Finite mixture models with predictive recursion marginal likelihood

Ryan Martin
Department of Mathematical Sciences
Indiana University–Purdue University Indianapolis
rgmartin@math.iupui.edu

January 30, 2011

Abstract

Estimation of finite mixture models when the mixing distribution support is unknown is an important and challenging problem. In this paper, a new approach is given based on the recently proposed predictive recursion marginal likelihood (PRML) method. By taking a sufficiently fine grid as a set of candidate support points, one may treat the support itself as an unknown parameter to be estimated. The PRML approach asymptotically integrates out the mixing distribution itself, leaving an approximate marginal likelihood for the support, which can be used for estimation. We employ a computationally efficient version of simulated annealing for the large-scale combinatorial optimization problem. Theory is given which shows that the PRML estimate will asymptotically identify the best mixture model supported on a subset of the candidate grid, where "best" is measured with respect to the Kullback–Leibler divergence on the mixture scale. Real and simulated data examples show that the PRML method compares favorably to existing Bayesian and non-Bayesian methods in terms of mixture density estimation accuracy and model parsimony.

Keywords and phrases: Density estimation; Dirichlet distribution; mixture complexity; simulated annealing; stochastic approximation.

1 Introduction

That complicated data sets can be described by a mixture of a few relatively simple models is an interesting and practically useful phenomenon. The relevant theoretical result is that any density can be well-approximated by an appropriate finite mixture. The catch, however, is that this finite mixing distribution is generally difficult to specify. In cluster analysis, for example, the configuration of the mixing distribution is exactly the quantity of interest. Likewise, for empirical Bayes inference, an estimate of the prior/mixing distribution is required. For this reason, using the observed data to estimate the finite mixing distribution is an important problem. Key references include Titterington et al. (1985), McLachlan and Basford (1988), Richardson and Green (1997), Roeder and

Wasserman (1997), McLachlan and Peel (2000), and Woo and Sriram (2006). An alternative approach, related to what we consider here, is in the context of nonparametric Bayes (Ghosh and Ramamoorthi 2003). A typical strategy is to model the mixing distribution itself as a random draw from a Dirichlet process distribution (Ferguson 1973). Discreteness properties of the Dirichlet process imply that the distribution of the observables is almost surely a finite mixture, where the number of mixture components, as well as the component-specific parameters, are random quantities. This flexible modeling strategy effectively allows the data to determine the mixture structure. Efficient Markov chain Monte Carlo algorithms (Escobar and West 1995; MacEachern and Müller 1998; Neal 2000) are now available to fit the Dirichlet process mixture model to data, and numerous density estimation, regression, and clustering applications have been considered; see, for example, Müller and Quintana (2004) and the references therein.

For general mixture models, the fast *predictive recursion* (PR) algorithm (Newton 2002; Newton et al. 1998) has received some recent attention. Martin and Ghosh (2008) use results from stochastic approximation theory to prove consistency of the PR estimates in finite mixtures with known support, and present numerical results showing that PR is competitive with classical methods. Martin (2011) extends the consistency result, giving a nearly root-n rate of convergence for PR in the finite mixture problem. For the general case of compact but known support, Tokdar et al. (2009) prove consistency of PR estimates of the mixing and mixture distributions, and Martin and Tokdar (2009) extend these results to the case of mis-specified mixtures and obtain a bound on the rate of convergence. But, for the most part, these convergence theorems do not apply when the mixing distribution support is unknown; see Martin and Tokdar (2009, Sec. 5) for the one exception. In fact, the PR algorithm itself is not directly applicable in the unknown support case. The main goal of this paper to develop a PR-based approach to estimate mixture models when the finite mixing distribution support is unknown.

Specifically, we assume that data $Y_1, \ldots, Y_n$ are independent observations from a common distribution with density $m(y)$, modeled as a finite mixture

$$m_{f,U}(y) = \sum_{u \in U} p(y \mid u) f(u), \quad y \in \mathscr{Y}, \quad U \subset \overline{\mathscr{U}}, \tag{1}$$

where $\overline{\mathscr{U}}$ is a known compact set and $(y, u) \mapsto p(y \mid u)$ is a known kernel on $\mathscr{Y} \times \overline{\mathscr{U}}$, such as Gaussian or Poisson, but the finite support set U and the mixing weights $f = \{f(u) : u \in U\}$ are unknown. A classical approach is nonparametric maximum likelihood, which estimates (f, U) by maximizing $\prod_{i=1}^n m_{f,U}(Y_i)$. The goal of maximum likelihood is to give an estimate $\hat{m}_{\text{MLE}}$ that fits the observed data well, so there are no built-in concerns about the size of the estimated support. In our experiments we find that maximum likelihood estimates the mixture well, but the support of the estimated mixing distribution tends to be too large. Moreover, maximum likelihood estimates are known to be sensitive to model mis-specification. An elegant and robust alternative is proposed by Woo and Sriram (2006). They use the HMIX algorithm (Cutler and Cordero-Brãna 1996) to minimize the Hellinger distance of a finite mixture with support size S to a nonparametric estimate of m. This distance, in turn, is used to construct a model selection criterion for choosing S. The procedure of James et al. (2001), based on minimizing a Kullback–Leibler divergence, is similar. However, our simulations indicate that the Woo-Sriram procedure can be too aggressive, often selecting S too small for accurate estimation of m.

In this paper we employ a version of the PR marginal likelihood (PRML) method of Martin and Tokdar (2011b) to estimate both U and f in (1) for a generic kernel $p(y \mid u)$. The main idea is to chop up $\overline{\mathscr{U}}$ into a finite but arbitrarily fine grid $\mathscr{U}$, and search for the best approximation to m by mixtures supported on subsets of $\mathscr{U}$. Thus, the essentially nonparametric problem is transformed to a very high-dimensional parametric one. The PRML approach, described in Section 3.1, takes advantage of a close connection between PR and the Bayes procedure with a finite-dimensional Dirichlet distribution on the mixing weights $f = \{f(u) : u \in U\}$, for a fixed support U, to construct an approximate marginal likelihood for U with f integrated out. Estimation of the support proceeds by maximizing this approximate marginal likelihood over $2^{\mathscr{U}}$, the collection of all subsets of the candidate support grid $\mathscr{U}$. For this high-dimensional combinatorial optimization problem, we propose, in Section 3.2, a fast version of the simulated annealing algorithm. Thus, this novel PRML approach can be viewed as a hybrid stochastic approximation–simulated annealing alternative to the Bayesian's Markov chain Monte Carlo.

