# Robust deterministic annealing based EM algorithm

B. Wang, F. Wan, P.-U. Mak, P.-I. Mak and M.-I. Vai

A deterministic annealing (DA)-based expectation-maximisation (EM) algorithm is proposed for robust learning of Gaussian mixture models. By combing the DA approach, trimmed likelihood function and Bayesian information criterion (BIC), the proposed algorithm can simultaneously perform model selection and outlier detection, and mitigate the problems of local optima and boundary of parameter space with the conventional EM algorithm. Experiments demonstrate that the proposed algorithm can determine the number of components correctly even though the data are contaminated by outliers.

*Introduction:* A general approach to parameter estimation of Gaussian mixture models (GMMs) is to iteratively calculate the maximum likelihood (ML) solution via the expectation-maximisation (EM) algorithm. However, the ML estimate calculated by the conventional EM algorithm for GMMs suffers from three problems. First, it may converge to local maxima of the log-likelihood function, or the boundary of parameter space. Secondly, the model order has to be assumed known beforehand, otherwise the estimation result may be poor. Unfortunately, this prior knowledge is not always available in many applications. The third problem is its sensitivity to outliers, which usually exist in practice owing to factors such as background noise or measurement inaccuracies.

In this Letter, a robust deterministic annealing EM algorithm (RDAEM) is presented, where the mixture models are learned based on maximum trimmed likelihood (MTL) [1] and Bayesian i nformation criterion [2, 3], with the top-down learning process controlled by the annealing approach [4]. As a result, the proposed method is less sensitive to local optima, avoids the boundary of parameter space, and can simultaneously perform model selection and outlier detection.

*Proposed RDAEM algorithm:* A general approach to parameter estimation of GMMs is the EM algorithm, which iteratively calculates the ML solution given by

$$F_{ML}(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\Theta}) = \frac{1}{(2\boldsymbol{\pi})^d|\boldsymbol{\Sigma}_k|} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \tag{1}$$

where $N$ is the number of samples, $K$ is the number of Gaussian components and $d$ is the dimension of samples. $\boldsymbol{\Theta} \equiv \{\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k\}$. $\pi_k$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, are the mixing coefficient, mean and covariance of $k$th Gaussian component, respectively.

Several model selection criteria have been proposed to estimate the number of Gaussian components. One of them is BIC, which is given by

$$F_{BIC}(\boldsymbol{\Theta}) = \frac{KJ}{2}\log N - \log p(\boldsymbol{X}|\boldsymbol{\Theta}) = \frac{KJ}{2}\log N - \sum_{n=1}^{N} \log p \sum_{k=1}^{N} \boldsymbol{\pi}_k p(\boldsymbol{x}_n|\boldsymbol{\theta}_k) \tag{2}$$

where $J$ is the number of free parameters specifying each component.

To estimate GMMs in a robust way, one approach is to calculate the MTL solution, which is given by

$$F_{MTL}(\boldsymbol{\Theta}, \boldsymbol{\omega}) = \sum_{n=1}^{N} \omega_n \log p(\boldsymbol{x}_n|\boldsymbol{\Theta}) \tag{3}$$

where $\omega_n \in \{0, 1\}\forall n= 1, \ldots, N$, and $\sum_{n=1}^{N}\omega_n = M$. $\omega_n$ takes the value 0 if the sample $\boldsymbol{x}_n$ is considered as an outlier. Otherwise, $\boldsymbol{\omega}_n = 1$. In other words, a subset of size $M$ out of $N$ original samples will be selected to maximise trimmed likelihood function.

To simultaneously perform model selection and outlier detection, we combine the cost functions (2) and (3), and let the parameter estimation be controlled by the DA approach. This leads to the following cost function

$$F_{RDAEM}(\boldsymbol{\Theta}, \boldsymbol{\omega}, \boldsymbol{\upsilon}) = \sum_{k=1}^{K} U_k \frac{J}{2}\log M - \sum_{n=1}^{N} \omega_n \log \sum_{k=1}^{N} U_k \boldsymbol{\pi}_k p(\boldsymbol{x}_n|\boldsymbol{\theta}_k) - T_\omega H_\omega - T_U H_U \tag{4}$$

where $U_k \in \{0, 1\}, \forall k = 1, \ldots, K$. $U_k$ takes value 0 if the $k$th component is removed, otherwise $U_k = 1$. $T_\omega$ and $T_U$ are Lagrange multipliers. $H_\omega$ and $H_U$ are Shannon entropy and binary entropy, respectively, given by

$$H_\omega = -\sum_{n=1}^{N} \omega_n \log \omega_n \text{ and } H_U = -\sum_{k=1}^{K}(U_k \log U_k + (1-U_k)\log(1-U_k)) \tag{5}$$

