Asymptotic properties of Predictive Recursion: robustness and rate of convergence

by

Ryan Martin and Surya T. Tokdar

PR # 09-04

This manuscript and other can be obtained via the World Wide Web from www.math.iupui.edu

September 2, 2009
Asymptotic properties of Predictive Recursion: robustness and rate of convergence

Ryan Martin  
Indiana University-Purdue University Indianapolis

Surya T. Tokdar  
Duke University

August 20, 2009

Abstract

Here we explore general asymptotic properties of Predictive Recursion (PR) for nonparametric estimation of mixing distributions. Under fairly mild conditions, we prove that, when the mixture model is mis-specified, the estimated mixture converges almost surely in total variation to the mixture that minimizes the Kullback-Leibler divergence, and a bound on the (Hellinger contrast) rate of convergence is obtained. Moreover, when the mixture model is identifiable, almost sure weak convergence of the mixing distribution estimate follows.

PR assumes that the support of the mixing distribution is known. To remove this requirement, we propose a generalization that incorporates a sequence of supports, increasing with the sample size, that combines the efficiency of PR with the flexibility of mixture sieves. Under mild conditions, we obtain a bound on the rate of convergence of these new estimates.

1 Introduction

Despite a well-developed theory and numerous applications of mixture models, estimation of a mixing distribution remains a challenging problem. However, some recent progress has been made through a computationally efficient non-parametric estimate due to Newton (2002); see also Newton et al. (1998) and Newton and Zhang (1999). A mixture model views data \((X_1, \cdots, X_n) \in \mathcal{X}^n\) as independent observations from a density \(m(x)\) of the form

\[
m_f(x) = \int_{\Theta} p(x|\theta)f(\theta)\,d\mu(\theta), \quad f \in \mathcal{F}
\]  

where \(\mathcal{F} = \mathcal{F}(\Theta, \mu)\) is the class of densities dominated by a \(\sigma\)-finite measure \(\mu\) on a parameter space \(\Theta\), and \(\mathcal{P} = \{p(\cdot|\theta) : \theta \in \Theta\}\) is a parametric family of densities.
on \( X \), dominated by a \( \sigma \)-finite measure \( \nu \). More succinctly, a mixture model assumes \( m \in \mathcal{M}_\Theta := \{ m_f : f \in \mathbb{F} \} \), the convex hull of \( \mathbb{P} \). To estimate \( f \) from the data \( X_1, \ldots, X_n \), Newton (2002) proposed the following \( n \)-step recursive algorithm, called Predictive Recursion (PR):

**Algorithm PR.** Choose an initial density \( f_0 \) on \( \Theta \) and a sequence of weights \( \{ w_i : i \geq 1 \} \subset (0, 1) \). For \( i = 1, \ldots, n \), compute

\[
f_i(\theta) = (1 - w_i)f_{i-1}(\theta) + w_i \frac{p(X_i|\theta)f_{i-1}(\theta)}{\int_\Theta p(X_i|\theta')f_{i-1}(\theta') d\mu(\theta')}, \quad \theta \in \Theta
\]

(1.2)

and produce \( f_n \) as the final estimate of \( f \).

Until recently, very little was known about the large-sample behavior of the estimates \( f_n \) and \( m_n \). Ghosh and Tokdar (2006) used a novel martingale argument to prove, when \( m = m_f \in \mathcal{M}_\Theta \) for finite \( \Theta \), that \( f_n \to f \) almost surely. Martin and Ghosh (2008) proved a slightly stronger consistency theorem via a stochastic approximation representation of \( f_n \). Most recently, Tokdar, Martin, and Ghosh (2009) (henceforth, TMG) handled the case of a more general parameter space \( \Theta \) by extending the martingale argument to the \( X \)-space, proving that the mixture density estimate

\[
m_n(x) := m_f_n(x) = \int p(x|\theta)f_n(\theta) d\mu(\theta), \quad x \in X,
\]

(1.3)

converges almost surely to \( m_f(x) \) in the \( L_1 \) topology. From this, consistency of \( f_n \) in the weak topology on \( \Theta \) is obtained.

In this paper, we further extend the martingale-based arguments of TMG, establishing two important new results. First, in the more general context, where the mixture model \( \mathcal{M}_\Theta \) need not contain the true density \( m \), we show that the estimated mixture \( m_n \) in (1.3) is asymptotically robust in the sense that it converges almost surely in the total variation (or Hellinger) topology to the mixture \( m_f \in \mathcal{M}_\Theta \) that minimizes the Kullback-Leibler (KL) divergence,

\[
K(m, m_\phi) = \int \log \frac{m}{m_\phi} m \, d\nu.
\]

When the mixing density is identifiable, we also obtain weak convergence of \( f_n \). Our second result is a bound on the rate of convergence for \( m_n \). For a unified treatment of the well- and mis-specified cases, we consider the Hellinger contrast

\[
\rho(m_n, m_f) = \left[ \frac{1}{2} \int \left( \sqrt{\frac{m_n}{m_f}} - 1 \right)^2 m \, d\nu \right]^{1/2},
\]