For flexibility, the finite set $\mathscr{U}$ of candidate support points should be large. But it is often the case that one believes that the true support size, or *mixture complexity*, is considerably smaller than $|\mathscr{U}|$. To account for these prior beliefs, we recommend a regularized version of PRML in Section 3.3 that includes an additional term in the PR marginal likelihood that penalizes supports $U \subseteq \mathscr{U}$ which are too large. In particular, we suggest a penalty determined by a binomial prior on $|U|$, with success probability parameter chosen to reflect the user's belief about the true mixture complexity.

Asymptotic convergence properties of the PRML estimates are presented in Section 3.4. We show that, for given $\mathscr{U}$, the PRML method will asymptotically identify the best mixture over all those supported on subsets of $\mathscr{U}$. In particular, if the mixture model is correctly specified, and the true mixing distribution support is a subset of $\mathscr{U}$, then the PRML estimate is consistent. Here "best" is measured in terms of the Kullback–Leibler divergence on the mixture scale; thus, PRML acts asymptotically like a minimum distance estimate (Cutler and Cordero-Braña 1996; Woo and Sriram 2006) albeit for a fixed candidate support grid. But unlike the Woo-Sriram estimates, the PRML estimate of the support size will always converge to a finite number, and will be consistent if the true mixing distribution support is a subset of $\mathscr{U}$. Two numerical examples are considered in Section 3.5 where it is shown that the PRML gives results comparable to those given elsewhere in the literature. A simulation study is presented in Section 3.6 in the context of a simple finite Gaussian mixture model. There we show that PRML outperforms a number of popular alternatives in terms of both accuracy of mixture density estimation and model parsimony.

In principle, the PRML approach can handle mixtures over any number of parameters, but the simulated annealing algorithm in Section 3.2 is time-consuming for mixtures over two or more parameters. In Section 4 we modify the proposed simulated annealing optimization algorithm to give a fast approximation to the PRML solution general finite location-scale mixtures. This approximation focuses on a justifiable class of admissible subsets and this restriction can substantially decrease the complexity of the combinatorial optimization problem to solve. We reconsider a simulation study in James et al. (2001) and Woo and Sriram (2006), showing that this approximate PRML procedure estimates well the mixture complexity in a challenging Gaussian mixture model.

2 PR and PRML for general mixtures

Consider the general problem where the common marginal density $m(y)$ for $Y_1, \ldots, Y_n$ is modeled as a nonparametric mixture

$$m_f(y) = \int_{\mathscr{U}} p(y \mid u) f(u)\, d\mu(u), \quad y \in \mathscr{Y}, \tag{2}$$

where $\mathscr{U}$ is a known set, not necessarily finite, 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 the σ-finite Borel measure μ on $\mathscr{U}$. Newton (2002) presents the following algorithm, called predictive recursion (PR), for nonparametric estimation of f.

PR Algorithm. Fix an initial guess $f_0(u)$ on $\mathscr{U}$ and a sequence of weights $\{w_i : i \geq 1\} \subset (0,1)$. Then, for $i = 1, \ldots, n$, compute

$$f_i(u) = (1 - w_i) f_{i-1}(u) + w_i \, \frac{p(Y_i \mid u) f_{i-1}(u)}{\int_{\mathscr{U}} p(Y_i \mid u') f_{i-1}(u')\, d\mu(u')} \tag{3}$$

Return f_n and $m_n = m_{f_n}$ as the final estimates of f and m, respectively.

Key properties of PR include its fast computation and its ability to estimate a mixing density f absolutely continuous with respect to any user-defined dominating measure μ. That is, unlike the nonparametric maximum likelihood estimate which is almost surely discrete (Lindsay 1995) regardless of μ, the PR estimate can be discrete, continuous, or both, depending on the user's choice of dominating measure. Throughout this paper, we take μ to be counting measure on a finite set $\mathscr{U}$, but see Martin and Tokdar (2011a) for an application of PR where μ is Lebesgue measure on $\mathbb{R}$ plus a point mass at the origin.

The PR estimates depend on the order in which the data $Y_1, \ldots, Y_n$ enter the recursion. This order-dependence is irrelevant asymptotically, but for finite samples, the estimates are mildly sensitive to the choice of ordering. To reduce this dependence in practice, the estimates are usually averaged over a set of randomly chosen permutations. The speed of PR for a fixed ordering makes this permutation averaging computationally feasible. In the numerical examples that follow, we consider averages over 25 data permutations.

Large-sample properties of the PR estimates can be obtained under fairly weak conditions on the kernel $p(y|u)$ and the true data-generating density $m(y)$. Let $\mathbb{M}$ denote the set of all mixtures m_f in (2) as f ranges over $\mathbb{F}$, and for two densities m and m' let $K(m, m') = \int \log\{m(y)/m'(y)\} m(y)\, dy$ denote the Kullback–Leibler divergence of m' from m. Then Tokdar et al. (2009) prove consistency of m_n and f_n when $m \in \mathbb{M}$. When $m \notin \mathbb{M}$, Martin and Tokdar (2009) show that there exists a mixing density $f^\star$ in $\overline{\mathbb{F}}$, the weak closure of $\mathbb{F}$, such that $K(m, m_{f^\star}) = \inf\{K(m, m_f) : f \in \overline{\mathbb{F}}\}$, and m_n converges almost surely to $m_{f^\star}$. As a corollary, if the mixture (2) is identifiable, then $f_n \to f^\star$ almost surely in the weak topology. Moreover, for a certain choice of weights $\{w_n\}$, Martin and Tokdar (2009) obtain a conservative $o(n^{-1/6})$ bound on the rate of convergence. Martin (2011) builds on the stochastic approximation representation of PR in Martin and Ghosh (2008) to prove that, in the case of finite mixtures with known support, $m_n \to m_{f^\star}$ almost surely at nearly a parametric $O(n^{-1/2})$ rate. The author believes that a nearly root-n rate can be achieved more generally, but this remains to be confirmed.

These asymptotic robustness properties of PR lead to an attractive construction of a new procedure for estimation in semiparametric mixture models. Martin and Tokdar (2011b) refer to this procedure PR marginal likelihood, or PRML for short. It is often difficult to fully specify the parametric kernel $p(y|u)$ in the nonparametric mixture (2), and an alternative is to consider a class of kernels $p(y|u,\theta)$, indexed by a parameter $\theta \in \Theta$, and allow the data to choose θ. This is a semiparametric mixture model

$$m_{f,\theta}(y) = \int_{\mathscr{U}} p(y \mid u, \theta) f(u)\, d\mu(u), \quad y \in \mathscr{Y}, \tag{4}$$

where both f and θ are unknown and to be estimated. Martin and Tokdar (2011b) prove, under certain conditions, that, with probability 1,

$$\ell_n(\theta) := \sum_{i=1}^{n} \log m_{i-1,\theta}(Y_i) = -nK^\star(\theta) + O(n), \tag{5}$$

where the dominating term $K^\star(\theta) = \inf\{K(m, m_{f,\theta}) : f \in \overline{\mathbb{F}}\}$ plays the role of a model selection tool for an oracle who knows that optimal mixing density f for each θ. The result (5) suggests that maximizing $\ell_n(\theta)$ is asymptotically equivalent to minimizing this oracle model selector. Consistency of the maximizer $\hat{\theta}_n = \arg\max \ell_n(\theta)$ for general parameter spaces Θ remains an open question, but Martin and Tokdar (2011b) demonstrate its good empirical performance in a variety of simulations.

