An approximate Bayesian marginal likelihood approach for estimating finite mixtures

Ryan Martin

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

February 13, 2012

Abstract

Estimation of finite mixture models when the mixing distribution support is unknown is an important problem. This paper gives a new approach based on a marginal likelihood for the unknown support. Motivated by a Bayesian Dirichlet prior model, a computationally efficient stochastic approximation version of the marginal likelihood is proposed and large-sample theory is presented. By restricting the support to a finite grid, a simulated annealing method is employed to maximize the marginal likelihood and estimate the support. Real and simulated data examples show that this novel stochastic approximation–simulated annealing procedure compares favorably to existing methods.

Keywords and phrases: Dirichlet distribution; mixture complexity; predictive recursion; simulated annealing; stochastic approximation.

1   Introduction

It is well-known that complicated data sets can be described by a mixture of a few relatively simple models. The catch, however, is that this finite mixing distribution is generally difficult to specify. For example, in clustering or empirical Bayes problems, the mixing distribution is exactly the quantity of interest. Therefore, estimating the unknown mixing distribution is an important problem. Key references include Titterington et al. (1985), McLachlan and Basford (1988), and McLachlan and Peel (2000).

To fix notation, we assume that data $Y_1, \dots, Y_n$ 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. The goal is to estimate the unknown finite support set U and the corresponding mixing weights $f = \{f(u) : u \in U\}$.

A classical approach to this problem is nonparametric maximum likelihood (Lindsay 1995), where the unknowns (f, U) are estimated by maximizing the likelihood function $\prod_{i=1}^{n} m_{f,U}(Y_i)$. The goal of maximum likelihood is simply to produce an estimate of $m_{f,U}$ that fits the data well, so there are no built-in concerns about the size of the estimated support. But as Priebe (1994) argues, there are compelling reasons to seek a well-fitting mixture with as few support points as possible. Since these two considerations—good fit and small support—are at odds with one another, a modification of maximum likelihood/minimum distance estimation is needed. The typical non-Bayesian strategy is to penalize the likelihood function of those models whose mixing distribution support is too large. This choice of penalty can be simple, like the AIC (Akaike 1973) or BIC (Schwarz 1978) penalties discussed by Leroux (1992), or can be more sophisticated, like the SCAD penalty (Fan and Li 2001) used in the mixture model context by Chen and Khalili (2008). Minimum distance methods with similar support size penalties have also been proposed, e.g., by James et al. (2001) and Woo and Sriram (2006, 2007). In the Bayesian context, there are a number of other methods. Ishwaran et al. (2001) give a solution based on Bayes factors, Richardson and Green (1997) presents a method in which each model is embedded into a larger parameter space on which a prior distribution is imposed, and Roeder and Wasserman (1997) describe a practical solution based on the posterior distribution of the number of mixture components. A typical nonparametric Bayesian 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 for fitting the Dirichlet process mixture model to data; see, e.g., Müller and Quintana (2004).

In this paper I introduce a new approach for fitting finite mixture models (1) with unknown support size. The starting point is the construction of a marginal likelihood for the mixing distribution support based on a Bayesian hierarchical model. For any fixed support set U, a computationally efficient approximation of the Bayesian marginal likelihood is available, based on a Robbins–Monro type of stochastic approximation algorithm called *predictive recursion*. But despite its efficiency, estimating U by maximizing this approximate marginal likelihood over *all possible U* is not feasible. The key idea is to chop up the bounding set $\overline{\mathscr{U}}$ into a large, suitably fine grid $\mathscr{U}$, and search for the best approximation to the underlying density m by mixtures supported on subsets of $\mathscr{U}$. Thus, the essentially nonparametric problem is transformed to a high-dimensional parametric one. The parameter U takes values in a very large finite set and for this high-dimensional combinatorial optimization problem, we propose a fast simulated annealing procedure. This novel combination of stochastic approximation and simulated annealing for finite mixture model estimation shall be called SASA.

Asymptotic convergence properties of the SASA estimates are presented in Section 3.3. Specifically, for given $\mathscr{U}$, the SASA procedure will asymptotically identify the best mixture over all those supported on subsets of $\mathscr{U}$. Here "best" is measured in terms of the Kullback–Leibler divergence, so that SASA acts asymptotically like a minimum distance estimation method. SASA also achieves the optimal rate of convergence obtained in Chen

(1995), although I suspect that the restriction to finitely many U's actually yields faster rates. Furthermore, unlike the estimates of James et al. (2001) or Woo and Sriram (2006), the SASA 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 $\mathscr{U}$.

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 is considerably smaller than $|\mathscr{U}|$. To account for these prior beliefs, I recommend a regularized version of the approximate marginal likelihood that penalizes supports U in $\mathscr{U}$ which are too large. In particular, I 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.

The SASA approach can, in principle, handle mixtures over any number of parameters, but computation can be relatively expensive for mixtures over two or more parameters. In Section 4, I modify the proposed algorithm to give a fast approximation to the SASA solution in 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.