(1.4)

where \( m_f \) is the mixture that minimizes \( K(m, m_\phi) \). The Hellinger contrast has been previously used to study asymptotics of maximum likelihood and Bayes estimates under model mis-specification (Kleijn and van der Vaart 2006; Patilea 2001). In Section 3, we show that if \( \Theta \) is compact, then \( \rho(m_n, m_f) \to 0 \) almost surely at a rate faster than \( a_n^{-1/2} \), where \( a_n = \sum_{i=1}^n w_i \). This establishes a direct connection between the choice of \( w_i \) and the performance of the resulting PR
estimate. Moreover, this bound is derived without using any structural knowledge about the mixands $p(x|\theta)$ and applies to a wide range of such densities, including Normal, Gamma, and Poisson. The conditions on $w_i$ required by this result are satisfied by $w_i \approx i^{-\gamma}$ for $\gamma \in (2/3, 1]$, leading to $a_n \approx n^{1-\gamma}$. Consequently, the Hellinger contrast convergence rate of $m_n$ is faster than $n^{-(1-\gamma)/2}$.

How this relates to the rate of convergence for $f_n$ remains an open question.

Our nearly $n^{-1/6}$ bound on the convergence rate for $m_n$ matches with the results in Genovese and Wasserman (2000) derived for the special case of finite Gaussian mixtures. But it falls short of the nearly parametric rates obtained in Li and Barron (2000) and Ghosal and van der Vaart (2001). However, Li and Barron’s bound involves constants which can be infinite, and Ghosal and van der Vaart’s proof seems to rely heavily on the smoothness of the Gaussian mixands. It is quite possible that the PR estimates converge faster than what our bounds suggest, but at this moment we are unsure how sharp our bounds are.

2 Kullback-Leibler projection

It is quite natural to use the KL divergence to study the large-sample properties of PR. Indeed, Martin and Ghosh (2008) remark that, for $\Theta$ finite, $\ell(\varphi) = K(m, m_\varphi)$ is a Lyapunov function for the differential equation that characterizes the asymptotics of PR: roughly, the KL divergence controls the dynamics of PR, driving the estimates toward a stable equilibrium. But when $m \notin M_\Theta$, this equilibrium cannot be $m$. In such cases, we consider the mixture $m_f$ that is “closest” to $m$ in a KL sense; that is,

$$K(m, m_f) = K(m, M_\Theta) := \inf\{K(m, m_\varphi) : \varphi \in \mathcal{F}\}$$

We call $m_f$ the KL projection of $m$ onto $M_\Theta$; see also Leroux (1992), Shyamalkumar (1996), Patilea (2001), and Kleijn and van der Vaart (2006).

Existence of the KL projection is an important issue, and various results are available; see, for example, Liese and Vajda (1987, Chap. 8). Here we prove a simple result which gives sufficient conditions for the existence of a KL projection in our special case of mixtures. Assume the following:

A1. $\mathcal{F}$ is compact with respect to the weak topology.

A2. $\theta \mapsto p(x|\theta)$ is bounded and continuous for $\nu$-almost all $x$.

Lemma 2.1. Under A1–A2, there exists a mixing density $f \in \mathcal{F}$ such that $K(m, m_f) = K(m, M_\Theta)$.

Proof. Choose any $\varphi \in \mathcal{F}$ and any $\{\varphi_s\} \subset \mathcal{F}$ such that $\varphi_s \rightarrow \varphi$ weakly. Then A2 and Scheffé’s theorem imply $m_{\varphi_s} \rightarrow m_\varphi$ in the $L_1(\nu)$ and, hence, the weak topology. Further,

$$\kappa(\varphi) := K(m, m_\varphi) \leq \liminf_{s \rightarrow \infty} K(m, m_{\varphi_s}) = \liminf_{s \rightarrow \infty} \kappa(\varphi_s),$$
where the inequality follows from weak lower semi-continuity of \( K(m, \cdot) \); see Liese and Vajda (1987), Theorem 1.47. Consequently, \( \kappa(\cdot) \) is weakly lower semi-continuous and, therefore, must attain its infimum on the compact \( F \). A similar proof may be given based on Lemma 4 of Brown et al. (2008).

Convexity of \( \mathcal{M}_\Theta \) and of \( K(m, \cdot) \) together imply that the KL projection \( m_f \) in Lemma 2.1 is unique. However, in general, there could be many mixing densities \( f \in \mathcal{F} \) that generate the KL projection \( m_f \). Identifiability is needed to guarantee uniqueness of the mixing density.

3 Robustness and rate of convergence

We begin with our assumptions and some preliminary lemmas—proofs can be found in the Appendix. Let \( \{w_i : i \geq 1\} \) be the user-specified weight sequence in the PR algorithm, and define the sequence of partial sums \( a_n = \sum_{i=1}^n w_i \) \((a_0 \equiv 0)\). In addition to A1–A2 in Section 2, assume the following:

A3. \( w_n > 0, w_n \downarrow 0, \sum_n w_n = \infty, \sum_n w_n^2 < \infty \), and \( \sum_n a_{n-1} w_n^2 < \infty \).

