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 propose a new method for fitting mixture models that performs component selection and does not require external initialization. The novelty of our approach includes: a minimum message length (MML) type model selection criterion; the inclusion of the criterion into the expectation-maximization (EM) algorithm (which also increases 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 tool. In pattern recognition, mixtures underlie formal approaches to unsupervised learning (clustering) [1, 2]. 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 [3].

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

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

where $\alpha_1, ..., \alpha_m$ 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 of the FM parameters, $\widehat{\boldsymbol{\theta}}_{(k)} = \arg\max_{\boldsymbol{\theta}_{(k)}} L(\boldsymbol{\theta}_{(k)}, \mathbf{y})$ 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 $L\big(\boldsymbol{\theta}_{(k)}, \mathbf{y}\big)$ or $[L\left(\boldsymbol{\theta}_{(k)}, \mathbf{y}\right) + \log p(\boldsymbol{\theta}_{(k)})]$ [4].

EM is supported on the interpretation of $\mathbf{y}$ as *incomplete* data [2, 4]. Here, the *missing* part is a set of labels $\mathbf{z} = \{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(n)}\}$, indicating which component produced each observation. The labels have the form $\mathbf{z}^{(i)} = [z_1^{(i)}, ..., z_k^{(i)}]$; if $\mathbf{y}^{(i)}$ was produced by the m-th component, then $z_m^{(i)} = 1$ and $z_p^{(i)} = 0$, for $p \neq m$. The (complete) log-likelihood (*i.e.*, if *complete* data $\mathbf{x} = \{\mathbf{y}, \mathbf{z}\}$ was observed) is [2, 4]

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

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 the current parameter estimate $\widehat{\boldsymbol{\theta}}_{(k)}^{(t)}$,

$$Q(\boldsymbol{\theta}_{(k)}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}) \equiv E\left[L_c\left(\boldsymbol{\theta}_{(k)}, \mathbf{y}, \mathbf{z}\right) \middle| \mathbf{y}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}\right]. \qquad (1)$$

Since L_c is linear in the missing $z_m^{(i)}$'s, this step reduces to the computation of their conditional expectations [2, 4]. Since the $z_m^{(i)}$'s are binary, $E[z_m^{(i)}|\cdot] = \Pr[z_m^{(i)} = 1|\cdot]$; then,

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

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

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

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

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

where $\mathcal{C}(\widehat{\boldsymbol{\theta}}_{(k)}, k)$ is some model selection criterion. Several of these methods (see references/comparisons in [5, 6]) have good model selection performance, but a major drawback remains: a whole set of $k_{\max}$ candidate models 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 best (highest $L(\widehat{\boldsymbol{\theta}}_{(k)}, \mathbf{y})$) estimate [3, 4, 5]; other schemes initialize the $w_m^{(i,t)}$ variables using clustering methods [3, 4]. **(b)** EM may converge to the boundary of the parameter space, *i.e.*, one of the α_m's approaches zero and the corresponding component becomes singular (unbounded likelihood); when the value of k is larger than the optimal/true one, this may happen frequently.

2. Proposed Approach

2.1. The Proposed Criterion

The *minimum description length* (MDL, [7]) and *minimum message length* (MML, [8, 6]) are two well known criteria which have been successfully used for FM model selection [5, 6]. However, the approach has been the one in Eq. (3), suffering from the draw-backs mentioned above.

To bypass these difficulties, we propose a shift of approach: we use a selection criterion that can be embedded in the steps of the EM algorithm, thus obtaining 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 $p(\boldsymbol{\theta}_{(k)}|k)$. Let $p(k) = 1/k_{\max}$, for some $k_{\max}$ known to be larger than the true k. The simultaneous selection of k and estimation of $\boldsymbol{\theta}_{(k)}$, denoted $\widehat{\boldsymbol{\theta}_{(k)}}$, is

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

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. Eq. (4) is an MML criterion (as used, *e.g.*, in [6]), the only difference being that we ignore the *optimal quantizing lattice* constants, as is done in MDL [7].

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 [2].

Since $|\mathbf{M}| = (\alpha_1\alpha_2\cdots\alpha_k)^{-1}$ (see, *e.g.*, [9]), we have

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

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

For the prior, we model the parameters of different components as independent and also independent from the mixing probabilities: $p(\boldsymbol{\theta}_{(k)}) = p(\boldsymbol{\theta}_1)\cdots p(\boldsymbol{\theta}_k)p(\alpha_1, .., \alpha_k)$. For each of these factors we adopt non-informative Jeffreys' priors [9]: $p(\boldsymbol{\theta}_i) \propto \sqrt{|\mathbf{I}(\boldsymbol{\theta}_i)|}$ and $p(\alpha_1, .., \alpha_k) \propto \sqrt{|\mathbf{M}|}$. Inserting this prior and Eq. (5) into Eq. (4) we obtain

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

2.2. Implementation via EM

From a Bayesian point of view, Eq. (6) includes, for each k, a Dirichlet-type prior for the α_m's, $p(\{\alpha_m\}) \propto \exp\{-(N/2)\sum_m \log\alpha_m\}$ (with negative parameters, thus improper [9]). Dirichlet priors are conjugate to multinomial likelihoods [9]; thus, in the M-step of EM, the α_m's are updated as (recall that $\alpha_m \geq 0$ 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 the Q-function (Eq. (1)) with respect to them. Note that this **M-step** performs component annihilation, thus being 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 several components have similar parameters are problematic. Under the criterion in Eq. (6), those configurations are unstable, with one of them eventually being annihilated. Another key feature is that the boundary of the parameter space, for each k, is no longer reachable: when one of the α_m's becomes too small, it is annihilated and the algorithm jumps to a smaller sub-space.