2 Likelihood for the mixing distribution support

The following model selection reasoning is ubiquitous in statistics: calculate a "score" for each model in consideration and then pick the model with the best score. For example, the model score is often the maximized likelihood function over the model's parameters. But the case can be made that this maximized (profile) likelihood for the model essentially ignores the uncertainty in estimating the model parameters. In such cases, a marginal likelihood, with model parameters "integrated out" may be more reasonable, justifying a sort of Bayesian perspective. Along these lines, take the mixing distribution support set U to be fixed for the time being and consider the following hierarchical model:

$$Y_1, \ldots, Y_n \mid (f, U) \overset{\text{iid}}{\sim} m_{f,U}, \quad \text{and} \quad f \mid U \sim \mathsf{P}_U, \tag{2}$$

where $m_{f,U}$ is as in (1), and P_U is a generic prior for the random discrete distribution f supported on U. While other choices are possible, I will henceforth assume that P_U is 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. For this model, the marginal likelihood for U is of the form

$$L_n^\dagger(U) = \int \left\{ \prod_{i=1}^n m_{f,U}(Y_i) \right\} d\mathsf{P}_U(f) = \prod_{i=1}^n \sum_{u \in U} p(Y_i \mid u) \hat{f}_{i-1,U}(u), \tag{3}$$

where $\hat{f}_{i-1,U} = \mathsf{E}_U(f \mid Y_1, \ldots, Y_{i-1})$ is the posterior mean. The second equality in (3) is a consequence of Fubini's theorem and the linearity of the mixture. This likelihood can be computed efficiently, without explicitly evaluating the $\hat{f}_{i-1,U}$'s, via the sequential imputation method of Liu (1996). The function $L_n^\dagger(U)$ is a genuine likelihood for U in the sense that it defines a reasonable, data-dependent ranking of candidate support sets, properly accounting for the uncertainty in the mixing distribution weights f. So $L_n^\dagger$ can be used to assign a preference between two supports U and U', but I claim that it can also be used, in a natural way, to estimate the support, up to an approximation.

Suppose, first, that the unknown support set U is known to be contained in a compact set $\overline{\mathscr{U}}$. This is a very weak assumption that can always be justified in practice. Now chop up $\overline{\mathscr{U}}$ into a sufficiently fine finite grid $\mathscr{U}$ such that assuming $U \subseteq \mathscr{U}$ is no practical restriction, in the sense that the data-generating density $m(y)$ can be closely approximated by some mixture supported on a subset of $\mathscr{U}$. In the examples that follow, good solutions can be found when the grid $\mathscr{U}$ is of only moderate size. The advantage of this approximation is that the essentially nonparametric problem of estimating the unknown support becomes a very-high-dimensional parametric problem. In fact, there are only finitely many possible parameter values so theoretical convergence of estimators follows from simple point-wise convergence of (a normalized version of) the marginal likelihood. The drawback, however, is that maximizing the marginal likelihood over $\mathscr{U}$ is a relatively challenging combinatorial optimization problem. Although (2) is a reasonable model, it turns out that this fully Bayesian framework is not completely satisfactory from a computational point of view; see Example 1 below. Next I propose a second approximation which closely mimics the Bayesian results at a fraction of the computational cost.

3 The SASA method for finite mixtures

3.1 A stochastic approximation-based likelihood

First consider the general problem where the common marginal density $m(y)$ for $\mathscr{Y}$-valued observations $Y_1, \ldots, Y_n$ is modeled as a nonparametric mixture

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

where $\mathscr{U}$ is a known set, not necessarily finite, and $f \in \mathbb{F}$ is the unknown mixing density to be estimated. Here $\mathbb{F} = \mathbb{F}(\mathscr{U}, \nu)$ is the set of all densities with respect to a σ-finite Borel measure ν on $\mathscr{U}$. Newton (2002) presents the following *predictive recursion* algorithm for nonparametric estimation of f.

Predictive recursion algorithm. Fix $f_0 \in \mathbb{F}$ and a sequence of weights $\{w_i : i \geq 1\} \subset (0, 1)$. For $i = 1, \ldots, n$, compute $m_{i-1}(y) = m_{f_{i-1}}(y)$ as in (4) and

$$f_i(u) = (1 - w_i) f_{i-1}(u) + w_i \, p(Y_i \mid u) f_{i-1}(u) \, / \, m_{i-1}(Y_i). \tag{5}$$

Then return f_n and m_n as the final estimates of f and m, respectively.

Martin and Ghosh (2008) showed that $\{f_n\}$ is a Robbins–Monro stochastic approximation process. Key properties of predictive recursion include its fast computation and its ability to estimate a mixing density f absolutely continuous with respect to any user-defined dominating measure ν. That is, unlike the nonparametric maximum likelihood estimate which is almost surely discrete (Lindsay 1995, Theorem 21), f_n can be discrete, continuous, or both, depending on ν. Herein I shall take ν to be counting measure on a finite set $\mathscr{U}$, but see Martin and Tokdar (2011, 2012) for applications of predictive recursion where ν is continuous or both discrete and continuous.

