Unsupervised Selection and Estimation of Finite Mixture Models

Mário A. T. Figueiredo
Instituto de Telecomunicações
Instituto Superior Técnico
1049-001 Lisboa, PORTUGAL
E-mail: mtf@lx.it.pt

Anil K. Jain
Dept. of Computer Science and Eng.
Michigan State University
East Lansing, MI 48824, USA
E-mail: jain@cse.msu.edu

Abstract

We describe a new method for fitting mixture models to multivariate data which performs component selection and does not require external initialization. The novelty of our approach includes: an MML-like (minimum message length) model selection criterion; inclusion of the criterion into the expectation-maximization (EM) algorithm (increasing its ability to escape from local maxima); an initialization strategy supported on the interpretation of EM as a self-annealing algorithm.

1. Introduction

1.1. Finite Mixtures and EM

Finite mixtures (FM) are a flexible and powerful modeling tool. In pattern recognition, mixtures underlie formal approaches to unsupervised learning (clustering) [1]. FM are also able to approximate arbitrary probability density functions (pdf's); this makes them well suited for modeling complex class-conditional pdf's in supervised learning [4].

Consider n i.i.d. samples of a (k-component) FM, $\mathbf{y} = \{\mathbf{y}^{(1)}, ..., \mathbf{y}^{(n)}\}$. The log-likelihood function is

$$L\left(\boldsymbol{\theta}_{(k)}, \mathbf{y}\right) = \log \prod_{i=1}^{n} \underbrace{\sum_{m=1}^{k} \alpha_m \overbrace{p(\mathbf{y}^{(i)}|\boldsymbol{\theta}_m)}^{\text{components}}}_{\text{mixture } p(\mathbf{y}^{(i)}|\boldsymbol{\theta}_{(k)})},$$

where $\alpha_1, ..., \alpha_k$ are the *mixing probabilities*, and $\boldsymbol{\theta}_{(k)} \equiv \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k, \alpha_1, ..., \alpha_{k-1}\}$; notice that $\alpha_k = 1 - \sum_{m=1}^{k-1} \alpha_m$.

The *maximum likelihood* (ML) estimate, defined as $\widehat{\boldsymbol{\theta}}_{(k)} = \arg\max_{\boldsymbol{\theta}_{(k)}} L(\boldsymbol{\theta}_{(k)}, \mathbf{y})$, can not be found analytically. The same is true for the Bayesian MAP estimate, $\widehat{\boldsymbol{\theta}}_{(k)} = \arg\max_{\boldsymbol{\theta}_{(k)}} [L(\boldsymbol{\theta}_{(k)}, \mathbf{y}) + \log p(\boldsymbol{\theta}_{(k)})]$, given some prior $p(\boldsymbol{\theta}_{(k)})$. The standard alternative is the EM algorithm which, under mild conditions, converges to a local maximum of $L(\boldsymbol{\theta}_{(k)}, \mathbf{y})$ or $[L(\boldsymbol{\theta}_{(k)}, \mathbf{y}) + \log p(\boldsymbol{\theta}_{(k)})]$ [2].

EM is supported on the interpretation of $\mathbf{y}$ as *incomplete* data [2]. The *missing* part is a set of labels $\mathbf{z} = \{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(n)}\}$, where $\mathbf{z}^{(i)} = [z_1^{(i)}, ..., z_k^{(i)}]$, with $z_m^{(i)} = 1$ and $z_p^{(i)} = 0$, for $p \neq m$, meaning that $\mathbf{y}^{(i)}$ is a sample of $p(\mathbf{y}^{(i)}|\boldsymbol{\theta}_m)$. The (complete) log-likelihood (*i.e.*, if *complete* data $\mathbf{x} = \{\mathbf{y}, \mathbf{z}\}$ was observed) is [2]

$$L_c\left(\boldsymbol{\theta}_{(k)}, \mathbf{y}, \mathbf{z}\right) = \sum_{i=1}^{n} \sum_{m=1}^{k} z_m^{(i)} \log \left[\alpha_m p(\mathbf{y}^{(i)}|\boldsymbol{\theta}_m)\right].$$

The EM algorithm proceeds by alternatingly applying two steps (until some convergence criterion is met):