To conclude this section, note that it is not necessary that the parametric part θ of the semiparametric mixture $m_{f,\theta}$ in (4) be an unknown characteristic of the kernel. In fact, the dependence of $m_{f,\theta}$ on θ can be mostly arbitrary. The only restriction is that the dominating measure μ cannot depend on θ.

3 PRML for finite mixtures

The problem of estimating finite mixture densities with unknown mixing distribution support is an important and challenging problem, and in this section we present a solution based on the PRML procedure in Section 2. To the author's knowledge, this is the first time the unknown support problem has been treated as a type of semiparametric mixture model estimation problem.

3.1 Unknown support problem and PRML

Recall the basic setup in Section 1. That is, we have a compact set $\overline{\mathscr{U}}$ over which the finite mixture is to be considered. A finite lattice $\mathscr{U} \subset \overline{\mathscr{U}}$ is chosen as a set of candidate mixture component locations. The motivation is that if $\mathscr{U}$ is sufficiently fine, then the data-generating density m, which is assumed approximable by some finite mixture, can in fact be well-approximated by a mixture supported on an appropriate subset U of $\mathscr{U}$.

Next we show that the PRML procedure, with the support $U \subset 2^{\mathscr{U}}$ playing the role of the structural parameter θ in (4). Choose a fixed support set U and consider the following hierarchical model:

$$Y_i \mid (f, U) \overset{\text{iid}}{\sim} m_{f,U}, \quad \text{and} \quad f \mid U \sim \mathsf{P}_U, \tag{6}$$

where P_U is a generic prior for the random discrete distribution f supported on U. For this model, it can be shown, using linearity of the mixture, that the marginal likelihood for U is of the form

$$L_{n,\text{marg}}(U) = \int \Big\{ \prod_{i=1}^{n} m_{f,U}(Y_i) \Big\} \, d\mathsf{P}_U(f) = \prod_{i=1}^{n} \sum_{u \in U} p(Y_i \mid u) \hat{f}_{i-1,U}(u), \tag{7}$$

where $\hat{f}_{i-1,U} = \mathsf{E}_U(f \mid Y_1, \ldots, Y_{i-1})$ is the posterior mean.

Equation (7) is the jumping off point for the PRML approximation. Towards this, take P_U to be a finite-dimensional Dirichlet distribution on U with precision parameter $\alpha_0 > 0$ and base measure $f_{0,U}$, a probability vector indexed by U. Then the Polya urn representation of the Dirichlet distribution (Ghosh and Ramamoorthi 2003, Sec. 3.1.2) implies that

$$\hat{f}_{1,U}(u) = \frac{\alpha_0}{\alpha_0 + 1} f_{0,U}(u) + \frac{1}{\alpha_0 + 1} \frac{p(Y_1 \mid u) f_{0,U}(u)}{\sum_{u' \in U} p(Y_1 \mid u') f_{0,U}(u')},$$

a mixture of the prior guess and a predictive distribution given Y_1. If $\alpha_0 = 1/w_1 - 1$, then $\hat{f}_{1,U}(u)$ is exactly $f_1(u)$ in (3). This correspondence holds exactly only for a single observation, but Martin and Tokdar (2011b) argue that PR acts as a dynamic, mean-preserving filter approximation to the Bayes estimate. Therefore, plugging in the PR estimate $f_{i-1,U}(u)$ for the Bayes estimate $\hat{f}_{i-1,U}(u)$ in (7) gives a PR-based approximation of the marginal likelihood $L_{n,\text{marg}}(U)$. In this case, the log PR marginal likelihood is

$$\ell_n(U) = \sum_{i=1}^{n} \log m_{i-1,U}(Y_i) = \sum_{i=1}^{n} \log \Big\{ \sum_{u \in U} p(Y_i \mid u) f_{i-1}(u) \Big\}. \tag{8}$$

Maximizing $\ell_n(U)$ over all possible U is a formidable task. Our simplification is to consider only those U in $2^{\mathscr{U}}$ for the finite set of candidate support points $\mathscr{U}$ defined above. When $|\mathscr{U}|$ is relatively large, this is still a challenging optimization problem. In the next section, we give a simulated annealing algorithm to solve this combinatorial optimization problem.

3.2 Computation

As described above, maximizing $\ell_n(U)$ over all subsets $U \in 2^{\mathscr{U}}$ is a combinatorial optimization problem. The challenge is that the finite set $2^{\mathscr{U}}$ is enormous, even if $|\mathscr{U}|$ is only of moderately large size, so it is not feasible to evaluate $\ell_n(U)$ for each U. The simulated annealing procedure is a stochastic algorithm where, at each iteration t, a move from the current state $U^{(t)}$ to a new state $U^{(t+1)}$ is proposed so that $\ell_n(U^{(t+1)})$ will tend to be larger than $\ell_n(U^{(t)})$. Next we describe our specific version of simulated annealing for the PRML application.

An important feature of simulated annealing is the decreasing temperature sequence $\{\tau_t : t \geq 0\}$. Following Hajek (1988) and Bélisle (1992), we take the default choice $\tau_t = a/\log(1+t)$ for a suitable a, chosen by trial-and-error. For the numerical examples that follow, $a = 1$ gives acceptable results.

To simplify the discussion, to each subset $U \subset \mathscr{U} = \{u_1, \ldots, u_S\}$, associate a binary S-vector $H \in \{0,1\}^S$, where $S = |\mathscr{U}|$. Then $H_s = 1$ if $u_s \in U$ and $H_s = 0$ otherwise. In other words, H_s determines whether u_s is in or out of the mixture. It clearly suffices to define the optimization of $\ell_n(U)$ over $2^{\mathscr{U}}$ in terms of the H vectors. Then the simulated annealing algorithm goes as follows.