Large-sample properties of f_n and m_n can be obtained under fairly mild conditions on the kernel $p(y \mid u)$ and the true data-generating density $m(y)$. Let $\mathbb{M}$ denote the

set of all mixtures m_f in (4) 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 almost sure L_1 and weak convergence of m_n and f_n, respectively, 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 in L_1 to $m_{f^\star}$. As a corollary, they show that if the mixture (4) is identifiable, then f_n converges weakly to $f^\star$ almost surely. Moreover, for a certain choice of weights $\{w_n\}$, they obtain a conservative $o(n^{-1/6})$ bound on the rate at which m_n converges to $m_{f^\star}$. The rate for f_n in the general case is unknown, but Martin (2012) obtains a near parametric $n^{-1/2}$ rate for f_n in the finite mixture case; see also Section 3.3 below.

Define a stochastic approximation-based marginal likelihood

$$L_n(U) = \prod_{i=1}^{n} m_{i-1,U}(Y_i) = \prod_{i=1}^{n} \sum_{u \in U} p(Y_i \mid u) f_{i-1}(u). \tag{6}$$

This is based on an interpretation of $m_{i-1,U}(Y_i)$ as the conditional density of Y_i given $Y_1, \ldots, Y_{i-1}$. I claim that $L_n(U)$ is an approximation of Dirichlet prior Bayes marginal likelihood $L_n^\dagger(U)$. Towards this, recall that $\hat{f}_{k,U} = \mathsf{E}_U(f \mid Y_1, \ldots, Y_k)$ is the posterior mean of the mixing distribution on fixed U, given $Y_1, \ldots, Y_k$. Then $\sum_{u \in U} p(Y_i \mid u)\hat{f}_{i-1,U}(u)$ is the conditional density of Y_i given $Y_1, \ldots, Y_{i-1}$ based on the Dirichlet hierarchical model. Also 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 on U given Y_1. If $\alpha_0 = 1/w_1 - 1$, then $\hat{f}_{1,U}(u)$ is exactly $f_1(u)$ in (5). This correspondence holds exactly only for a single observation, but Martin and Tokdar (2011) argue that $f_{i-1,U}$, for any i, acts as a dynamic, mean-preserving filter approximation to the Bayes estimate $\hat{f}_{i-1,U}$. Then $L_n(U)$ in (6) can be viewed as a plug-in approximation of the Bayes marginal likelihood $L_n^\dagger(U)$ in (3), with $f_{i-1,U}$ in place of the Bayes estimate $\hat{f}_{i-1,U}$. See, also, Example 1 below.

In what follows, I will work with the marginal likelihood on the log-scale,

$$\ell_n(U) = \log L_n(U) = \sum_{i=1}^{n} \log\Big\{\sum_{u \in U} p(Y_i \mid u) f_{i-1}(u)\Big\}. \tag{7}$$

The goal is to estimate U by maximizing $\ell_n(U)$ over U. Restricting U to be a subset of the finite set $\mathscr{U}$ is a helpful first step, but even when $|\mathscr{U}|$ is only moderately large, the set of possible supports is still enormous, cardinality $2^{|\mathscr{U}|} - 1$. So despite the fact that the search space is finite, its size makes this is challenging problem. In Section 3.2, I give a simulated annealing algorithm to solve this combinatorial optimization problem.

3.2 Optimization with simulated annealing

As described above, maximizing $\ell_n(U)$ over all subsets $U \in 2^{\mathscr{U}}$ is a combinatorial optimization problem. The challenge is that $2^{\mathscr{U}}$ is so large that it is not feasible to evaluate

$\ell_n(U)$ for each U. Simulated annealing is a stochastic algorithm where, at 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 $\ell_n(U^{(t)})$. An important feature of simulated annealing is the decreasing temperature sequence $\{\tau_t : t \geq 0\}$. Following Hajek (1988) and Bélisle (1992), I 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 $U \subset \mathscr{U} = \{u_1, \ldots, u_S\}$, where $S = |\mathscr{U}|$, associate a binary S-vector $H \in \{0,1\}^S$. 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.

Simulated annealing algorithm. Choose a starting point $H^{(0)}$ and a maximum number of iterations T. Set $t = 1$ and generate a sequence $\{H^{(t)} : t = 1, \ldots, T\}$ as follows:

1. Simulate H_{new} from a probability distribution $\pi^{(t)}$ on $\{0,1\}^S$, possibly depending on t and the current iterate $H^{(t)}$.

2. Define the acceptance probability

$$\alpha(t) = 1 \wedge \exp\big[\{\ell_n(H_{\text{new}}) - \ell_n(H^{(t)})\}/\tau_t\big],$$

where $\ell_n(H)$ is the PR marginal likelihood defined in (6), written as a function of the indicator H that characterizes U, and set

$$H^{(t+1)} = \begin{cases} H_{\text{new}} & \text{with probability } \alpha(t) \\ H^{(t)} & \text{with probability } 1 - \alpha(t) \end{cases}$$

3. If $t < T$, set $t \leftarrow t + 1$ and return to Step 1; else, exit the loop.

Then return the visited $H^{(t)}$ with the largest log-likelihood $\ell_n(H^{(t)})$.