- **E-step:** Computes the conditional expectation of L_c, given $\mathbf{y}$ and $\widehat{\boldsymbol{\theta}}_{(k)}^{(t)}$ (the current parameter estimate): $E[L_c\left(\boldsymbol{\theta}_{(k)}, \mathbf{y}, \mathbf{z}\right)|\mathbf{y}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}] \equiv Q(\boldsymbol{\theta}_{(k)}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)})$. Since L_c is linear in the missing $z_m^{(i)}$'s, this step reduces to the computation of their conditional expectations. Moreover, because they are binary, $E[z_m^{(i)}|\cdot] = \Pr[z_m^{(i)} = 1|\cdot]$; then,

$$E\left[z_m^{(i)}|\mathbf{y}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}\right] = \frac{\widehat{\alpha}_m^{(t)} p(\mathbf{y}^{(i)}|\widehat{\boldsymbol{\theta}}_m^{(t)})}{\sum_{j=1}^{k} \widehat{\alpha}_j^{(t)} p(\mathbf{y}^{(i)}|\widehat{\boldsymbol{\theta}}_m^{(t)})} \equiv w_m^{(i,t)}. \quad (1)$$

- **M-step:** Updates the parameter estimates according to

$$\widehat{\boldsymbol{\theta}}_{(k)}^{(t+1)} = \arg\max_{\boldsymbol{\theta}_{(k)}} \{Q(\boldsymbol{\theta}_{(k)}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}) + \log p(\boldsymbol{\theta}_{(k)})\}. \quad (2)$$

If we are looking for ML estimates, rather than MAP, $\log p(\boldsymbol{\theta}_{(k)})$ is flat and is removed from Eq. (2).

1.2. Model Selection for Finite Mixtures

Model selection (*i.e.*, choosing the *optimal* number of components) is a central question in FM fitting. Most approaches to model selection for FM obtain a set of candidate models (usually by EM), for a range of values of k, and then select one according to

$$\widehat{k} = \arg\min_{k} \{\mathcal{C}(\widehat{\boldsymbol{\theta}}_{(k)}, k), k = 1, ..., k_{\max}\}, \quad (3)$$

where $\mathcal{C}(\widehat{\boldsymbol{\theta}}_{(k)}, k)$ is some cost function. Several of these methods (see [5, 6, 7]) have good model selection performance, but a major drawback remains: a whole set of $k_{\max}$ candidates has to be obtained, and well-known problems associated with EM emerge. **(a)** EM is highly dependent on initialization; a common (time-consuming) solution uses several random starts, and then chooses the highest likelihood estimate [2, 4, 6]; other schemes initialize the $w_m^{(i,t)}$ variables using clustering methods [2, 4]. Smarter methods based on merge [5], or split and merge [8], operations were recently proposed. **(b)** EM may converge to the boundary of the parameter space, *i.e.*, one of the α_m's goes to zero and the corresponding component becomes singular (the likelihood is unbounded); when k is larger than the optimal/true value, this may happen frequently.

Of course there is also the fully Bayesian alternative, via MCMC, which does not suffer from these drawbacks [9]; however, despite their formal appeal, we think that MCMC-based techniques are still far too computationally demanding to be useful in pattern recognition applications.

2. Proposed Approach

2.1. The Criterion

The *minimum description length* (MDL [10]) and *minimum message length* (MML [7, 11]) are two well known criteria which have been used for FM model selection [6, 7] in the form of Eq. (3), thus having the drawbacks mentioned above. We propose a new approach: a selection criterion that can be embedded in the EM algorithm, leading to an integrated model selection and estimation procedure.

Consider a prior $p(\boldsymbol{\theta}_{(k)}, k) = p(\boldsymbol{\theta}_{(k)})\, p(k)$, where $p(\boldsymbol{\theta}_{(k)})$ is short for $p(\boldsymbol{\theta}_{(k)}|k)$. With $p(k) = 1/k_{\max}$, for $k_{\max}$ known to be larger than the true k, let the simultaneous selection of k and estimation of $\boldsymbol{\theta}_{(k)}$, denoted $\widehat{\boldsymbol{\theta}_{(k)}}$, be

$$\widehat{\boldsymbol{\theta}_{(k)}} = \arg\min_{k,\,\boldsymbol{\theta}_{(k)}} \left\{ \frac{\log|\mathbf{I}(\boldsymbol{\theta}_{(k)})|}{2} - L(\boldsymbol{\theta}_{(k)}, \mathbf{y}) - \log p(\boldsymbol{\theta}_{(k)}) \right\}, \tag{4}$$

where $\mathbf{I}(\boldsymbol{\theta}_{(k)}) \equiv E[-\nabla^2_{\boldsymbol{\theta}_{(k)}} L(\boldsymbol{\theta}_{(k)}, \mathbf{y})]$ is the (expected) Fisher information matrix, and $|\mathbf{I}(\boldsymbol{\theta}_{(k)})|$ its determinant. This is an MML criterion (as used, *e.g.*, in [7]), the only difference being that we ignore the *optimal quantizing lattice* constants, as is done in MDL [10].