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

As a final remark, it can be shown that $\sum_i \log \alpha_i \propto -\mathcal{D}_{\text{KL}}[\{1/k\} \parallel \{\alpha_m\}]$, the Kullback-Leibler divergence between a uniform distribution and the one specified by the α_m's. That is, we are favoring less uniform (lower entropy) distributions, sharing the spirit of recent work in [10]. However, unlike [10], we have closed-form updates for the α_m's and explicit component annihilation (no additional tests).

3. The Self Annealing Behavior of EM

Deterministic annealing (DA) versions of EM (DAEM) have been proposed as a means of overcoming its initialization dependence [11, 12]. DA is a fast surrogate of simulated annealing which has been successfully applied in many problems, namely in clustering [13, 14].

The DA approach to k-means clustering is similar to EM for Gaussian mixtures [13]; in fact, k-means clustering coincides with Gaussian mixture fitting when all k components share a common covariance, $T\,\mathbf{I}$ (where $\mathbf{I}$ is the identity matrix), with vanishing T (called *temperature*) [13]. In DA, the hard clusters are "softened" by starting with a high temperature (high entropy assignments); T is then lowered according to some *cooling schedule* until $T \rightarrow 0$. The heuristic behind DA is that by forcing the entropy (softness) of the assignments to decrease slowly, premature (hard) decisions that may lead to poor local minima are avoided.

When estimating a finite mixture via EM, the entropy (softness) of the assignments is given (at iteration t) by

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

DAEM schemes work by artificially forcing this entropy to stay higher, and then controlling its (slow) decay [11, 12].

In another front, self annealing (SA) was described in [15] 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 a vector parameter $\boldsymbol{\phi}$, consider the iteration

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

where $d(\boldsymbol{\phi}, \boldsymbol{\phi}') \geq 0$, and $d(\boldsymbol{\phi}, \boldsymbol{\phi}') = 0 \Leftrightarrow \boldsymbol{\phi} = \boldsymbol{\phi}'$ [15]. The key observation in [15] is: if $\boldsymbol{\phi}$ contains T, and we use a "high T" initialization, this iterative procedure exhibits "self annealing". That is, the temperature does not decrease "too fast" (due to $d(\cdot,\cdot)$); the cooling is self-controlled.

It is easy to show that EM iterations do have the form of Eq. (9), with $E(\boldsymbol{\theta}_{(k)}) = -L\big(\boldsymbol{\theta}_{(k)}, \mathbf{y}\big)$ and

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

A related result is found in [16]. This $d(\cdot,\cdot)$ function is a relative entropy involving distributions of the missing variables which control the assignment of data points to mixture components: as in DA and SA, also in EM it is the entropy of this assignment that is being controlled. Observe also that the function being minimized in each step is analogous to a *free energy* (see [13, 14]), for unit temperature, with the relative entropy $\mathcal{D}_{\text{KL}}[p(\mathbf{z}|\mathbf{y}, \widehat{\boldsymbol{\theta}}_{(k)}^{(t)}) \parallel p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}_{(k)})]$ playing the role of entropy. Accordingly, EM behaves like a SA algorithm, and all that is required is high-entropy initialization; in the case of mixtures, this simply means $w_m^{i,0} \simeq 1/k$.

In summary, we propose: use EM with the **M-step** in Eq. (7), starting with some k known to be larger than the true/optimal one, and initialized with $w_m^{i,0} \simeq 1/k$.

4. Examples

Fig. 1 shows 900 samples of a mixture used in [12]: three equiprobable Gaussian components with means $[0, -2]^T$, $[0,\ 0]^T$, $[0,\ 2]^T$, and equal covariances diag$\{2,\ 0.2\}$. Starting with $k = 10$, initialization with $w_m^{i,0} \simeq 1/10$ leads to almost coincident components, as shown. Intermediate estimates ($k = 8$, $k = 5$) and the final result are presented. We also plot the evolution of the criterion function (Eq. (6)) and of the entropy $H(t)$ (observe its controlled decay). In conclusion, for this mixture, our method successfully overcomes the initialization issue, like DAEM in [12]; however, our method (i) does not require a cooling schedule, and (ii) autonomously found the correct number of components.