Herein, the initial choice is $H_s^{(0)} = 1$ for each s, which corresponds to the full mixture. 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. 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 a proposal distribution $\pi^{(t)}$ in Step 1. I shall assume that 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, I 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

$$\pi_s^{(t)} \propto 1 + \Big(\frac{S}{\sum_{s=1}^{S} H_s^{(t)}}\Big)^r \cdot H_s^{(t)}, \quad s = 1, \ldots, S, \quad r \geq 1. \tag{8}$$

Note 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.

Next I discuss a few miscellaneous computational details.

- The log marginal likelihood $\ell_n(U)$ depends on the order in which the data $Y_1, \ldots, Y_n$ are processed. To reduce this dependence, I take $\ell_n(U)$ to be the average of the log marginal likelihoods over several random data permutations. The speed of predictive recursion for fixed U makes this permutation-averaging computationally feasible. Herein I use 100 permutations but, in my experience, the SASA estimates are relatively stable for as few as 25 permutations. These permutations are chosen once and kept fixed throughout optimization process.

- To avoid various degeneracies, specifically in (8), I set $\ell_n(\varnothing) = -\infty$. This means that if a move to an empty mixing distribution support is proposed, then it will surely be rejected since the corresponding $\alpha(t)$ would be zero.

- For all examples, the "distance" k between two consecutive support sets, is taken to be 1. Also, choosing $r = 1$ in (8) works well.

- For all examples, I run simulated annealing for $T = 5000$ iterations. As with the choice of permutations, my choice here is rather conservative, as the estimates are often relatively stable for $T = 2000$.

R codes to implement the SASA procedure are available at my website `www.math.uic.edu/~rgmartin/research.html`. The R function `optim`, with the option `method="SANN"`, is the driving force behind the simulated annealing. A C subroutine used to efficiently evaluate the log marginal likelihood $\ell_n(U)$ for any fixed U.

3.3 Large-sample theory

Suppose that $\mathscr{U}$ is a fixed finite set and $U \subseteq \mathscr{U}$ is any subset. Assume that the predictive recursion weights $\{w_i : i \geq 1\}$ are given by $w_i = (i+1)^{-\gamma}$ for some $\gamma \in (0.5, 1)$. With U fixed, convergence of $f_{n,U}$ at a $n^{-(1-1/2\gamma)}$ rate is established in Martin (2012). This result holds only for fixed U, while the present focus is on the case of unknown U. Towards this, consider a normalized version of the log marginal likelihood $\ell_n(U)$, namely

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

Also define $K^\star(U) = \inf\{K(m, m_{f,U}) : f \in \mathbb{F}\}$, where $\mathbb{F} = \mathbb{F}_U$ is the probability simplex in $\mathbb{R}^{|U|}$. This is the smallest Kullback–Leibler divergence of a mixture in the class $m_{f,U}$ from m. Since $K(m, m_{n,U}) \to K^\star(U)$ for each U and $K_n(U)$ is in some sense similar to $K(m, m_{n,U})$, one might expect that $K_n(U)$ also converges to $K^\star(U)$. This will imply that maximizing the likelihood $\ell_n(U)$ in (6) to estimate U is a reasonable strategy. Indeed, Martin (2012) proves that $K_n(U) \to K^\star(U)$ almost surely, as $n \to \infty$. Furthermore, since the collection of all U's is finite, the convergence is uniform, i.e., $K_n(\widehat{U}_n) \to K^\star(U^\star)$, where $\widehat{U}_n$ is the maximizer of $\ell_n(U)$ and $U^\star$ is the minimizer of $K^\star(U)$. It follows that $\widehat{U}_n \to U^\star$ in the sense that, eventually, both sets will have the same elements. Furthermore, Theorem 3 in Martin (2012) implies that $f_{n,\widehat{U}_n}$ converges at a nearly $n^{-1/4}$ rate.

Three remarks about this result are in order. First, I had originally motivated SASA as a computationally efficient approximation to a Bayesian marginal likelihood procedure. The convergence results above give SASA a secondary interpretation as a minimum distance method, not unlike those of James et al. (2001) and Woo and Sriram (2006). Second,

recall that Chen (1995) shows that the optimal rate for estimating finite mixing distributions with unknown support is $n^{-1/4}$, nearly matched by the rate for $f_{n,\widehat{U}_n}$. However, I expect that this can be improved to $n^{-1/2}$, although the proof eludes me. The difference comes from the fact that the essentially nonparametric problem of estimating the finite mixing distribution support is, here, first reduced to a very-high- but ultimately finite-dimensional parametric subproblem. In fact, the rate $n^{-1/2}$ is available for the Bayesian marginal likelihood version. Third, regarding estimation of the mixture complexity $|U|$, the results here differ considerably from those in, say, James et al. (2001) and Woo and Sriram (2006). In particular, once the "parameter space" $\mathscr{U}$ is specified, the SASA estimate of the support size is bounded by $|\mathscr{U}|$, whether the model is correctly specified or not, and is guaranteed to converge. In contrast, the James et al. (2001) and Woo and Sriram (2006) estimates of the mixture complexity explode to infinity in the misspecified case. I believe that, in the misspecified case, SASA's asymptotic identification of the best finite mixture in a sufficiently large class is more meaningful. That is, one would arguably prefer the estimates to converge to the closest approximation of $m(y)$ within the postulated class of finite mixtures.