A4. There exists \( B < \infty \) such that

\[
\sup_{\theta_1, \theta_2 \in \Theta} \int_{\mathcal{F}} \left[ \frac{p(x|\theta_1)}{p(x|\theta_2)} \right]^2 m(x) d\nu(x) < B.
\]

Condition A3 is satisfied if \( w_n \approx n^{-\gamma} \), for \( \gamma \in (2/3, 1] \). The square-integrability condition A4 is the strongest assumption, but it does hold for many common mixands, such as Normal or Poisson, provided that \( \Theta \) is compact and \( m \) admits a moment-generating function on \( \Theta \). If one is willing to assume that \( m \in \mathcal{M}_\Theta \), then A4 can be replaced by a less restrictive condition, depending only on \( p(x|\theta) \); cf. assumption A5 in TMG.

Our new results are partially based on calculations in TMG, and we begin by recording a few of these for future reference. First, let \( R(x) \) be the remainder term of a first-order Taylor approximation of \( \log(1 + x) \) at \( x = 0 \); that is, \( \log(1 + x) = x - x^2 R(x), \ x > -1 \), where \( R(x) \) satisfies

\[
0 \leq R(x) \leq \max\{1, (1 + x)^{-2}\}, \ x > -1.
\]  

Next, write the PR estimate \( m_n \in \mathcal{M}_\Theta \) in (1.3) as

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

where

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

For notational convenience, also define the function

\[
H_{n,y}(x) = \frac{h_{n,y}(x)}{m_{n-1}(x)} - 1, \ x, y \in \mathcal{X}.
\]
Then the KL divergence $K_n = K(m, m_n)$ satisfies

$$K_n - K_{n-1} = \int m \log(m_{n-1}/m_n) \, d\nu$$

$$= - \int m \log(1 + w_n H_{n,X_n}) \, d\nu$$

$$= -w_n \int m H_{n,X_n} \, d\nu + w_n^2 \int m H_{n,X_n}^2 R(w_n H_{n,X_n}) \, d\nu$$

where $R(x)$ satisfies (3.1). Let $\mathcal{A}_{n-1}$ be the $\sigma$-algebra generated by the data sequence $X_1, \ldots, X_{n-1}$. Since $K_{n-1}$ is $\mathcal{A}_{n-1}$-measurable, upon taking conditional expectation with respect to $\mathcal{A}_{n-1}$ we get

$$E(K_n|\mathcal{A}_{n-1}) - K_{n-1} = -w_n T(f_{n-1}) + w_n^2 E(Z_n|\mathcal{A}_{n-1}) \quad (3.2)$$

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

$$T(\varphi) = \int_\Theta \left\{ \int_X \frac{m(x)}{m_\varphi(x)} p(x|\theta) \, d\nu(x) \right\}^2 \varphi(\theta) \, d\mu(\theta) - 1, \quad \varphi \in \mathcal{F} \quad (3.3)$$

$$Z_n = \int m H_{n,X_n}^2 R(w_n H_{n,X_n}) \, d\nu \quad (3.4)$$

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

**Lemma 3.1.** (a) $T(\varphi) \geq 0$ with equality iff $K(m, m_\varphi) = K(m, M_\Theta)$.
(b) Under A1, A2 and A4, $T$ is continuous in the weak topology on $\mathcal{F}$.

For some further insight into the relationship between $T(\varphi)$ and the KL divergence $K(m, m_\varphi)$, define

$$D(\theta; \varphi) = \int \frac{m(x)}{m_\varphi(x)} p(x|\theta) \, d\nu(x) - 1$$

and notice that $T(\varphi) = \int D^2(\theta; \varphi) \varphi(\theta) \, d\mu(\theta)$. Some analysis shows that $D(\theta; \varphi)$ is the Gâteaux derivative of $K(m, \eta)$ at $\eta = m_\varphi$ in the direction of $p(\cdot|\theta)$. Now, if $T(\varphi) = 0$, then $D(\theta; \varphi) = 0$ for $\mu$-almost all $\theta$ and, hence,

$$D(\psi; \varphi) = \int D(\theta; \varphi) \psi(\theta) \, d\mu(\theta),$$

the Gâteaux derivative of $K(m, \eta)$ at $\eta = m_\varphi$ in the direction of $m_\varphi$, is zero for all $\psi \in \mathcal{F}$. The fact that the Gâteaux derivative vanishes in all directions suggests that $m_\varphi$ is a point at which the infimum $K(m, M_\Theta)$ is attained, and this is exactly the conclusion of Lemma 3.1(a). A similar characterization of the NPMLE is given in Lindsay (1983).

According to (3.2) and Lemma 3.1(a), $K_n$ would be a non-negative supermartingale were it not for the term involving $Z_n$. Fortunately, since $Z_n$ is almost surely bounded, its presence causes no major difficulties.
Lemma 3.2. Under A4, $Z_n \leq 1 + B$ a.s. for all $n$.

Our last preliminary result of this section gets us closer to a convergence rate for $K(m, m_n)$. Define $K_n = K(m, m_n) - K(m, M_{n\Theta}) \geq 0$.

Lemma 3.3. Under A3–A4, $\sum_n w_n K_n < \infty$ a.s.

We are ready state and prove our main result. Those convergence properties advertised in Section 1 are corollaries to Theorem 3.4 that follows.

Theorem 3.4. Under A1–A4, $a_n K_n \to 0$ a.s.