1. Choose a starting point $H^{(0)}$, and a number of iterations T.
2. Generate a sequence $\{H^{(t)} : t = 0, 1, \ldots, T\}$:
   (a) Simulate H_{new} from a probability distribution $\pi^{(t)}$ on $\{0,1\}^S$, possibly depending on t and the current iterate $H^{(t)}$.
   (b) Define the acceptance probability
   $$\alpha(t) = 1 \wedge \exp\big[\big\{\ell_n(H_{\text{new}}) - \ell_n(H^{(t)})\big\}/\tau_t\big],$$
   where $\ell_n(H)$ is the PR marginal likelihood defined in (8), written as a function of the indicator H that characterizes U, and set
   $$H^{(t+1)} = \begin{cases} H_{\text{new}} & \text{with probability } \alpha(t) \\ H^{(t)} & \text{with probability } 1 - \alpha(t) \end{cases}$$
   (c) If $t < T$, set $t \leftarrow t + 1$ and return to Step 2(a); else, go to Step 3.
3. Return the $H^{(t)}$ visited with largest $\ell_n(H^{(t)})$.

In our implementation, the initial choice is $H_s^{(0)} = 1$ for each s, which corresponds to the full mixture. Also, in what follows, $T = 2000$ iterations of the simulated annealing algorithm above seems sufficient to identify a good clustering. The key to the success of simulated annealing is that while all uphill moves are taken, some downhill moves, to "less likely" U_{new}, are allowed through the flip of a $\alpha(t)$-coin in Step 2(b). This helps prevent the algorithm from getting stuck at local modes. But the vanishing temperature τ_t makes these downhill moves less likely when t is large.

It remains to specify the proposal distribution $\pi^{(t)}$ in Step 2(a). In our examples, a draw H_{new} from $\pi^{(t)}$ differs from $H^{(t)}$ in exactly $k \geq 1$ positions. In other words, k of the S components of $H^{(t)}$ are chosen and then each is flipped from 0 to 1 or from 1 to 0. The choice of components is not made uniformly, however. To encourage a solution with a relatively small number of mixture components, we want $\pi^{(t)}$ to assign greater mass to those components $H_s^{(t)}$ in $H^{(t)}$ such that $H_s^{(t)} = 1$. The particular choice of weights is

$$\pi_s^{(t)} \propto 1 + \Big(\frac{S}{\sum_{s=1}^S H_s^{(t)}}\Big)^r \cdot H_s^{(t)}, \quad s = 1, \ldots, S, \quad r \geq 1.$$

Here we see that when most of the components of $H^{(t)}$ are 1, equivalently, when $|U^{(t)}|$ is large, the sampling is near uniform, whereas, when $H^{(t)}$ is sparse, those components with value 1 have a greater chance of being selected. For the examples that follow, we take $k = 1$ and $r = 1$.

Note that the stochastic approximation representation of PR, combined with the simulated annealing algorithm described above for PRML optimization, gives a hybrid stochastic approximation–simulated annealing algorithm, call it SASA, for estimating finite mixture models. This SASA procedure serves as an alternative to the Bayesian's MCMC for finite mixtures.

3.3 Regularization

In the hierarchical model (6), it would not be unnatural to introduce a prior for U to complete the hierarchy. Martin and Tokdar (2011a) propose an extension of the PRML framework in which priors for structural parameters are incorporated into the model, effectively replacing the marginal likelihood with a marginal posterior. They refer to this procedure as "regularized" PRML and show the advantage of such regularization in large-scale hypothesis testing applications.

In our context, a prior $\Pi(U)$ for the support U should be designed to reflect the degree of sparsity in the mixture representation. Since $S = |\mathscr{U}|$ will typically be large—much larger than the unknown support is likely to be—we want to penalize those U with many components. To accomplish this, we recommend a prior for U with a binomial prior for the size $|U|$. The parameters of the binomial prior are (S, ρ), where ρ denotes the prior probability that an element of $\mathscr{U}$ will be included in U. The parameter ρ can be adjusted to penalize candidate supports which are too large. For example, one might be able to guess/elicit an *a priori* reasonable expected number of components, say 5. In which case, one should choose $\rho = 5/S$. In the absence of such information, a crude choice of ρ follows from estimating m with some standard density estimate $\hat{m}$ and taking $\rho =$ (number of modes of $\hat{m}$)$/S$.

3.4 Asymptotic theory

Suppose that $\mathscr{U}$ is a fixed finite set and $U \subseteq \mathscr{U}$ is any subset. Throughout this discussion, we shall assume that $u \mapsto p(y \mid u)$ is continuous for each y, and that the PR weight sequence $\{w_i : i \geq 1\}$ is given by $w_i = (i+1)^{-\gamma}$ for some γ in $(1/2, 1)$. In this case, Martin (2011) proves the following theorem.

Theorem 1. *Suppose the mapping $f \mapsto m_{f,U}$ is one-to-one. Then there exists a unique minimizer $f^\star = f^\star_U$ of $K(m, m_{f,U})$, for the given U, and the PR estimate f_n converges almost surely to $f^\star$, as $n \to \infty$, at a rate $O(n^{-(1-1/2\gamma)})$.*

From Theorem 1 it follows that f_n and the corresponding mixture $m_{n,U}$ both converge at a nearly parametric root-n rate when the exponent γ of the PR weights is close to 1. In the numerical examples in Sections 3.5 and 3.6 we take $\gamma = 1$. This rate is a drastic improvement over the conservative bounds given in Martin and Tokdar (2009).

But our focus here is on the problem of unknown mixing distribution support. Specifically, we use the PR marginal likelihood to estimate the unknown finite support. Towards a convergence theorem, consider a normalized version of the PR marginal likelihood, namely

$$K_n(U) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{m(Y_i)}{m_{i-1,U}(Y_i)} = -\frac{\ell_n(U) - \sum_{i=1}^{n} \log m(Y_i)}{n}. \tag{9}$$

It follows from the analysis of Martin and Tokdar (2011b) that $K_n(U)$ converges almost surely to

$$K^\star(U) = \inf\{K(m, m_{f,U}) : f \in \overline{\mathbb{F}}\}, \tag{10}$$

where $\overline{\mathbb{F}}$ is the closed probability simplex in $\mathbb{R}^{|U|}$. But since U ranges over the finite set $2^{\mathscr{U}}$, this pointwise convergence is actually uniform. Theorem 2 states this asymptotic result formally. First we give two important assumptions.

Assumption 1. There exists a finite constant A such that

$$\max_{U \subseteq \mathscr{U}} \max_{u_1, u_2 \in U} \int \left\{ \frac{p(y \mid u_1)}{p(y \mid u_2)} \right\}^2 m(y)\, dy \leq A.$$

Assumption 2. There exists a finite constant B such that

$$\max_{U \subseteq \mathscr{U}} \int \left\{ \log \frac{m(y)}{m_{f^\star, U}(y)} \right\}^2 m(y)\, dy \leq B,$$

where $f^\star = f^\star_U$ is where the infimum in (10) is attained.

Assumption 1 holds for many common kernels, such as Gaussian or Poisson, provided that $m(y)$ admits a moment generating function. Assumption 2 is a statement about the quality of the mixture model for m; it follows from Assumption 1 when the model is correctly specified. Then the following is an immediate consequence of Theorem 2 in Martin and Tokdar (2011b).