In the next example we use Gaussian mixtures to model class-conditional densities; this is called *mixture discriminant analysis* (MDA) in [3]. The specific problem we address is one with three (equiprobable) classes in 21-dimensional space, studied in [3]. Each observation is defined as $\mathbf{y}^{(i)} = [y_1^{(i)}, ... y_{21}^{(i)}]^T$, with

$$\begin{aligned}
y_j^{(i)} &= u^{(i)} h_1(j) + (1 - u^{(i)}) h_2(j) + n_j^{(i)}, \quad \text{Class 1,} \\
y_j^{(i)} &= u^{(i)} h_1(j) + (1 - u^{(i)}) h_3(j) + n_j^{(i)}, \quad \text{Class 2,} \\
y_j^{(i)} &= u^{(i)} h_2(j) + (1 - u^{(i)}) h_3(j) + n_j^{(i)}, \quad \text{Class 3,}
\end{aligned}$$

where the $u^{(i)}$ are i.i.d. uniform in $(0,1)$, the $n_j^{(i)}$ are i.i.d. zero-mean unit-variance Gaussian, and the h_i functions are shifted triangular waveforms: $h_1(j) = \max(0, 6 - |j - 11|)$, $h_2(j) = h_1(j-4)$, and $h_3(j) = h_1(j+4)$. Like in [3], the mixtures representing the class-conditional densities are fitted to sets of 300 samples (roughly 100 per class); the resulting *maximum a posteriori* classifier is then tested on samples of size 500. Table 1 reports error rates for four methods: **(a)** MDA based on our new method, with diagonal covariances and initialized with $k = 7$; **(b)** MDA with standard EM, using 3 components per class, common covariance matrix, and initialized as described in [3] (*i.e.*, the k-means algorithm is run from 10 random starts and the results used to initialize EM; the best final result is then chosen); **(c)** *linear discriminant analsyis* (LDA – classes modelled as Gaussian with different means but common covariance); and **(d)** *quadratic discriminant analysis* (QDA

– Gaussian classes with different means and different covariances). MDA based on our method has the best performance; moreover, it does not suffer from the initialization difficulties of standard EM and it does not require the user to specify the number of components of each mixture.

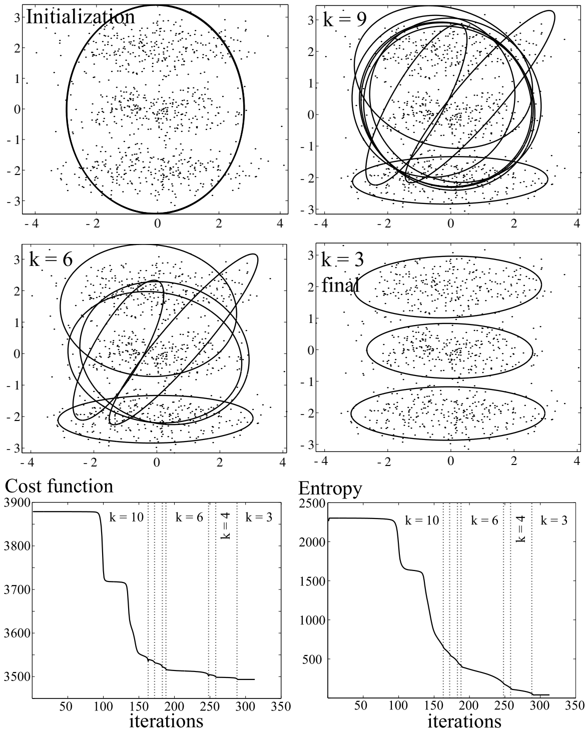

Figure 1. A 3-component Gaussian mixture. The ellipses represent isodensity curves of each component. The vertical dotted lines signal the annihilation of one component.

Method	Average (standard error)
MDA - new method	0.158 (0.005)
MDA - EM	0.167 (0.005)
LDA	0.195 (0.008)
QDA	0.211 (0.008)

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

5. Conclusions

A new unsupervised algorithm for selection and estimation of finite mixture models was proposed. It is based on a 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 experimental evaluation (*e.g.*, on non-Gaussian mixtures).

References

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

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

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

[4] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: John Wiley & Sons, 1997.

[5] S. Roberts, D. Husmeier, I. Rezek, and W. Penny, “Bayesian approaches to Gaussian mixture modelling,” *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 20, pp. 1133-1142, 1998.

[6] 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.

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

[8] 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.

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

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

[11] M. Kloppenburg and P. Tavan, “Deterministic annealing for density estimation by multivariate normal mixtures,” *Physical Review E*, vol. 55, pp. R2089–R2092, 1997.

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

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

[14] T. Hofmann and J. Buhmann, “Pairwise data clustering by deterministic annealing,” *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 19, pp. 1–14, January 1997.

[15] A. Rangarajan, “Self annealing,” in *Energy Minimization Methods in Comp. Vis. and Patt. Rec.* (M. Pellilo and E. Hancock, eds.), pp. 229–244, Springer Verlag, 1997.

[16] R. Neal and G. Hinton, “A view of the EM algorithm that justifies incremental, sparse, and other variants,” in *Learning in Graphical Models* (M.I. Jordan, ed.), pp. 355–368, Kluwer, 1998.