Proof. Define two sequences of random variables,

$$\delta_n = \sum_{i=n+1}^{\infty} a_{i-1} w_i^2 E(Z_i | \mathcal{F}_n) \quad \text{and} \quad \varepsilon_n = \sum_{i=n+1}^{\infty} w_i E(K_i^{*} | \mathcal{F}_n),$$

and define $\tilde{K}_n = a_n K_n + \delta_n + \varepsilon_n \geq 0$. Using (3.2) we find that

$$E(K_n | \mathcal{F}_{n-1}) - \tilde{K}_n = E(a_n K_n + \delta_n + \varepsilon_n | \mathcal{F}_{n-1}) - a_{n-1} K_{n-1} - \delta_{n-1} - \varepsilon_{n-1} - w_n K(m, M_{n\Theta}) \leq a_n E(K_n | \mathcal{F}_{n-1}) + \sum_{i=n+1}^{\infty} a_{i-1} w_i^2 E[Z_i | \mathcal{F}_n] | \mathcal{F}_{n-1}] + \sum_{i=n+1}^{\infty} E[E(K_i | \mathcal{F}_n) | \mathcal{F}_{n-1}] - a_{n-1} K_{n-1} - \sum_{i=n}^{\infty} a_{i-1} w_i^2 E[Z_i | \mathcal{F}_{n-1}] - \sum_{i=n}^{\infty} w_i E(K_i | \mathcal{F}_{n-1}).$$

After some simplification, using the fact that $\mathcal{F}_{n-1} \subset \mathcal{F}_n$, we have

$$E(K_n | \mathcal{F}_{n-1}) - \tilde{K}_n \leq a_n E(K_n | \mathcal{F}_{n-1}) - a_{n-1} K_{n-1} - a_{n-1} w_n^2 E(Z_n | \mathcal{F}_{n-1}) - w_n E(K_n | \mathcal{F}_{n-1}) = a_{n-1} [E(K_n | \mathcal{F}_{n-1}) - K_{n-1} - w_n^2 E(Z_n | \mathcal{F}_{n-1})] = -a_{n-1} w_n T(f_{n-1}) \leq 0$$

Thus, $\tilde{K}_n$ forms a non-negative supermartingale so, by the martingale convergence theorem, there is a limit $\tilde{K}_\infty \geq 0$ with finite expectation. But assumption A3, Lemma 3.2, and the dominated convergence theorem ensure that $\delta_n$ and $\varepsilon_n$ both converge a.s. to zero; therefore, $a_n K_n \to \tilde{K}_\infty$ a.s.

It remains to show that $\tilde{K}_\infty = 0$ a.s. Suppose, on the other hand, that $\tilde{K}_\infty > 0$ with positive probability. From the previous display we have

$$\tilde{K}_{i-1} - E(\tilde{K}_i | \mathcal{F}_{i-1}) \geq a_{i-1} w_i T(f_{i-1}) \quad \forall i \geq 2.$$  \hspace{1cm} (3.5)
Fix any two integers \( N \) and \( n \). Taking conditional expectation with respect to \( \mathcal{A}_{N-1} \) in (3.5) and summing gives

\[
\sum_{i=N}^{N+n} a_{i-1} w_i E[T(f_i-1)|\mathcal{A}_{N-1}] \leq \sum_{i=N}^{N+n} E[\tilde{K}_{i-1} - E(\tilde{K}_i|\mathcal{A}_{i-1}) | \mathcal{A}_{N-1}]
\]

\[
= \sum_{i=N}^{N+n} \left\{ E(\tilde{K}_{i-1}|\mathcal{A}_{N-1}) - E[E(\tilde{K}_i|\mathcal{A}_{i-1}) | \mathcal{A}_{N-1}] \right\}
\]

\[
= \sum_{i=N}^{N+n} E(\tilde{K}_{i-1} - \tilde{K}_i | \mathcal{A}_{N-1})
\]

\[
= \tilde{K}_{N-1} - E(\tilde{K}_{N+n}|\mathcal{A}_{N-1})
\]

\[
\leq \tilde{K}_{N-1}
\]

We are assuming \( \tilde{K}_\infty > 0 \), so there exists \( \varepsilon > 0 \) such that

\[
K(m,m_n) > K(m,M_\Theta) + \varepsilon
\]

for all but perhaps finitely many \( n \). Recall the proof of Lemma 2.1, which shows that the mapping \( \kappa(\varphi) = K(m,m_\varphi) \) is lower semi-continuous with respect to the weak topology on \( F \). Consequently,

\[
G_\varepsilon := \{ \varphi \in F : \kappa(\varphi) > K(m,M_\Theta) + \varepsilon \} \subset F.
\]

is a weakly open set, and its closure \( \overline{G}_\varepsilon \) is compact by A1. Since \( f_n \in \overline{G}_\varepsilon \) for all but finitely many \( n \), Lemma 3.1 implies that \( T(f_n) \) is bounded away from zero. This and A3 imply that the left-most sum in (3.6) goes to \( \infty \) with positive probability as \( n \to \infty \). But the upper bound \( \tilde{K}_{N-1} \) is almost surely finite—a contradiction. Therefore, \( \tilde{K}_\infty = 0 \) almost surely.

Next we show that \( K_n^* \to 0 \) implies \( \|m_n - m_f\| \to 0 \), where \( \| \cdot \| \) denotes the \( L_1(\nu) \) norm.