Theorem 2. *Under Assumptions 1–2, $K_n(U)$ converges almost surely to $K^\star(U)$, uniformly in $U \subseteq \mathscr{U}$. Consequently, $\widehat{U}_n = \arg\max \ell_n(U)$ converges almost surely to $U^\star = \arg\min K^\star(U)$.*

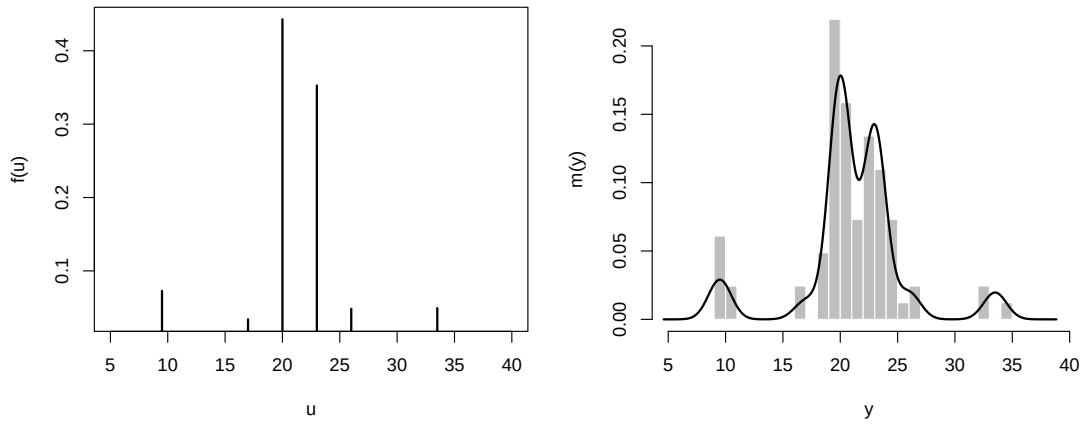
In words, the maximum PRML estimate $\widehat{U}_n$ converges to the "best" support $U^\star \subseteq \mathscr{U}$ in the sense that $\widehat{U}_n$ and $U^\star$ will eventually have the same elements. If m is indeed a mixture with mixing distribution supported on a subset of $\mathscr{U}$, then $\widehat{U}_n$ is a consistent estimate of the true support. Regarding estimation of the mixture complexity our results here differ considerably from those in, say, James et al. (2001) and Woo and Sriram (2006). In particular, once $\mathscr{U}$ is specified, the PRML estimate of the mixture complexity is bounded by $|\mathscr{U}|$, whether the model is correctly specified or not, and is guaranteed to converge. The estimates of James et al. (2001) and Woo and Sriram (2006) of the mixture complexity explode to infinity in the mis-specified case. The author believes that, in the mis-specified case, PRML's asymptotic identification of the best finite mixture in a sufficiently large class is more meaningful.

The initial choice of $\mathscr{U}$ and, in particular, $|\mathscr{U}|$ is not obvious, however. An interesting proposition is to let $\mathscr{U} = \mathscr{U}_n$ depend on the sample size n, like a sieve. The idea is that, if the set $\mathscr{U}$ of candidate support points is sufficiently large, then the class of mixtures supported on subsets of $\mathscr{U}$ should be rich enough to closely approximate m. For example, suppose that m is a finite mixture with support points somewhere in the compact bounding set $\overline{\mathscr{U}}$. Then it should be possible to choose $\mathscr{U}_n$ to saturate the bounding set $\overline{\mathscr{U}}$ at a suitable rate so that $K(m, m_{n,\widehat{U}_n}) \to 0$ almost surely. To prove such a result in our context, bounds on the constants associated with the rate in Theorem 1 are needed, since these would most likely depend on $|U|$. We leave this for future work.

3.5 Illustrative examples

3.5.1 Galactic velocity data

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



(a) Mixing distribution (b) Mixture distribution

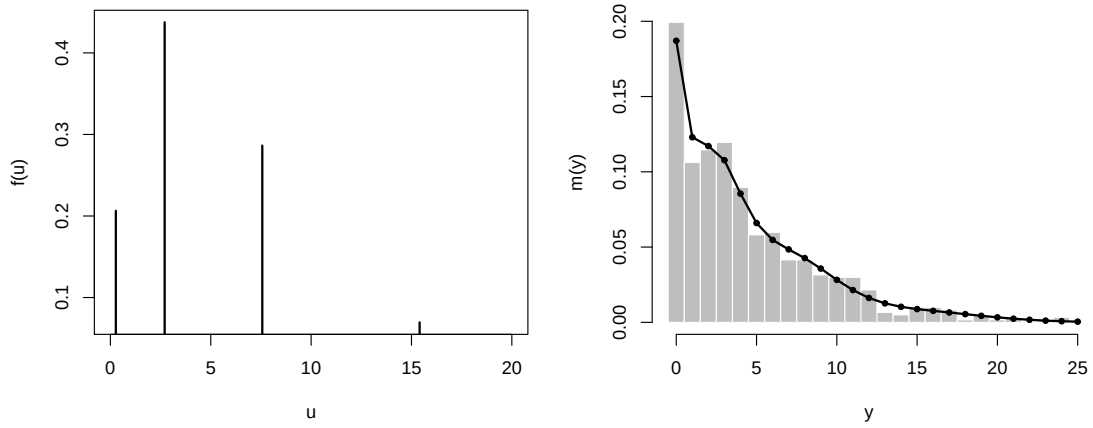
Figure 1: Plots of the PRML estimates for the galactic velocity data in Section 3.5.1.

$n = 82$ galaxies. She models this data as a finite mixture of Gaussian distributions, with the number and location of 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 $\sigma^2 = 1$. This choice is based on the *a priori* considerations of Escobar and West (1995): their common prior for the variance of each Gaussian component has unit mean. From the observed velocities, it is apparent that the mixture components must 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 PRML estimates of the mixing and mixture distribution based on the simulated annealing optimization procedure. The PR marginal likelihood $\ell_n(U)$ in (8) is averaged over 25 random permutations of the data within the simulated annealing optimization to reduce dependence on the data-ordering. Despite the permutations and the slow convergence of simulated annealing, the full estimation procedure here takes only a few seconds to complete. The results here are based on the regularized PRML with the binomial prior chosen to have expected value 5, but the conclusions seem to be fairly robust to this choice. The PRML estimate of the mixing distribution clearly identifies six galaxy clusters, closely matching the conclusions in Roeder (1990), Escobar and West (1995), and Richardson and Green (1997), and the PRML mixture density also provides a very good fit to the observed velocities.