3.4   Regularized SASA

In the hierarchical model (2), it would be natural to introduce a prior for U to complete the hierarchy. Martin and Tokdar (2012) propose a regularized version of the the approximate marginal likelihood in which priors for structural parameters are incorporated into the model, effectively replacing the marginal likelihood with a marginal posterior.

Here, a prior for U should 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—it is reasonable to penalize those U with too many components. To accomplish this, I recommend a prior for U consisting of a binomial prior for the size $|U|$ and a conditionally uniform prior on U, given its size. 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 elicit an *a priori* reasonable expected number of components, say 5, and then one may choose $\rho = 5/S$. In the absence of such information, ρ can be chosen by first estimating m with some standard density estimate $\hat{m}$ and taking $\rho = (\text{number of modes of } \hat{m})/S$.

3.5   Numerical results

For the simple univariate mixture problem, typical kernels are Gaussian with fixed scale and Poisson. Example 1 gives a quick comparison, in the context of Gaussian location mixtures, of SASA with the Bayesian method it is meant to approximate. Example 2 gives the details of a large-scale simulation study for Poisson mixtures. Location-scale mixtures (of Gaussians) will be the focus of the next section.

Example 1. Consider the galaxy data set from Roeder (1990). This one consists of velocity measurements for $n = 82$ galaxies. Based on the *a priori* considerations of Escobar and West (1995), it is reasonable to model these data as a location mixture of Gaussians with common scale $\sigma = 1$. The results for SASA, presented in Figure 1 of Martin

Model	$(u_1, f(u_1))$	$(u_2, f(u_2))$	$(u_3, f(u_3))$	$(u_4, f(u_4))$
1	(1, .50)	(9, .50)		
2	(1, .80)	(9, .20)		
3	(1, .95)	(10, .05)		
4	(1, .45)	(5, .45)	(10, .10)	
5	(1, .33)	(5, .33)	(10, .34)	
6	(1, .30)	(5, .40)	(9, .25)	(15, .05)
7	(1, .25)	(5, .25)	(10, .25)	(15, .25)

Table 1: Parameters for the Poisson mixture simulations in Example 2.

(2012), using $\mathscr{U} = \{5.0, 5.5, \ldots, 39.5, 40.0\}$, are obtained in roughly 3.5 seconds. The fully Bayes version, using Jun Liu's sequential imputation algorithm to approximate the marginal likelihood $L_n^\dagger(U)$ in (3), gives identical results but takes more than 30 seconds. Therefore, the approximate marginal likelihood $L_n(U)$ in (6) is reasonable.

Example 2. Here I consider a simulation study presented in Chen and Khalili (2008) for Poisson mixtures. There are seven models under consideration, and their respective parameters are listed in Table 1. The resulting estimates of the mixture support size for a host of methods over 500 random samples each of size $n = 100, 500$ are summarized in Table 2. For the SASA implementation, I take $\overline{\mathscr{U}} = [0, 20]$ and $\mathscr{U}$ a grid of $S = 101$ equispaced points. The SASA regularization parameter is taken as $\rho = 15/S$. For the most complicated Model 7, with $n = 500$, the SASA estimate took about 15 seconds to compute, on average. In addition to SASA, the methods compared are those based on AIC, BIC, and likelihood ratio test (LRT) model selection procedures, two minimum Hellinger distance (HD) methods from Woo and Sriram (2007), and the mixture SCAD method of Chen and Khalili (2008). The most striking observation is in Models 2–3 with $n = 500$. Note that existing methods estimate the support as a single point, while SASA is able to correctly identify two support points in a large proportion of the runs. In Models 6–7, where the number of support points is relatively large, SASA tends to underestimate the support size. But, in these examples, only MSCAD is successful, and SASA's performance better than BIC and the Woo–Sriram estimates.

