thanks Professor Surya Tokdar for a number of helpful suggestions, and the Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis, for their hospitality when a portion of this work was completed.

A Convergence rates for stochastic approximation

Consider a stochastic approximation process $\{X_n : n \geq 0\}$ which, for fixed initial value $X_0 = x_0$, is defined recursively as follows:

$$X_n = X_{n-1} + a_n \varphi(X_{n-1}) + a_n Z_n, \quad n \geq 1.$$

The process is designed so that $X_n \to x^\star$ almost surely, where $x^\star$ satisfies $\varphi(x^\star) = 0$. We shall assume that $\{X_n\}$ bounded; otherwise, some truncation or projection techniques are needed (Chen 2002; Kushner and Yin 2003). The PR estimates f_n are constrained to the simplex, so they satisfy this boundedness condition trivially. Next are the main assumptions of the theorem.

A1. The weights $\{a_n\}$ satisfy $a_n > 0$, $a_n \to 0$, $\sum_n a_n = \infty$, and $a_{n+1}^{-1} - a_n^{-1} \to \alpha$ for some $\alpha \geq 0$.

A2. There exists a Lyapunov function $\ell(x)$ at the equilibrium point $x^\star$ of the ODE $dx_t/dt = \varphi(x_t)$.

A3. $\sum_n a_n^{1-\delta} Z_n < \infty$ almost surely for some $\delta \in (0, 1/2)$.

A4. $\varphi(x)$ is continuously differentiable, and all eigenvalues of $J + \alpha\delta I$ have negative real parts, where $J = D\varphi(x^\star)$.

Chen's Theorem. *Under A1–A4,* $\|X_n - x^\star\| = o(a_n^\delta)$ *almost surely.*

References

Breiman, L. (1992), *Probability*, vol. 7 of *Classics in Applied Mathematics*, Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Chen, H.-F. (2002), *Stochastic approximation and its applications*, vol. 64 of *Nonconvex Optimization and its Applications*, Dordrecht: Kluwer Academic Publishers.

Chen, J. and Khalili, A. (2008), "Order selection in finite mixture models with a nonsmooth penalty," *J. Amer. Statist. Assoc.*, 103, 1674–1683.

Chen, J. H. (1995), "Optimal rate of convergence for finite mixture models," *Ann. Statist.*, 23, 221–233.

Diaconis, P. and Stroock, D. (1991), "Geometric bounds for eigenvalues of Markov chains," *Ann. Appl. Probab.*, 1, 36–61.

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

Genovese, C. R. and Wasserman, L. (2000), "Rates of convergence for the Gaussian mixture sieve," *Ann. Statist.*, 28, 1105–1127.

Ghosal, S. and van der Vaart, A. W. (2001), "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities," *Ann. Statist.*, 29, 1233–1263.

Ghosh, J. K. and Tokdar, S. T. (2006), "Convergence and consistency of Newton's algorithm for estimating mixing distribution," in *Frontiers in Statistics*, eds. Fan, J. and Koul, H., London: Imp. Coll. Press, pp. 429–443.

Ishwaran, H., James, L. F., and Sun, J. (2001), "Bayesian model selection in finite mixtures by marginal density decompositions," *J. Amer. Statist. Assoc.*, 96, 1316–1332.

Karlis, D. and Xekalaki, E. (2001), "Robust inference for finite Poisson mixtures," *J. Statist. Plann. Inference*, 93, 93–115.

Kushner, H. J. and Yin, G. G. (2003), *Stochastic approximation and recursive algorithms and applications*, New York: Springer-Verlag, 2nd ed.

LaSalle, J. and Lefschetz, S. (1961), *Stability by Liapunov's Direct Method with Applications*, New York: Academic Press.

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

Martin, R. (2011), "A hybrid stochastic approximation–simulated annealing approach for estimating finite mixtures," *Submitted manuscript.* Preprint at `arXiv:1106.4432`.

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

Martin, R. and Tokdar, S. T. (2009), "Asymptotic properties of predictive recursion: robustness and rate of convergence," *Electron. J. Stat.*, 3, 1455–1472.

— (2011a), "A nonparametric empirical Bayes framework for large-scale multiple testing," *Biostatistics*, to appear. Preprint at `arXiv:1106.3885`.

— (2011b), "Semiparametric inference in mixture models with predictive recursion marginal likelihood," *Biometrika*, 98, 567–582.

Newton, M. A. (2002), "On a nonparametric recursive estimator of the mixing distribution," *Sankhyā Ser. 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*, eds. Dey, D., Müller, P., and Sinha, D., New York: Springer, vol. 133 of *Lecture Notes in Statist.*, pp. 45–61.

Robbins, H. and Monro, S. (1951), "A stochastic approximation method," *Ann. Math. Statistics*, 22, 400–407.

Roeder, K. (1990), "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *J. Amer. Statist. Assoc.*, 617–624.

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009), "Consistency of a recursive estimate of mixing distributions," *Ann. Statist.*, 37, 2502–2522.

Woo, M.-J. and Sriram, T. N. (2007), "Robust estimation of mixture complexity for count data," *Comput. Statist. Data Anal.*, 51, 4379–4392.