3.5.2 Thailand illness data

The observed data comes from a cohort study in north-east Thailand where the number of illness spells for $n = 602$ pre-school children is monitored between June 1982 and September 1985. A histogram is given in panel (b) of Figure 2. Böhning (2000) points out that an ordinary Poisson model is inadequate for this data due to over-dispersion. A Poisson mixture, therefore, seems more appropriate. Upon investigation of the data, it is



(a) Mixing distribution (b) Mixing distribution

Figure 2: Plots of the PRML estimates for the Thailand illness data in Section 3.5.2. In (b) the dots represent the tips of the mixture probability mass function.

apparent that the support of the finite mixture must be within $\overline{\mathscr{U}} = [0, 20]$. Therefore, we take our candidate support $\mathscr{U}$ to be a grid of $|\mathscr{U}| = 75$ equispaced points from 0 to 20. The PRML estimates are displayed in Figure 2. We use the same regularization and permutation averaging as in Section 3.5.1. Here we see that the PRML solution provides a very good fit to the observed counts. Moreover, our mixing density estimate closely matches the nonparametric maximum likelihood results Wang (2007), both in support values and in weights. Compared to the maximum likelihood estimation in Wang (2007), PRML estimation for this example is a bit more time consuming; this is because PRML does not allow a reduction of the large data set to sufficient statistics, a frequency table in this case.

3.6 A simulation experiment

Consider the case where the data-generating density $m(y)$ is a finite Gaussian mixture:

$$m(y) = 0.11\mathsf{N}(y \mid -5.0, 1) + 0.56\mathsf{N}(y \mid 0.0, 1) + 0.33\mathsf{N}(y \mid 3.5, 1), \tag{11}$$

where $\mathsf{N}(y \mid u, \sigma^2)$ is a Gaussian density with mean u and variance σ^2. Here we shall investigate the performance of several methods of estimating this finite mixture distribution based on independent data $Y_1, \ldots, Y_n$ sampled from (11). The four methods are:

- The PRML method described above. Specifically, we take $\overline{\mathscr{U}} = [-6, 5]$ and $\mathscr{U}$ a grid of 50 equispaced points in $\overline{\mathscr{U}}$. Note that the true support $\{-5.0, 0.0, 3.5\}$ is *not* contained in $\mathscr{U}$. We choose a fixed set of 25 data permutations over which the PRML is averaged in the optimization step. The regularization parameter ρ is chosen by counting modes of a standard Gaussian kernel density estimate, as described in Section 3.3.

- A nonparametric Bayes method where f is modeled as a random draw from a Dirichlet process distribution. Specifically, we take $f \sim \mathsf{DP}(\alpha, f_0)$, where the pre-

cision parameter α is taken to be 1 and base measure f_0 is taken to be $\mathsf{N}(\mu, \tau\sigma^2)$. The parameters μ and τ are estimated via standard empirical Bayes methods. We employ the Gibbs sampling algorithm of Escobar and West (1995) to generate 1000 clustering configurations to estimate the number of mixture components as well as the mixture density. We shall refer to this as method as DPM.

- Nonparametric maximum likelihood using the algorithm of Wang (2007). Specifically, we estimate (f, U) by maximizing $\prod_{i=1}^{n} \sum_{u \in U} \mathsf{N}(Y_i \mid u, 1) f(u)$ over all (f, U). We shall refer to this method as MLE.

- The method of Woo and Sriram (2006). The main idea is to compare mixtures of various sizes to a nonparametric kernel density estimate of m via Hellinger distance, and choose the smallest mixture with satisfactory fit. We shall refer to this method as WS. Note that, for the model in question, with constant scale parameter, the adaptive kernel density estimate in Woo and Sriram (2006, Sec. 4) is not needed.

In this experiment, we consider 100 independent samples of size $n = 100$ taken from the mixture (11). We are primarily interested in the respective estimates of the mixture complexity and the mixture densities, and Table 1 summarizes the results. The first panel tallies $|U|$, the number of clusters, for each method over the 100 replications. Both DPM and MLE frequently over-estimate. PRML and WS each hit the right number of clusters close to 90% of the time, but how they miss the target is also important. WS is aggressive in choosing the number of clusters, tending to under-estimate, while PRML never misses an existing cluster. That PRML never misses a cluster pays off in the estimation of m. The second panel in Table 1 summarizes the values of $K(m, \hat{m})$, scaled by a factor of 100, for the various estimates $\hat{m}$'s. Performance is similar across methods in terms of the median $K(m, \hat{m})$. However, that WS tends to miss existing clusters, means that $\hat{m}_{\mathrm{ws}}$ nearly vanishes in regions where m does not, which causes the Kullback–Leibler divergence to be large in such cases; see the large entry for $\max K(m, \hat{m}_{\mathrm{ws}})$. The other methods do not exhibit this instability and, in particular, PRML turns out to be a hair better than DPM and MLE, on average, in this experiment.

Regarding computation time, MLE and WS, in that order, are the fastest. PRML, which takes about 2 seconds on average in this experiment, is still noticeably faster than DPM, despite the permutation averaging and the need to optimize over a 2^{50}-dimensional space. So, overall, it seems that PRML is a strong competitor among these powerful existing methods. Similar conclusions are reached when the mixture model is mis-specified, for example, when $m(y)$ is a mixture of Student-t densities but is modeled as a mixture of Gaussians.

4 Approximate PRML for finite location-scale mixtures

4.1 Setup and algorithm

In principle, the PRML procedure is able to handle finite mixtures of any type of kernel. However, when $\mathscr{U}$ is a lattice in a higher-dimensional space, the computations become

	Number of clusters							Summary of $100K(m, \hat{m})$				
	2	**3**	4	5	6	7	8	Min	Q1	Med	Q3	Max
PRML		**88**	11			1		0.33	1.43	2.69	3.70	10.80
DPM				7	67	25	1	0.56	1.81	2.80	3.79	11.90
MLE		**21**	54	23	2			0.40	1.71	2.64	3.92	10.70
WS	13	**87**						0.32	1.49	2.47	4.59	120.70

Table 1: Summary of the 100 estimates of $|U|$ and $K(m, \hat{m})$ for the four methods in the simulation experiment in Section 3.6.

somewhat costly. For a two-parameter kernel, for example, the approach outlined above would be to construct a lattice in the two-dimensional u-space and use the same in/out simulated annealing algorithm as in Section 3.2 for pairs $u = (u_1, u_2)$. The collection $2^{\mathscr{U}}$ of all such pairs is, in general, quite large so it is advantageous to introduce a simpler approximation of the two-parameter mixture model. Our approach starts with the observation that, in general, the full two-parameter model could potentially have pairs (u_1, u_2) and (u_1, u_2') both entering the mixture. Our simplification is to rule out such cases, allowing at most one instance of, say, u_1 in the mixture. This reduces the size of the search space and, in turn, accelerates the simulated annealing optimization step. Here we develop this modification for the important special case of location-scale mixtures.