4 SASA for location-scale mixtures

4.1 Setup and modified algorithm

In principle, the SASA procedure is able to handle any type of finite mixture. However, when $\mathscr{U}$ is a lattice in a higher-dimensional space, the computations become 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. My 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. The simplification is to rule out such cases, allowing at

			n = 100					n = 500				
Model	$|U|$	Method	1	2	3	4	5	1	2	3	4	5
1	2	AIC		.938	.062				.942	.076		
		BIC		.998	.002				.998	.002		
		HD$_{2/n}$		1.00					1.00			
		HD$_{\log n/n}$		1.00					1.00			
		LRT		.950	.050				.960	.040		
		MSCAD		.988	.012				1.00			
		SASA		.982	.028				.998	.002		
2	2	AIC		.958	.042			.950	.042	.008		
		BIC		.994	.006			1.00				
		HD$_{2/n}$		.998	.002			1.00				
		HD$_{\log n/n}$	.002	.998				1.00				
		LRT		.950	.050			.960	.040			
		MSCAD	.002	.986	.012			.990	.010			
		SASA		.988	.012				.972	.028		
3	2	AIC	.012	.948	.036			.950	.048	.002		
		BIC	.026	.972	.002			.998	.002			
		HD$_{2/n}$	.616	.384				1.00				
		HD$_{\log n/n}$	.946	.054				.994	.006			
		LRT		.930	.070			.950	.050			
		MSCAD	.052	.868	.080			.994	.004			
		SASA	.024	.974	.002				.932	.068		
4	3	AIC		.410	.590				.006	.972	.022	
		BIC		.778	.222				.100	.900		
		HD$_{2/n}$		.966	.034				.162	.838		
		HD$_{\log n/2}$		1.00					.846	.154		
		LRT		.390	.580	.020			.940	.060		
		MSCAD		.280	.692	.028			.082	.896	.022	
		SASA		.670	.330				.040	.958	.002	
5	3	AIC		.274	.720	.006				.974	.026	
		BIC		.684	.316				.026	.974		
		HD$_{2/n}$		.840	.160				.018	.982		
		HD$_{\log n/n}$		.988	.012				.462	.538		
		LRT		.300	.660	.030				.940	.060	
		MSCAD		.200	.780	.020			.016	.964	.020	
		SASA		.436	.554	.010			.010	.904	.086	
6	4	AIC		.080	.878	.042				.644	.356	
		BIC		.316	.680	.004			.974	.026		
		HD$_{2/n}$		.718	.282				.956	.044		
		HD$_{\log n/n}$		.962	.038			.060	.940			
		LRT		.090	.780	.130			.590	.380	.030	
		MSCAD		.010	.666	.320	.004		.366	.624	.010	
		SASA		.194	.804	.002			.892	.108		
7	4	AIC		.010	.918	.072				.592	.408	
		BIC		.134	.858	.008			.970	.030		
		HD$_{2/n}$		.182	.812	.006			.924	.076		
		HD$_{\log n/n}$		.718	.282			.060	1.00			
		LRT		.020	.860	.120			.590	.400	.010	
		MSCAD			.512	.460	.028		.110	.812	.078	
		SASA		.062	.914	.024			.888	.112		

Table 2: Results of the Poisson mixture simulations in Example 2. Values are the proportion of estimates of the given size in 500 samples. All but the SASA results are taken from Chen and Khalili (2008, Tables 6–8)

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 I develop this modification for the important special case of location-scale mixtures.

Let $\overline{\mathscr{U}}_1$ and $\overline{\mathscr{U}}_2$ be closed intervals in $\mathbb{R}$ and $\mathbb{R}^+$, respectively, assumed to contain the range of values the location μ and scale σ can take. Chop up these intervals into sufficiently fine grids $\mathscr{U}_1$ and $\mathscr{U}_2$ of sizes $S_1 = |\mathscr{U}_1|$ and $S_2 = |\mathscr{U}_2|$, respectively. Define the rectangle $\overline{\mathscr{U}} = \overline{\mathscr{U}}_1 \times \overline{\mathscr{U}}_2$ and the two-dimensional lattice $\mathscr{U} = \mathscr{U}_1 \times \mathscr{U}_2$. Then the finite mixture model is just as before

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

where the kernel $p(y \mid \mu, \sigma)$ equals $\sigma^{-1} p\big((y-\mu)/\sigma\big)$ for some symmetric unimodal density function p. 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. Here I will focus on the Gaussian case only.

What makes this different from before is that U can contain at most one (μ, σ) pair on each vertical line $\mathscr{U}_1 \times \mathscr{U}_2$. To accommodate this restriction, I shall modify the simulated annealing algorithm proposed in Section 3.2. The key idea 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 I use the notation $H = (H_1, \ldots, H_{S_1})$, where each H_s takes values in $\{0, 1, \ldots, S_2\}$ to characterize the support set U. The interpretation is

$$H_s = \begin{cases} 0 & \text{if } \mu_s \text{ is not in the mixture} \\ h & \text{if pair } (\mu_s, \sigma_h) \text{ is in the mixture, } h = 1, \ldots, S_2. \end{cases} \tag{9}$$

In other words, location μ_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, I can formulate the SASA optimization problem in terms of H. By restricting the estimates to this collection of admissible subsets, the state space to search goes from $2^{S_1 \times S_2}$ down to $(S_2 + 1)^{S_1}$, which can be a drastic reduction. To maximize the approximate log marginal likelihood $\ell_n(H)$ over the set of all admissible H, I 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 $\beta(t) = S_1^{-1} \sum_{s=1}^{S_1} I\{H_s^{(t)} = 0\}$, the proportion of zero entries in $H^{(t)}$. Now sample an entry in $H^{(t)}$ with probabilities

$$\pi_s^{(t)} \propto 1 + \big(1 - \beta(t)\big)^{-r} \cdot I\{H_s^{(t)} > 0\}, \quad s = 1, \ldots, S(\theta). \tag{10}$$

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 to consider:

- If $H_s^{(t)} = 0$, take $H_{\text{new}} \sim \mathsf{Unif}\{1, \ldots, S_2\}$.