**Corollary 3.5.** Under A1–A4, \( m_n \) converges to \( m_f \) in \( L_1(\nu) \).

**Proof.** Suppose this is not true—that is, there is an \( \varepsilon > 0 \) and a subsequence \( \{n_s\} \) such that \( \|m_{n_s} - m_f\| > \varepsilon \) for all \( s \). Then, by A1, this sequence has a further subsequence \( \{n_{s(t)}\} \) such that \( f_{n_{s(t)}} \) converges weakly to some \( f_\infty \).

By A2, \( m_{n_{s(t)}} \) converges to \( m_\infty := m_f \) pointwise and in \( L_1(\nu) \) by Scheffé’s theorem; therefore, \( m_\infty \neq m_f \). Define \( u_t = m_{n_{s(t)}}/m_\infty - 1 \). Then \( u_t \to 0 \) pointwise and \( \{u_t\} \) is uniformly integrable with respect to \( md\nu \) by A4 and Jensen’s inequality; indeed,

\[
\sup_t \int u_t^2 m \, d\nu = \sup_t \int \left( \frac{m_{n_{s(t)}}}{m_\infty} - 1 \right)^2 m \, d\nu < 1 + B.
\]
Then Theorem 3.4, together with Vitali’s theorem, implies that
\[
0 \leq K(m, m_\infty) - K(m, m_f) = K(m, m_\infty) - \lim_{t \to \infty} K(m, m_{n,t}) = \lim_{t \to \infty} \int m \log(m_{n,t}/m_\infty) \, d\nu \leq \lim_{t \to \infty} \int u_{t} \, d\nu = 0
\]

Therefore, \( K(m, m_\infty) = K(m, m_f) \), and uniqueness of the KL projection implies \( m_\infty = m_f \), which contradicts our supposition.

Theorem 3.4 and Corollary 3.5 seem to indicate that the estimates \( f_n \) converge to some \( f \in \mathcal{F} \) at which the infimum \( K(m, M_\Theta) \) is attained. However, to conclude weak convergence of \( f_n \) from \( L_1 \) convergence of \( m_n \), we need two additional conditions:

A5. Identifiability: \( m_\varphi = m_\psi \) \( \nu \text{-a.e.} \) implies \( \varphi = \psi \) \( \mu \text{-a.e.} \).

A6. For any \( \varepsilon > 0 \) and any compact \( A \subset X \), there exists a compact \( B \subset \Theta \) such that \( \int_A p(x|\theta) \, d\nu(x) < \varepsilon \) for all \( \theta \not\in B \).

With conditions A5–A6 and Theorem 3 of TMG, the next result follows immediately from Corollary 3.5.

**Corollary 3.6.** Under A1–A6, \( f_n \to f \) a.s. in the weak topology, where \( f \in \mathcal{F} \) is the unique mixing density that satisfies \( K(m, m_f) = K(m, M_\Theta) \). In particular, if \( m \in M_\Theta \), then \( f_n \) is a consistent estimate of the true mixing density, in the weak topology.

In the mis-specified case, even though \( K_\star^* \to 0 \) implies \( ||m_n - m_f|| \to 0 \), the \( L_1(\nu) \) rate does not easily follow without extra assumptions, such as boundedness of \( m_f/m \). But a Hellinger contrast rate is a direct consequence of Theorem 3.4. In the well-specified case, when \( m_f = m \), the Hellinger contrast reduces to the usual Hellinger distance, so our convergence rate results are comparable to those of, say, Genovese and Wasserman (2000).

**Corollary 3.7.** Choose \( w_n \approx n^{-\gamma} \), \( \gamma \in (2/3, 1] \), and assume A1–A4.

(a) \( \rho(m_n, m_f) = o(n^{-(1-\gamma)/2}) \) a.s.

(b) If \( m_f/m \in L_\infty(m \, d\nu) \), then \( ||m_n - m_f|| = o(n^{-(1-\gamma)/2}) \) a.s.

**Proof.** Let \( \Gamma_n = \int (m_n/m_f) \, d\nu \). An important well known property of KL projections that will be used in the proof is \( \Gamma_n \leq 1 \); see, e.g., Theorem 3 of Barron (2000) or Lemma 2.1 of Patilea (2001).
Part (a) follows from Lemma 2.4 of Patilea (2001). Indeed,

\[ 2\rho^2(m, m_f) = \int \left( \sqrt{\frac{m}{m_f}} - 1 \right)^2 m \, d\nu \]

\[ = (\Gamma_n - 1) + 2 \int \left( 1 - \sqrt{\frac{m}{m_f}} \right) m \, d\nu \]

\[ \leq 2 \int \log \sqrt{\frac{m_f}{m}} m \, d\nu \]

\[ = K^*_n \]

For part (b), let \( q_n = \frac{mm_n}{m_f}\Gamma_n \), and notice that \( K^*_n \geq K(m, q_n) \). Then, Pinsker’s inequality gives

\[ \sqrt{2K^*_n} \geq m - q_n = \frac{1}{\Gamma_n} \left\| \frac{m_n}{m_f} - \Gamma_n \right\|_{L_1(m \, d\nu)} \geq \frac{m_n}{m_f} - \Gamma_n \]