Since $\mathbf{I}(\boldsymbol{\theta}_{(k)})$ can not, in general, be obtained analytically, we replace it by the complete-data Fisher information matrix $\mathbf{I}_c(\boldsymbol{\theta}_{(k)}) \equiv E[-\nabla^2_{\boldsymbol{\theta}_{(k)}} L_c(\boldsymbol{\theta}_{(k)}, \mathbf{y}, \mathbf{z})]$, which upper-bounds[1] $\mathbf{I}(\boldsymbol{\theta}_{(k)})$. This matrix has block-diagonal structure,

$$\mathbf{I}_c(\boldsymbol{\theta}_{(k)}) = n \text{ block-diag}\{\alpha_1 \mathbf{I}(\boldsymbol{\theta}_1), \ldots, \alpha_k \mathbf{I}(\boldsymbol{\theta}_k), \mathbf{M}\},$$

where $\mathbf{I}(\boldsymbol{\theta}_m)$, for $m = 1, ..., k$, is the Fisher matrix for a single observation produced by the m-th component, and $\mathbf{M}$ is the Fisher matrix of a multinomial distribution.

Since $|\mathbf{M}| = (\alpha_1 \alpha_2 \cdots \alpha_k)^{-1}$ (see, *e.g.*, [12]), we have

$$\log|\mathbf{I}_c(\boldsymbol{\theta}_{(k)})| = \sum_{i=1}^{k} \log|\mathbf{I}(\boldsymbol{\theta}_i)| + \log n^{k(N+1)} + \sum_{i=1}^{k} \log \alpha_i^{(N-1)}, \tag{5}$$

where N is the dimension of the $\boldsymbol{\theta}_i$'s.

For $p(\boldsymbol{\theta}_{(k)})$, we model the parameters of different components as *a priori* mutually independent and also independent from the mixing probabilities. Formally, $p(\boldsymbol{\theta}_{(k)}) = p(\boldsymbol{\theta}_1) \cdots p(\boldsymbol{\theta}_k)\, p(\alpha_1, .., \alpha_k)$, where each factor is the corresponding non-informative Jeffreys' prior [12], *i.e.*, $p(\boldsymbol{\theta}_i) \propto \sqrt{|\mathbf{I}(\boldsymbol{\theta}_i)|}$ and $p(\alpha_1, .., \alpha_k) \propto \sqrt{|\mathbf{M}|}$. Inserting this prior and Eq. (5) into Eq. (4) we finally obtain

$$\widehat{\boldsymbol{\theta}_{(k)}} = \arg\min_{\boldsymbol{\theta}_{(k)}} \left\{ \frac{N}{2} \sum_{i=1}^{k} \log \alpha_i + \frac{kN + k}{2} \log n - L(\boldsymbol{\theta}_{(k)}, \mathbf{y}) \right\} \tag{6}$$

From a Bayesian viewpoint, Eq. (6) is a MAP criterion, for each k, with a Dirichlet prior $p(\{\alpha_m\}) \propto \exp\{-(N/2) \sum_m \log \alpha_m\}$ on the mixing probabilities (with negative parameters, thus improper [12]).

2.2. Implementation via EM

Since Dirichlet priors are conjugate to multinomial likelihoods [12], to implement Eq. (6) via EM, the M-step becomes (recall the constraints $\alpha_m \geq 0$ and $\sum \alpha_m = 1$)

$$\widehat{\alpha}_m^{(t+1)} = \frac{\max\left\{0, \left(\sum_{i=1}^{n} w_m^{(i,t)}\right) - \frac{N}{2}\right\}}{\sum_{j=1}^{k} \max\left\{0, \left(\sum_{i=1}^{n} w_j^{(i,t)}\right) - \frac{N}{2}\right\}}. \tag{7}$$

The $\boldsymbol{\theta}_m$'s are updated by simply maximizing EM's Q-function with respect to them. Note that this M-step may annihilate components; it is an explicit rule for moving from a certain value of k to a smaller one. Accordingly, we propose to start with a large value of k, and let EM, via Eq. (7), annihilate redundant components. Moreover, this new M-step provides increased robustness against local minima. For example, configurations where, say, two components have similar parameters are problematic. Under the criterion in Eq. (6), such configurations are unstable; one of them is eventually killed. Another key feature is that the boundary of the parameter space, for a given k, is no longer reachable: when one of the α_m's becomes too small, it is annihilated and the algorithm jumps to a mixture with $k - 1$ components.