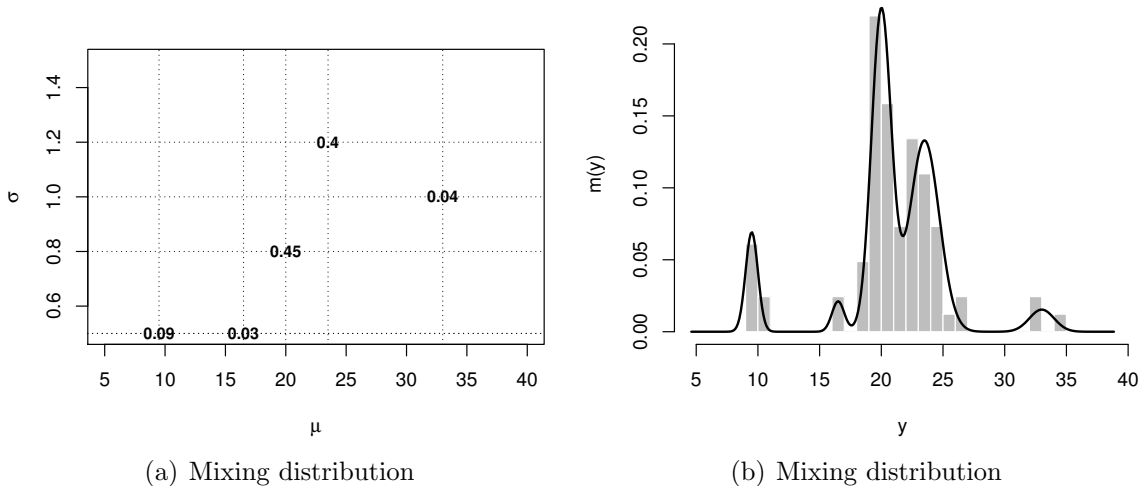(a) Mixing distribution   (b) Mixing distribution

Figure 1: Plot of the SASA estimates of the location-scale mixture density for the galactic velocity data in Example 3. In panel (a), the numbers are the mixing weights.

- If $0 < H_s^{(t)} < S_2$, take $H_{\text{new}} = 0$ with probability $\beta(t)$ and

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

If $H_s^{(t)} = 1$ or S_2, then H_{new} would be 2 or $S_2 - 1$, 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 $\mathscr{U}_1$ should remain in the mixture as t increases.

Once H_{new} has been sampled, the simulated annealing algorithm decides to take $H^{(t+1)}$ as H_{new} or $H^{(t)}$ depending on the flip of the $\alpha(t)$-coin as in Step 2 in Section 3.2. As before, if H_{new} is a better candidate support than $H^{(t)}$ 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 Numerical results

Example 3. Here I use SASA to estimate a Gaussian location-scale mixture for the galaxy data in Example 1. Take $\mathscr{U}_1 = \{5.0, 5.5, \ldots, 39.5, 40.0\}$ and $\mathscr{U}_2 = \{0.5, 0.6, \ldots, 1.4, 1.5\}$. SASA estimates five Gaussian components with varying scales, and Figure 1 shows the resulting estimate of the density. In this case, the overall fit is good—similar to that in Ishwaran et al. (2001) and elsewhere—but only five components are needed compared to six in Example 1 in Martin (2012). Here the computation took roughly 6 seconds.

Example 4. Next I present a simulation experiment in which the focus is 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

$$m(y) = 0.25\mathsf{N}(y \mid -0.3, 0.05) + 0.5\mathsf{N}(y \mid 0, 10) + 0.25\mathsf{N}(y \mid 0.3, 0.05). \tag{11}$$

	Method	Number of components

	Method	1	2	3	4	5	6	7	8
$n = 50$	WS	80	20						
	JPM	44	53	3					
	RW	22	7	59	10	1	1		
	SASA		23	44	25	8			
$n = 250$	WS	16	39	45					
	JPM		87	11	1	1			
	RW			60	22	18			
	SASA	15	28	47	17	1			
$n = 500$	WS		35	65					
	JPM		58	34	6	2			
	RW			22	12	61	5		
	SASA		17	48	32	2	1		
$n = 1000$	WS		26	74					
	JPM		18	63	10	2	3	1	3
	RW				1	89	10		
	SASA		10	50	24	13	3		

Table 3: Summary of the 100 estimates of $|U|$ in Example 4. The true mixture complexity is 3. All but the SASA results are taken from Woo and Sriram (2006, Table 1).