To maximise (4), we modify the EM algorithm and the RDAEM algorithm for GMMs is described as follows:

```
1. Initialise the parameter set $\boldsymbol{\Theta}$ of the mixture models, scaling
   factors $\boldsymbol{\alpha}_\omega$, $\boldsymbol{\alpha}_U$, $T_{\omega\min}$, $T_{U\min}$ and $T_{\omega\max}$, $T_{U\max}$. $T_\omega = T_{\omega\max}$, $T_U =$
   $T_{U\max}$, $t = 0$.
2. Repeat
      Repeat
            $t = t + 1$
            E-step: Calculate posterior probability using current parameters

            $$p(k|\mathbf{x}_n) = \frac{U_k \boldsymbol{\pi}_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} U_j \pi_j N(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{6}$$

            M-step: update parameter set $\boldsymbol{\Theta}$ using current posterior
            probability

            $$\boldsymbol{\pi}_k = \frac{1}{M}\sum_{n=1}^{N} \boldsymbol{\omega}_n p(k|\boldsymbol{x}_n), \quad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \boldsymbol{\omega}_n p(k|\boldsymbol{x}_n)\boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{\omega}_n p(k|\boldsymbol{x}_n)} \tag{7}$$

            $$\boldsymbol{\Sigma}_k \frac{\sum_{n=1}^{N} \boldsymbol{\omega}_n p(k|\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \boldsymbol{\omega}_n p(k|\boldsymbol{x}_n)} \tag{8}$$

            Update $\boldsymbol{\omega}$ and $\boldsymbol{\upsilon}$
      Until a stop criterion is met
      $T_\omega = \boldsymbol{\alpha}_\omega T_\omega$, $T_U = \boldsymbol{\alpha}_U T_U$
   Until $T_\omega < T_{\omega\min}$ and $T_U < T_{U\min}$
3. Return $\boldsymbol{\Theta}$, $\boldsymbol{\omega}$, $\boldsymbol{\upsilon}$
```

The parameter set $\boldsymbol{\omega}$ is calculated as the solution of the following objective function minimisation:

$$\boldsymbol{\omega}_n = \arg\min_{\omega_n}\left\{\begin{array}{l} F_{RDAEM}(\boldsymbol{\Theta}, \boldsymbol{\omega}, \boldsymbol{\upsilon}) \\ \text{s.t.} \sum_{n=1}^{N} \boldsymbol{\omega}_n = M, \omega_n \in [0, 1] \end{array}\right\} \tag{9}$$

This is a bound-constrained convex optimisation which can be solved by some standard optimisation tools (e.g. cvx [5]).

The parameter set $\boldsymbol{\upsilon}$ is calculated by computing the solution of

$$\frac{J}{2}v_k \log M - \frac{N_k}{\ln 2} + V_k T_U \log\frac{V_k}{1-V_k} = 0 \tag{10}$$

where $N_k$ is given by

$$N_k = \sum_{n=1}^{N} \frac{V_k W_n \boldsymbol{\pi}_k p(\boldsymbol{x}_n|\boldsymbol{\theta}_k)}{\sum_{j=1}^{N} V_j \boldsymbol{\pi}_j p(\boldsymbol{x}_n|\boldsymbol{\theta}_j)} \tag{11}$$

*Experiments:* We have applied the algorithm to a synthetic data set, and the learning process is demonstrated in Fig. 1. The data set consists of 1000 samples generated from eight Gaussians with equal mixing coefficients. In addition, 300 samples are added from a uniform distribution. The colour of each sample indicates the value of weight $\boldsymbol{\omega}_n \in [0, 1]$: a sample is in cyan if $\boldsymbol{\omega}_n = 0$ (i.e. it is identified as an outlier) or in magenta if $\boldsymbol{\omega}_n = 1$ (i.e. it is identified as a typical sample). The parameters are set as $\boldsymbol{\alpha}_\omega = \boldsymbol{\alpha}_U = 0.9$, $T_\omega = T_U = T$, $T_{\max} = 100$, $T_{\min} = 0.01$. During the learning process, the singular components (Fig. 1*b*, Fig. 1*d*) and the overlapped components (Fig. 1*d*, Fig. 1*f*) can be annihilated (Fig. 1*c*, Fig. 1*e*, Fig. 1*g*), indicating that RDAEM can escape the boundary of parameter space and local optima. In addition, the smooth change of the samples colours illustrates that the outliers are detected and removed. Finally, RDAEM can successfully perform model selection and outlier detection.