It can be shown that $\sum_i \log \alpha_i \propto -\mathcal{D}_{\text{KL}}[\{1/k\} \,\|\, \{\alpha_m\}]$, the Kullback-Leibler divergence between a uniform distribution and the one specified by the α_m's. In other words, our criterion favors less uniform (lower entropy) distributions, sharing the spirit of recent work in [13]. However, unlike [13], we have a closed-form M-step and explicit component annihilation (no additional tests).

[1] In matrix sense, *i.e.*, $\mathbf{I}_c(\boldsymbol{\theta}_{(k)}) - \mathbf{I}(\boldsymbol{\theta}_{(k)})$ is positive definite [2].

3. The Self Annealing Behavior of EM

Deterministic annealing (DA) is a fast surrogate of (stochastic) simulated annealing; it has been successfully applied in many problems, namely in clustering [15, 16].

The DA approach to k-means clustering (as described in [15]) leads to an algorithm that is similar to EM for fitting mixtures of Gaussians with a common covariance matrix of the form $T\,\mathbf{I}$ (where $\mathbf{I}$ is the identity matrix and T is called *temperature*). DA clustering starts at high temperature (forcing high entropy assignments); T is then lowered according to some *cooling schedule* until $T \simeq 0$. The heuristic behind DA is that forcing the entropy of the assignments to decrease slowly avoids premature (hard) decisions that may correspond to poor local minima.

DA versions of EM (DAEM) have been proposed as a means of overcoming its initialization dependence [14]. For finite mixture fitting via EM, the average entropy of the assignments is given (at iteration t) by

$$H(t) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{k} w_m^{(i,t)} \log w_m^{(i,t)}. \qquad (8)$$

In DAEM, this entropy is initially held high (by modifying Eq. (1)), and then forced to decrease slowly [14].

In another front, self annealing (SA) was proposed in [17] as a means of obtaining DA algorithms without prespecified cooling schedules. Formally, given some cost function $E(\boldsymbol{\phi})$, whose minimum is to be found with respect to $\boldsymbol{\phi}$, consider the iteration

$$\boldsymbol{\phi}^{(t+1)} = \arg\min_{\boldsymbol{\phi}} \left\{ E(\boldsymbol{\phi}) + \beta\, d(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t)}) \right\}, \qquad (9)$$

where $d(\boldsymbol{\phi}, \boldsymbol{\phi}') \geq 0$, and $d(\boldsymbol{\phi}, \boldsymbol{\phi}') = 0 \Leftrightarrow \boldsymbol{\phi} = \boldsymbol{\phi}'$. The key observation in [17] is: if $\boldsymbol{\phi}$ controls the entropy of the assignments, and a high entropy initialization is used, this iterative procedure exhibits "self annealing". That is, due to the presence of $d(\cdot,\cdot)$, the "cooling" is self-controlled.

It turns out that Eq. (9) defines a so-called *proximal point algorithm* (PPA) and it can be shown (see [18]) that EM is a PPA with $E(\boldsymbol{\theta}_{(k)}) = -L\big(\boldsymbol{\theta}_{(k)}, \mathbf{y}\big)$, $\beta = 1$, and

$$d(\boldsymbol{\theta}_{(k)}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}) = \mathcal{D}_{\text{KL}}\left[p(\mathbf{z}|\mathbf{y}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}) \;\|\; p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(k)}) \right].$$

This $d(\cdot,\cdot)$ function is a relative entropy involving distributions of the (missing) assignment variables: as in DA and SA, also in EM it is the entropy of these assignments that is being controlled. Observe also that the function being minimized in Eq. (9) is analogous to the *Helmholtz free energy* (see [15, 16]), for unit temperature, with the relative entropy $\mathcal{D}_{\text{KL}}[p(\mathbf{z}|\mathbf{y}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}) \;\|\; p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(k)})]$ playing the role of entropy. Accordingly, EM behaves like a SA algorithm, as long as a high-entropy initialization is used; in the mixture case, this simply means $w_m^{i,0} = 1/k + \varepsilon_m$, where the ε_m are small random perturbations (of course we can't use $w_m^{i,0} = 1/k$ because that is a fixed point of EM).