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 I investigate the performance of the approximate SASA procedure outlined in Section 4.1.

Table 3 summarizes the SASA estimates of the mixture complexity based on 100 random samples from the mixture model $m(y)$ in (11) with four different sample sizes: $n = 50$, 250, 500, and 1000. In particular, I take $\overline{\mathscr{U}}_1 = [-2, 2]$, $\overline{\mathscr{U}}_2 = [0.1, 4.0]$ and $\mathscr{U}_1$ and $\mathscr{U}_2$ are equispaced grids of length $S_1 = 40$ and $S_2 = 25$, respectively. Note that the true location and scale parameters in (11) do not belong to $\mathscr{U}_1 \times \mathscr{U}_2$. The simulated annealing optimization procedure in Section 4.1 is used to optimize the approximate 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 $\mathscr{U}_1 \times \mathscr{U}_2$, 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. The average computation time for SASA ranges from 4.5 seconds for $n = 50$ and 52 seconds for $n = 1000$. For comparison, I also include minimum distance estimates of Woo and Sriram (2006) and James et al. (2001), and the Bayesian estimates of Roeder and Wasserman (1997); these shall be denoted by *WS*, *JPM*, and *RW*, respectively. The RW method performs well for small n but seems to falter as n increases, while the JPM method does well for large n. SASA 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 WS method is as good or better than SASA at correctly identifying the true mixture complexity across simulations.

5   Discussion

This paper presents a novel hybrid stochastic approximation–simulated annealing algorithm for estimating finite mixtures. The method is based, first, on a marginal likelihood function for the support based on a Bayesian hierarchical model. Then two approximations are introduced to estimate the unknown support U—the first is an approximation of the bounding set $\overline{\mathscr{U}} \supseteq U$ by a finite grid $\mathscr{U}$, and the second is an efficient approximation of the marginal likelihood. Then a simulated annealing algorithm is used to maximize this approximate marginal likelihood over the finite set of candidate U's. There may be some theoretical benefits, in terms of rates of convergence, to approximating $\overline{\mathscr{U}}$ with the finite set $\mathscr{U}$, but the details remain to be worked out. Examples in both the Poisson and Gaussian case indicate that SASA is competitive with existing methods. In my experience, SASA is generally a bit more expensive computationally than the minimum distance methods of Woo and Sriram (2006) or James et al. (2001). But, on the other hand, it tends to be faster than the Bayesian methods it is meant to approximate.

In applications, the initial choice of $\mathscr{U}$ and, in particular, $|\mathscr{U}|$ is not obvious. In practice, one should make this choice based on the shape/spread of the data and the chosen kernel; this was the approach taken in Examples 1 and 3. 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 $\mathscr{U}$ 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 this, bounds on the constants associated with the rate in Martin (2012) would be needed, since these would likely depend on $|\mathscr{U}|$.

An explicit penalty on the size of the mixing distribution support was introduced in Section 3.4. And the location-scale adjustment to SASA's simulated annealing proposal can also be viewed as an implicit penalty on U. An anonymous reviewer pointed out the potential for incorporating even more sophisticated penalties in the approximate marginal likelihood for U. For example, one could further penalize candidate supports that contain points which are too close in some sense. This extreme regularization was not necessary in the examples considered here, but if the grid $\mathscr{U}$ is very fine, then the the closeness of nearby support points may become a concern.

In some cases, one might want to consider, say, a Gaussian location mixture with fixed but unknown scale σ. It is straightforward to implement an intermediate step in the algorithm in Section 3.2 whereby one replaces the joint marginal likelihood $\ell_n(U, \sigma)$ by a profile version $\ell_n(U, \hat{\sigma})$. In my experience, this was actually more expensive computationally than the location-scale approach, so I did not pursue this direction.

Acknowledgments

References

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, Budapest: Akadémiai Kiadó, pp. 267–281.

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

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.

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

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, 96, 1348–1360.

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.

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.

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

Leroux, B. G. (1992), "Consistent estimation of a mixing distribution," *Ann. Statist.*, 20, 1350–1360.

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

Liu, J. S. (1996), "Nonparametric hierarchical Bayes via sequential imputations," *Ann. Statist.*, 24, 911–930.

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

Martin, R. (2012), "Convergence rate for predictive recursion estimation of finite mixtures," *Statist. Probab. Lett.*, 82, 378–384.

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.

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

— (2012), "A nonparametric empirical Bayes framework for large-scale multiple testing," *Biostatistics*, to appear. Preprint at `arXiv:1106.3885`.

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

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.

Priebe, C. E. (1994), "Adaptive mixtures," *J. Amer. Statist. Assoc.*, 89, 796–806.

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.

Schwarz, G. (1978), "Estimating the dimension of a model," *Ann. Statist.*, 6, 461–464.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, 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.

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

— (2007), "Robust estimation of mixture complexity for count data," *Comput. Statist. Data Anal.*, 51, 4379–4392.