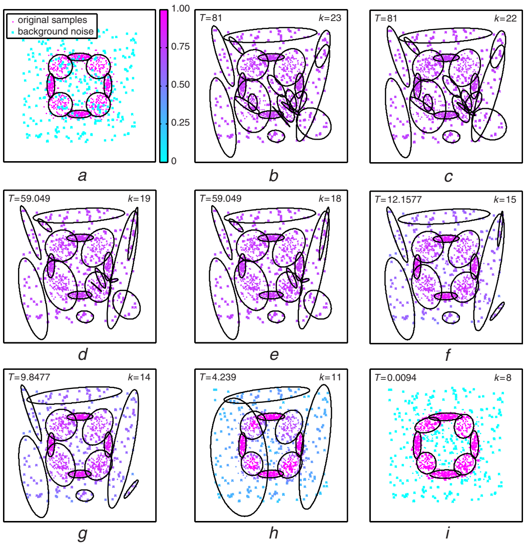

**Fig. 1** *Learning process of RDAEM*

*a* True model
*b–h* Intermediate results
*i* Final estimate

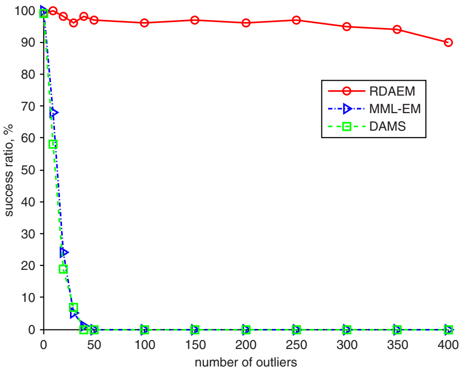

**Fig. 2** *Percentages of correctly identified numbers of components by RDAEM, DAMS and MML-EM*

To demonstrate the negative effects of outliers, we have also compared the proposed method with other two well-known model selection algorithms: the deterministic annealing based model selection (DAMS) [3] and the minimum message length (MML) based EM algorithm (MML-EM) [6]. We repeat the example in Fig. 1 100 times at different noise levels, and the performances in terms of percentages of correctly identified number of components are shown in Fig. 2. Owing to the background noise, the performances of MML-EM and DAMS are degraded dramatically, and thus their learning results are unreliable. However, RDAEM achieves robust results.

Finally, we have applied RDAEM to the electroencephalography (EEG) signal classification task. Noise is ubiquitous in EEG signals owing to factors such as the muscle and eye blink artefacts, measurement inaccuracies, and physiological variations in background EEG. Therefore, it is necessary to prune some samples to achieve more reliable results. The data set used is the data set IV in BCI competition II [7], which consists of 316 training samples and 100 testing samples. We prune 5%, 10%, 15% training samples and fit each class using RDAEM. We repeat the experiments 50 times and the average classification accuracies are shown in Table 1. It can been seen that, after sample pruning, the classification accuracy is improved.

**Table 1:** Average accuracies of different algorithms for EEG signal classification

| Algorithms | Accuracies (%) |
|---|---|
| DAMS | 78.1 ± 3.5 |
| MML-EM | 79.2 ± 2.7 |
| RDAEM (5%) | 82.7 ± 3.8 |
| RDAEM (10%) | 83.5 ± 3.3 |
| RDAEM (15%) | 81.9 ± 2.1 |

*Conclusion:* The RDAEM algorithm is proposed to simultaneously deal with three problems including boundary of parameter space, local optima, model selection and outlier detection. The experimental results demonstrate that RDAEM achieves more reliable results than standard algorithms, especially when the data are contaminated by noise.

*Acknowledgments:* This work was supported in part by the Macau Science and Technology Development Fund (Grant FDCT/036/2009/A) and the University of Macau Research Fund (Grants RG059/08-09S/FW/FST, RG080/09-10S/WF/FST and MYRG139 (Y1-L2)-FST11-WF).

*2 September 2011*
doi: 10.1049/el.2011.2797
One or more of the Figures in this Letter are available in colour online.

B. Wang, F. Wan, P.-U. Mak, P.-I. Mak and M.-I. Vai (*Department of Electrical and Computer Engineering, University of Macau, Av. Padre Tomás Pereira Taipa, Macau, People's Republic of China*)

E-mail: fwan@umac.mo

B. Wang is also with the School of Computer Science, McGill University, Montreal, Quebec, H3A 2A7, Canada

## References

1 Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P.: 'Robust fitting of mixtures using trimmed likelihood estimator', *Comput. Stat. Data. Anal.*, 2007, **52**, (1), pp. 299–308

2 Schwarz, G.: 'Estimating the dimension of a model', *Ann. Stat.*, 1978, **6**, (2), pp. 461–464

3 Zhao, Q., and Miller, D.J.: 'A deterministic annealing-based approach for learning and model selection in finite mixture models'. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004, pp. V–457-60

4 Rose, K.: 'Deterministic annealing for clustering, compression, classification, regression, and related optimisation problems', *Proc. IEEE*, 1998, **86**, (11), pp. 2210–2239

5 Grant, M., and Boyd, S.: 'CVX: Matlab software for disciplined convex programming', version 1.21. http://cvxr.com/cvx, 2011

6 Figueiredo, M.A.T., and Jain, A.K.: 'Unsupervised learning of finite mixture models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (3), pp. 381–396

7 BCI competition II. Available: http://www.bbci.de/competition/ii/