4. Experiments

Fig. 1 shows 900 samples of a mixture used in [14]: three equiprobable Gaussians (means $[0, -2]^T$, $[0,\, 0]^T$, $[0,\, 2]^T$; same covariance diag$\{2,\, 0.2\}$). Initialization ($k = 10$), two intermediate estimates ($k = 8$, $k = 5$), and the final result are presented. The evolution of the criterion function (Eq. (6)) and of the entropy $H(t)$ (note its controlled decay) are also shown. In conclusion, for this mixture, our method successfully overcomes the initialization issue, like DAEM in [14]; however, it **(i)** does not require a cooling schedule, and **(ii)** autonomously found the correct number of components.

The next example considers Gaussian mixtures to model class-conditional densities (*mixture discriminant analysis*, MDA, [4]). The specific problem we address is one with 3 (equiprobable) classes in 21-dimensional space, studied in [4] (for details, see [4]). As in [4], the class-conditional mixtures are fitted to sets of 300 samples (∼100 per class); the resulting MAP classifier is then tested on 500 samples. We have compared two methods: **(a)** MDA based on our new method, with diagonal covariances and initialized with $k = 7$; **(b)** MDA with 3 common covariance components per class, estimated via EM initialized as described in [4]: k-means clustering is run from 10 random starts and the results used to initialize EM; the best final result is then chosen. Method **(b)** was shown in [4] to clearly outperform both linear and quadratic discriminant analysis. The results in Table 1 show that MDA based on our method beats MDA as used in [4]; moreover, it does not require external initialization and it adaptively selects the number of components.

Table 1. Average error rates (over 10 simulations) for the methods described in the text.

Method	Average (standard error)
MDA - new method	0.158 (0.005)
MDA - as in [4]	0.167 (0.005)

The final example, reported in Fig. 2, illustrates the good performance of our method in fitting a Gaussian mixture to an arbitrary probability density. The 900 data points were generated with the *noisy shrinking spiral* model described in [8]:

$$\begin{bmatrix} x_1^i \\ x_2^i \\ x_3^i \end{bmatrix} = \begin{bmatrix} (13 - 0.5t_i)\cos t_i \\ (0.5t_i - 13)\sin t_i, \\ t_i \end{bmatrix} + \begin{bmatrix} n_1^i \\ n_2^i \\ n_3^i \end{bmatrix}$$

for $i = 1...900$, the t_i uniformly distributed in $[0, 4\pi]$, and n_1^i, n_2^i, and n_3^i i.i.d. zero-mean Gaussian samples.

5. Conclusions

A new unsupervised algorithm for selection and estimation of finite mixture models is proposed. It is based on an MML-type criterion and on the observation that EM exhibits self-annealing. Examples have shown the good performance of the approach. Future work includes further

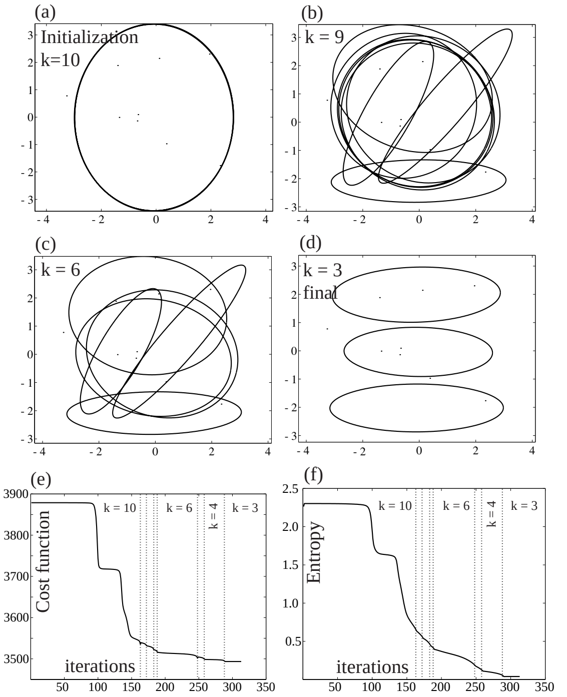

Figure 1. A 3-component Gaussian mixture. The ellipses are isodensity curves of each component. In (e) and (f), vertical dotted lines signal the annihilation of one component.

experimental evaluation (*e.g.*, with non-Gaussian mixtures, and with latent variable models such as mixtures of factor analyzers [8]).