Let $\overline{\Theta}$ and $\overline{\Sigma}$ be closed intervals in $\mathbb{R}$ and $\mathbb{R}^+$, respectively, assumed to contain the range of values the location θ and scale σ can take. Chop up these intervals into sufficiently fine grids Θ and Σ of sizes $S(\theta) = |\Theta|$ and $S(\sigma) = |\Sigma|$, respectively. Define the rectangle $\overline{\mathscr{U}} = \overline{\Theta} \times \overline{\Sigma}$ and the two-dimensional lattice $\mathscr{U} = \Theta \times \Sigma$. Then the finite mixture model is just as before

$$m(y) = \sum_{(\theta,\sigma)\in U} p(y \mid \theta, \sigma) f(\theta, \sigma), \quad U \subset \mathscr{U},$$

where the kernel $p(y \mid \theta, \sigma)$ equals $\sigma^{-1} p\big((y-\theta)/\sigma\big)$ for some symmetric unimodal density function $p(\cdot)$. This covers the case of finite location-scale Gaussian mixtures, but also the robust class of finite Student-t mixtures with a common fixed degrees of freedom. We will focus on the Gaussian case.

What makes this different from before is that U can contain at most one pair in each "row" of $\Theta \times \Sigma$. To accommodate this restriction, we shall modify the simulated annealing algorithm proposed in Section 3.2. The key idea here is to continue to use the location as the main parameter, but adjust the in/out scheme from before to allow for various levels of "in." Recall the indicators H_s in Section 3.2. Here we use the notation $H = (H_1, \ldots, H_{S(\theta)})$, where each H_s takes values in $\{0, 1, 2, \ldots, S(\sigma)\}$ to characterize the support set U. The interpretation is

$$H_s = \begin{cases} 0 & \text{if } \theta_s \text{ is not in the mixture} \\ h & \text{if pair } (\theta_s, \sigma_h) \text{ is in the mixture, } h = 1, \ldots, S(\sigma). \end{cases} \tag{12}$$

In other words, location θ_s enters the mixture only if $H_s > 0$, but can enter paired with any of the scales σ_h depending on the non-zero value of H_s. Since there is a one-to-one correspondence between admissible subsets $U \subset \mathscr{U}$ and vectors H of this form,

we shall formulate the PRML optimization problem in terms of H. By restricting the estimates to this collection of admissible subsets, the state space to search goes from $2^{S(\theta)\times S(\sigma)}$ down to $[S(\sigma)+1]^{S(\theta)}$, which can be a drastic reduction. To maximize the PR marginal likelihood $\ell_n(H)$ over the set of all admissible H, we propose a modification of the foregoing simulated annealing algorithm. In particular, the structure of the algorithm presented in Section 3.2 remains the same; all that changes is the proposal distribution.

At iteration t, define $\beta(t) = S(\theta)^{-1}\sum_{s=1}^{S(\theta)} I\{H_s^{(t)} = 0\}$, the proportion of zero entries in $H^{(t)}$. Now sample an entry in $H^{(t)}$ with probabilities

$$\pi_s^{(t)} \propto 1 + \big(1-\beta(t)\big)^{-r} \cdot I\{H_s^{(t)} > 0\}, \quad s = 1, \ldots, S(\theta). \tag{13}$$

When $H^{(t)}$ has many zero entries, $1-\beta(t)$ will be small, so the non-zero entries will have greater chance of being sampled. Let $H_s^{(t)}$ be the chosen entry. To define H_{new}, there are two cases:

- If $H_s^{(t)} = 0$, take $H_{\text{new}} \sim \mathsf{Unif}\{1, \ldots, S(\sigma)\}$.
- If $H_s^{(t)} > 0$, take $H_{\text{new}} = 0$ with probability $\beta(t)$ and

$$H_{\text{new}} \sim \mathsf{Unif}\{H_s^{(t)} - 1, H_s^{(t)} + 1\} \quad \text{with probability } 1-\beta(t).$$

  If $H_s^{(t)} = 1$ or $S(\sigma)$, then H_{new} would be 2 or $S(\sigma)-1$, respectively.

The idea is to maintain the entry sampling that encourages a sparse mixture. This is accomplished by, first, encouraging the selection of non-zero $H^{(t)}$ entries. Second, these selected non-zero entries will likely be set to zero as the algorithm proceeds because $\beta(t)$ will tend to increase with t. Thus, only the crucial components of Θ should remain in the mixture as t increases.

Once H_{new} has been sampled, the simulated annealing algorithm decides to take $H_s^{(t+1)}$ as H_{new} or $H_s^{(t)}$ depending on the flip of the $\alpha(t)$-coin as in Step 2(b) in Section 3.2. As before, if H_{new} is a better candidate support than $H_s^{(t)}$ then the move will be accepted. But allowing some moves to worse candidates helps prevent the simulated annealing procedure from getting stuck at local modes.

4.2 Galactic velocity example, revisited

Consider again the galaxy velocity data example described in Section 3.5.1. We produce the PRML estimate of a location-scale mixture of Gaussians. Here Θ is the same as $\mathscr{U}$ in Section 3.5.1, and $\Sigma = \{0.5, 0.6, \ldots, 1.4, 1.5\}$. As before we average the PR marginal likelihood over 25 randomly chosen permutations. But, in this case, we choose $r = 3$ in (13) to further discourage sampling of supports with too many components. The PRML method estimates five Gaussian components with varying scales, and Figure 3 shows the resulting estimate of the density. In this case, the overall fit is as good as before, but only five components are needed. It is interesting and somewhat counter-intuitive that when the scale parameters are allowed to vary across mixture components, we actually get fewer components than for fixed $\sigma = 1$ as in Section 3.5.1.

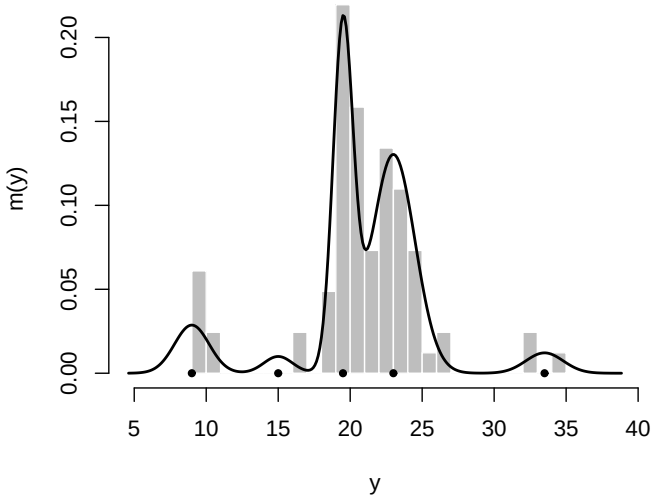


Figure 3: Plot of the PRML estimates of the location-scale mixture density for the galactic velocity data in Section 4.2, with the estimated support points plotted (•) on the horizontal axis.