The triangle inequality for \( \| \cdot \|_{L_1(m \, d\nu)} \) implies

\[ \left\| \frac{m_n}{m_f} - 1 \right\|_{L_1(m)} \leq \left\| \frac{m_n}{m_f} - \Gamma_n \right\|_{L_1(m \, d\nu)} + (1 - \Gamma_n), \]

and hence

\[ \sqrt{2K^*_n} + (1 - \Gamma_n) \geq \left\| \frac{m_n}{m_f} - 1 \right\|_{L_1(m \, d\nu)}. \]

The right-hand side is related to the \( L_1(\nu) \) error via Holder’s inequality

\[ \|m_n - m_f\| = \int |m_n - m_f| \, d\nu \]

\[ = \int \frac{m_f}{m} |\frac{m_n}{m_f} - 1| m \, d\nu \leq C \left\| \frac{m_n}{m_f} - 1 \right\|_{L_1(m)} \]

where \( C := \|m_f/m\|_{L_\infty(m \, d\nu)} \) is finite by assumption. Therefore,

\[ \|m_n - m_f\| \leq C \left\{ 2K^*_n + (1 - \Gamma_n) \right\}, \]

and part (b) follows from the fact that \( (1 - \Gamma_n) \leq K^*_n. \)

\[ \square \]

4 A generalized PR algorithm

To satisfy the conditions of Theorem 3.4, one typically would need the mixing parameter space \( \Theta \) to be compact. Also, in order to calculate the estimate in practice, this \( \Theta \) must be known since integration over \( \Theta \) is required in (1.2). These requirements are similar to those in Ghosal and van der Vaart (2001) and can be somewhat restrictive, particularly when there is no natural choice.
of \( \Theta \). A potential solution is to use a *mixture sieve*, which allows the support of the estimated mixing distribution to grow with the sample size. A motivation for this dynamic choice of support would be that, eventually, the support will be large enough so that the class of all mixtures over that support will be sufficiently rich. Borrowing on this idea, we propose a sieve-like extension of the PR algorithm which, instead of requiring \( \Theta \) to be fixed and known, incorporates a sequence of compact mixing parameter spaces that increases with \( n \).

Let \( \Theta \) denote the mixing parameter space, which may or may not be compact. For example, if the model is a Gaussian location-scale mixture, then \( \Theta = \mathbb{R} \times \mathbb{R}_+ \).

The generalized PR algorithm is as follows.

**Algorithm GPR.** Choose an increasing sequence of compact sets \( \Theta_n \) such that \( \Theta_n \uparrow \Theta \), a bounded, strictly positive, \( \mu \)-measurable function \( g(\theta) \), and a sequence \( c_n \geq 0 \) that satisfies

\[
\sum_n \log(1 + c_n) < \infty.
\]

Define \( g_n(\theta) = g(\theta) I_{\Theta_n \setminus \Theta_{n-1}}(\theta)/d_n \), where \( d_n = \int_{\Theta_n \setminus \Theta_{n-1}} g \, d\mu \) is the normalizing constant. Start with an initial estimate \( f_0 \) on \( \Theta_0 \) and, for \( n \geq 1 \), define

\[
f_n^*(\theta) = (1 - w_n) f_{n-1}(\theta) + w_n \frac{p(X_n|\theta)f_{n-1}(\theta)}{m_{n-1}(X_n)}, \quad \theta \in \Theta_{n-1}
\]

and then

\[
f_n(\theta) = \begin{cases} 
\frac{1}{1 + c_n} f_n^*(\theta), & \theta \in \Theta_{n-1} \\
\frac{c_n}{1 + c_n} g_n(\theta), & \theta \in \Theta_n \setminus \Theta_{n-1} \\
0, & \theta \in \Theta_n^c
\end{cases}
\]  \hspace{1cm} (4.1)

As in (1.3), define \( m_n := m_{f_n} \) as the final estimate of \( m \).

The motivation for using \( g_n \) is that \( \Theta_n \) is small compared to \( \Theta \), so the estimate should be padded near the boundary of \( \Theta_n \) to compensate for the possibly heavy tails of \( f \) which might assign considerable mass to \( \Theta_n^c \). Asymptotically, any function \( g \) would suffice, so we recommend a default choice \( g(\theta) \equiv 1 \), making \( g_n(\theta) \) a Uniform density on \( \Theta_n \setminus \Theta_{n-1} \).

Next we derive convergence properties of the GPR estimate \( m_n \). The primary obstacle in extending the results in Section 3 is that now the support of the mixing densities is changing with \( n \)—this makes the comparisons of the mixing densities in the proof of Lemma 3.3 more difficult. Here we will consider only the case where \( m = m_f \in \mathcal{M}_\Theta \), but the mixing density \( f \) and its support \( \Theta_f \subset \Theta \) are both unknown. Theorem 4.1 below establishes a bound on the rate of convergence in the case where \( \Theta_f \) is a compact subset of \( \Theta \). It is important to point out that, while we are restricting ourselves to the compact case, we need not assume \( \Theta_f \) to be known.