References

[1] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs: Prentice Hall, 1988.

[2] G. McLachlan and K. Basford, *Mixture Models: Inference and Application to Clustering*. Marcel Dekker, 1988.

[3] D. Titterington, A. Smith, and U. Makov, *Statistical Analysis of Finite Mixture Distributions*. Chichester (U.K.): John Wiley & Sons, 1985.

[4] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society (B)*, vol. 58, pp. 155–176, 1996.

[5] M. Figueiredo, J. Leitão, and A. K. Jain, "On fitting mixture models," in *Energy Minimization Methods in Computer Vision and Pattern Recognition* (E. Hancock and M. Pellilo, eds.), pp. 54–69, Springer Verlag, 1999.

[6] S. Roberts, D. Husmeier, I. Rezek, W. Penny, "Bayesian approaches to Gaussian mixture modelling," *IEEE Trans. on PAMI*, vol. 20, pp. 1133-1142, 1998.

[7] J. Oliver, R. Baxter, and C. Wallace, "Unsupervised learning using MML," in *Proc. of the 13th Int. Conf. on Machine Learning*, (San Francisco), pp. 364–372, 1996.

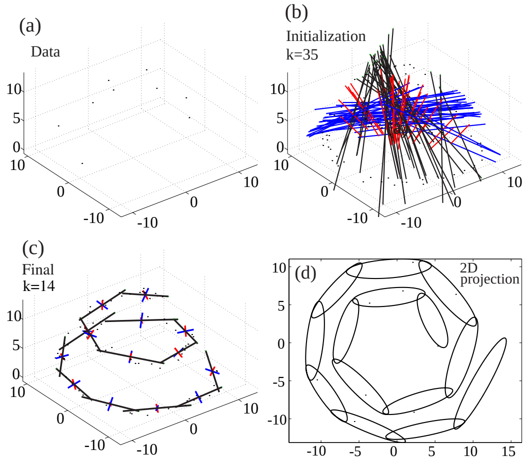

Figure 2. The 3D noisy shrinking spiral. The line segments in (b) and (c) are the axes of the isodensity ellipsoids. A 2D projection of (c) (with isodensity curves) is shown in (d).

[8] N. Ueda, R. Nakano, Z. Gharhamani, and G. Hinton, "SMEM algorithm for mixture models," *Neural Computation*. To appear.

[9] S. Richardson and P. Green, "On Bayesian analysis of mixtures with unknown number of components," *Journal of the Royal Statistical Society B*, vol. 59, pp. 731–792, 1997.

[10] J. Rissanen, *Stochastic Complexity in Stastistical Inquiry*. Singapore: World Scientific, 1989.

[11] C. Wallace and P. Freeman, "Estimation and inference via compact coding," *Journal of the Royal Statistical Society (B)*, vol. 49, no. 3, pp. 241–252, 1987.

[12] J. Bernardo and A. Smith, *Bayesian Theory*. Chichester, UK: J. Wiley & Sons, 1994.

[13] M. Brand, "Structure learning in conditional probability models via entropic prior and parameter extinction," *Neural Computation*, vol. 11, pp. 1155–1182, 1999.

[14] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," *Neural Networks*, vol. 11, pp. 271–282, 1998.

[15] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. of the IEEE*, vol. 86, pp. 2210–2239, 1998.

[16] T. Hofmann and J. Buhmann, "Pairwise data clustering by deterministic annealing," *IEEE Trans. on PAMI*, vol. 19, pp. 1–14, January 1997.

[17] A. Rangarajan, "Self annealing: unifying deterministic annealing and relaxation labeling," in *Energy Minimization Methods in Computer Vision and Pattern Recognition* (M. Pellilo and E. Hancock, eds.), pp. 229–244, Springer Verlag, 1997.

[18] S. Chretien and A. Hero, "Kullback proximal algorithms for maximum likelihood estimation," *Submitted to IEEE Trans. on Info. Theo.*, 1999.