4.3 Another simulation experiment

In this section we present a simulation experiment in which we focus on estimating the number of components in a challenging Gaussian mixture model considered in James et al. (2001) and Woo and Sriram (2006). The particular model is

$$m(y) = 0.25\mathsf{N}(y \mid -0.3, 0.05) + 0.5\mathsf{N}(y \mid 0, 10) + 0.25\mathsf{N}(y \mid 0.3, 0.05). \tag{14}$$

The two components with variance 0.05 makes for two nearby but dramatic modes. With small sample sizes especially, it should be relatively difficult to detect these two distinct components. For this model, accurate estimation of the number of components requires varying scale parameters, and we investigate the performance of the approximate PRML procedure outlined in Section 4.1.

Table 2 summarizes the PRML estimates of the mixture complexity based on 100 random samples from the mixture model $m(y)$ in (14) with four different sample sizes: $n = 50$, 250, 500, and 1000. In particular, we take $\overline{\Theta} = [-2, 2]$, $\overline{\Sigma} = [0.1, 4.0]$ and Θ and Σ are equispaced grids of length $S(\theta) = 40$ and $S(\sigma) = 25$, respectively. Note that the true location and scale parameters in (14) do not belong to $\Theta \times \Sigma$. The simulated annealing optimization procedure in Section 4.1 is used to optimize the PR marginal likelihood over the collection of admissible subsets, which provides an estimate of the mixture complexity. In this case, there are $2^{40\times 25} \approx 10^{301}$ subsets of $\Theta \times \Sigma$, compared to $26^{40} \approx 4\times 10^{56}$ admissible subsets, so there is a substantial computational savings in using the approximation in Section 4.1. For comparison, we also include estimates based on the methods of Woo and Sriram (2006), denoted by *MHDE*, the estimates of James et al. (2001), denoted by *MKE* and *NKE*, the Bayesian estimates of Roeder and Wasserman (1997), denoted by *RW*, the estimates of McLachlan (1987), denoted by *Bootstrap*, and finally the estimates of Henna (1985), denoted by *Henna*. The RW method performs well

for small n but seems to falter as n increases, while the MKE method does well for large n. The PRML method does quite well for $n = 50$ and, although it is not the best, it is competitive in all other cases. In particular, it seems that only the MHDE method of Woo and Sriram (2006) is as good or better than PRML at correctly identifying the true mixture complexity across simulations. But, depending on the application, one might prefer PRML because it does not have such a strong tendency as the MHDE method to under-estimate the true mixture complexity.

Acknowledgments

The author thanks Professors J. K. Ghosh, Surya T. Tokdar, and Chuanhai Liu for a number of helpful discussions and suggestions.

References

Bélisle, C. J. P. (1992), "Convergence theorems for a class of simulated annealing algorithms on $\mathbf{R}^d$," *J. Appl. Probab.*, 29, 885–895.

Böhning, D. (2000), *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping, and Others*, Boca Raton: Chapman and Hall–CRC.

Cutler, A. and Cordero-Braña, O. I. (1996), "Minimum Hellinger distance estimation for finite mixture models," *J. Amer. Statist. Assoc.*, 91, 1716–1723.

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

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, 1, 209–230.

Ghosh, J. K. and Ramamoorthi, R. V. (2003), *Bayesian Nonparametrics*, New York: Springer-Verlag.

Hajek, B. (1988), "Cooling schedules for optimal annealing," *Math. Oper. Res.*, 13, 311–329.

Henna, J. (1985), "On estimating of the number of constituents of a finite mixture of continuous distributions," *Ann. Inst. Statist. Math.*, 37, 235–240.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent estimation of mixture complexity," *Ann. Statist.*, 29, 1281–1296.

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

MacEachern, S. and Müller, P. (1998), "Estimating mixture of Dirichlet process models," *J. Comput. Graph. Statist.*, 7, 223–238.

		Number of components							
		1	2	**3**	4	5	6	7	8
$n = 50$	PRML	2	52	**35**	8	3			
	MHDE	80	20						
	NKE	44	56						
	MKE	44	53	**3**					
	RW	22	7	**59**	10	1	1		
	Bootstrap	0	96	**4**					
	Henna	25	68	**6**	1				
$n = 250$	PRML	15	19	**44**	19	3			
	MHDE	16	39	**45**					
	NKE	0	99	**1**					
	MKE	0	87	**11**	1	1			
	RW	0	0	**60**	22	18			
	Bootstrap	0	83	**16**	1				
	Henna	0	90	**10**					
$n = 500$	PRML	1	32	**55**	11	1			
	MHDE	0	35	**65**					
	NKE	0	97	**3**					
	MKE	0	58	**34**	6	2			
	RW	0	0	**22**	12	61	5		
	Bootstrap	0	74	**20**	6				
	Henna	0	85	**15**	1				
$n = 1000$	PRML	0	38	**45**	16	1			
	MHDE	0	26	**74**					
	NKE	0	86	**14**					
	MKE	0	18	**63**	10	2	3	1	3
	RW	0	0	**0**	1	89	10		
	Bootstrap	0	79	**15**	4	2			
	Henna	0	78	**15**	5	1	0	1	

Table 2: Summary of the 100 estimates of the mixture complexity in the simulation experiment in Section 4.3. The true mixture complexity is 3.

Martin, R. (2011), "Convergence rate for predictive recursion estimation of finite mixtures," *Submitted*. Preprint `arXiv:1106.4223`.

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," *Submitted*. Preprint `arXiv:1106.3885`.

— (2011b), "Semiparametric inference in mixture models with predictive recursion marginal likelihood," *Biometrika*, to appear. Preprint `arXiv:1106.3352`.

McLachlan, G. and Peel, D. (2000), *Finite mixture models*, Wiley-Interscience, New York.

McLachlan, G. J. (1987), "On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture," *J. Roy. Statist. Soc. Ser. C*, 36, 318–324.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture models*, vol. 84, New York: Marcel Dekker Inc., inference and applications to clustering.

Müller, P. and Quintana, F. A. (2004), "Nonparametric Bayesian data analysis," *Statist. Sci.*, 19, 95–110.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *J. Comput. Graph. Statist.*, 9, 249–265.

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.

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

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

Roeder, K. and Wasserman, L. (1997), "Practical Bayesian density estimation using mixtures of normals," *J. Amer. Statist. Assoc.*, 92, 894–902.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Chichester: John Wiley & Sons Ltd.

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

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

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