Recall condition A4 in Section 3 requiring that the likelihood ratio be square integrable uniformly over \( \Theta \). In this case, we have a sequence \( \Theta_n \), and we require a bound similar to that of A4 for each \( \Theta_n \). To this end, define the sequence

\[
B_n = \sup_{\theta_1, \theta_2, \theta_3 \in \Theta_n} \int \left[ \frac{p(x|\theta_1)}{p(x|\theta_2)} \right]^2 p(x|\theta_3) \, d\nu(x), \quad n \geq 0.
\]  \hspace{1cm} (4.2)
Since $\Theta_n \subset \Theta_{n+1}$, the sequence $B_n$ is clearly increasing; if we are to push the proof of Theorem 3.4 through in this more general situation, we will need to control how fast $B_n$ increases.

**Theorem 4.1.** Assume that $\Theta_f \subset \Theta$ is compact and that conditions A2–A3 hold. Furthermore, assume that $\sum_n a_{n-1} w_n^2 B_n < \infty$. Then the GPR estimates $m_n$ satisfy $a_n K (m_f, m_n) \to 0$ a.s.

See the Appendix for the proof. In general, the conditions of the theorem are satisfied if $w_n \asymp (n^\alpha \log n)^{-1}$, $\alpha \in [2/3, 1]$, and $B_n = O(\log n)$. (4.3)

**Example 4.2.** Suppose that $p(x|\theta) = e^{-\theta x^2} / x!$ is a Poisson density (with respect to counting measure). Then

$$\sum_x \left[ \frac{p(x|\theta_1)}{p(x|\theta_2)} \right]^2 \int p(x|\theta_3) = \exp\{2(\theta_2 - \theta_1) - \theta_3 + \theta_1^2 \theta_3 / \theta_2^2\}.$$

Take $\Theta_n = [\alpha_n, \beta_n]$ where $\alpha_n$ and $\beta_n$ are to be determined. Then,

$$B_n = \sup_{\theta_1, \theta_2, \theta_3 \in \Theta_n} \exp\{2(\theta_2 - \theta_1) - \theta_3 + \theta_1^2 \theta_3 / \theta_2^2\} \leq \exp\{\beta_n^3 / \alpha_n^2\}.$$

If $\beta_n \asymp (c + \log \log n)^{1/5}$, for some constant $c > 0$ such that $\Theta_0$ suitably large, and $\alpha_n = \beta_n^{-1}$, then it is easy to check that $B_n = O(\log n)$. Therefore, the conditions of Theorem 4.1 are satisfied with $w_n$ as in (4.3).

**Acknowledgments**

The authors thank Professor J. K. Ghosh for his support and guidance throughout this and other projects.

**A Proofs**

**Proof of Lemma 3.1.** For part (a), treat $\theta$ as a random element in $\Theta$, whose distribution has density $\phi \in F$, and define the random variable

$$g_{\phi}(\theta) = \int \frac{m(x)}{m_{\phi}(x)} p(x|\theta) d\nu(x).$$

Then $E_{\phi}\{g_{\phi}(\theta)\} = \int g_{\phi} \phi d\mu = 1$ and $T(\phi) = V_{\phi}\{g_{\phi}(\theta)\} \geq 0$, with equality if and only if $g_{\phi} = 1$ $\mu$-a.e. Next define

$$\Lambda(\phi) = \log \left\{ \int g_{\phi}(\theta) f(\theta) d\mu(\theta) \right\}.$$
where \( f \in \mathcal{F} \) is such that \( K(m, m_f) = K(m, M_\Theta) \). Note that \( T(\varphi) = 0 \) implies \( \Lambda(\varphi) = 0 \). By Jensen’s inequality

\[
\Lambda(\varphi) = \log \left\{ \int \frac{m_f(x)}{m_\varphi(x)} m(x) \, d\nu(x) \right\} \\
\geq \int \log \left( \frac{m_f(x)}{m_\varphi(x)} \right) m(x) \, d\nu(x) \\
= K(m, m_\varphi) - K(m, m_f) \\
\geq 0
\]

so that \( \Lambda(\varphi) = 0 \) implies \( K(m, m_\varphi) = K(m, m_f) \).

For part (b), take a sequence \( \{ \varphi_s \} \subset \mathcal{F} \) such that \( \varphi_s \) converges weakly to some \( \varphi \in \mathcal{F} \), and let \( g_{\varphi_s} \) and \( g_\varphi \) be as in (A.1). Let \( r_s(x, \theta) = p(x|\theta)/m_\varphi(x) \) so that \( g_{\varphi_s}(\theta) = \int r_s(x, \theta) m(x) \, d\nu(x) \). By A4 and Jensen’s inequality,

\[
\sup_s \int r_s^2(x, \theta) m(x) \, d\nu(x) \leq B,
\]

which implies \( \{ r_s(\cdot, \theta) \} \) is uniformly integrable with respect to \( m \, d\nu \); therefore, \( g_{\varphi_s} \to g_\varphi \) \( \mu \)-a.e. Since the weak topology on \( \mathcal{F} \) is metrizable, to prove continuity of \( T \) it suffices to show that \( T(\varphi_s) \to T(\varphi) \). We have

\[
|T(\varphi_s) - T(\varphi)| = \left| \int g_{\varphi_s}^2 \varphi \, d\mu - \int g_\varphi^2 \varphi \, d\mu \right| \\
\leq \int (g_{\varphi_s}^2 - g_\varphi^2) \varphi \, d\mu + \int g_\varphi^2 (\varphi_s - \varphi) \, d\mu. \tag{A.2}
\]

The second term on the right-hand side of (A.2) goes to zero by definition of weak convergence, since \( g_\varphi^2 \leq B \) (by A4) and \( g_\varphi^2 \) is continuous (by A1). Next, we know the following:

- \(|g_{\varphi_s}^2 - g_\varphi^2| \varphi_s \to 0 \) \( \mu \)-a.e.,
- \(|g_{\varphi_s}^2 - g_\varphi^2| \varphi_s \leq 2B \varphi_s \), and
- \( \varphi_s \to \varphi \) pointwise \( \mu \)-a.e. and \( \lim \int \varphi_s \, d\mu = 1 = \int \varphi \, d\mu \).

Now, for the first term on the right-hand side of (A.2),

\[
\left| \int (g_{\varphi_s}^2 - g_\varphi^2) \varphi_s \, d\mu \right| \leq \int |g_{\varphi_s}^2 - g_\varphi^2| \varphi_s \, d\mu \to 0,
\]

where convergence follows from the above properties and Theorem 1 of Pratt (1960). Therefore, \( T(\varphi_s) \to T(\varphi) \) and since the sequence \( \{ \varphi_s \} \) was arbitrary, \( T \) must be continuous.

**Proof of Lemma 3.2.** Note that for \( a > 0 \) and \( b \in (0, 1) \), we have

\[
(a - 1)^2 \max\{1, (1 + b(a - 1))^{-2}\} \leq \max\{(a - 1)^2, (1/a - 1)^2\}. \tag{A.3}
\]
Combining inequalities (3.1) and (A.3) we see that

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

and, since both \( h_{n,X_n} \) and \( m_{n-1} \) belong to \( M_\Theta \) for each \( n \), we conclude from A4 and Jensen’s inequality that \( Z_n \leq 1 + B \) a.s.

Proof of Lemma 3.3. This was proved in TMG for the case \( m \in M_\Theta \). Here we only sketch the details for the more general case. Let \( f \in F \) be such that \( K(m, m_f) = K(m, M_\Theta) \). TMG’s Taylor approximation and telescoping sum argument for the KL divergence between mixing densities reveals

\[ K(f, f_n) - K(f, f_0) = -\sum_{i=1}^{n} \int_{\Theta} f \log(f_i / f_{i-1}) \, d\mu \]

\[ = \sum_{i=1}^{n} w_i V_i - \sum_{i=1}^{n} w_i M_i + \sum_{i=1}^{n} w_i^2 E_i \]

where, with a slight difference in notation compared to TMG,

\[ M_i = \int \left( \frac{m_f}{m_{i-1}} - 1 \right) m \, d\nu, \quad V_i = 1 - \frac{m_f(X_i)}{m_{i-1}(X_i)} + M_i, \]

and the \( E_i \)’s are uniformly bounded (by A4). From \( \log(x) \leq x - 1 \) and the proof of Lemma 3.1(a), we get

\[ M_i \geq \int \log(m_f / m_{i-1}) m \, d\nu = K^*_i - 1 \geq 0. \]

To prove the claim, it suffices to show that \( \sum_{i=1}^{n} w_i M_i \) converges almost surely to a finite random variable. The same arguments in TMG show that both \( V_\infty := \sum_{i=1}^{\infty} w_i V_i \) and \( E_\infty := \sum_{i=1}^{\infty} w_i^2 E_i \) are almost surely finite, and non-negativity of the KL divergence implies

\[ \sum_{i=1}^{n} w_i M_i \leq K(f, f_0) + V_\infty + E_\infty < \infty \quad \forall \ n. \]

Then the left-hand side converges because the \( M_i \)’s are non-negative.

Proof of Theorem 4.1. The proof is essentially the same as that of Theorem 3.4; we simply need to check that Lemmas 3.2 and 3.3 continue to hold in this new context.

Since \( B_n \) is increasing, the condition \( \sum_{n} a_{n-1} w_n^2 B_n < \infty \) implies that the conclusion of Lemma 3.2 holds, and the remainder term \( Z_n \) in (3.4) satisfies \( \sum_{n} w_n^2 E(Z_n) < \infty \). The key observation is that, since \( \Theta_f \) is compact, there is a number \( N = N(f) \) such that \( \Theta_f \subset \Theta_n \) for all \( n \geq N \). Consequently, there are only \( N \) iterates \( f_0, \ldots, f_{N-1} \) such that \( K(f, f_i) = \infty \) so, after appropriately
shifting the time scale, we can apply the telescoping sum argument in the proof of Lemma 3.3 to get

\[
K(f, f_{N+n}) - K(f, f_N) = - \sum_{i=N+1}^{N+n} \int f \log(f_i / f) \, d\mu
\]

\[
= \sum_{i=N+1}^{N+n} w_i V_i - \sum_{i=N+1}^{N+n} w_i M_i + \sum_{i=N+1}^{N+n} w_i^2 E_i + \sum_{i=N+1}^{N+n} \log(1 + c_i)
\]

The right-most sum in the display converges by assumption and the others can be handled as in the proof of Lemma 3.3. Therefore, we can conclude that \( \sum_{n=N}^\infty w_n K(m_f, m_n) < \infty \) a.s. To finish the proof, simply throw away the first \( N \) iterations and apply the argument used to prove Theorem 3.4.

\[\square\]

References


