with the estimation of the parameters. One possible approach is to choose a best k^* by the Akaike's information criterion [1] or its extensions. But the process of evaluating a criterion incurs a large computational cost since we need to repeat the entire parameter learning process at a number of different values of k.

Proposed in 1995 [2] and systematically developed in past years [3–5], Bayesian Ying–Yang (BYY) harmony learning acts as a general statistical learning framework not only for understanding several existing major learning approaches but also for tackling the learning problem with a new learning mechanism that makes model selection automatically during parameter learning. In the following, we implement this mechanism on a bi-directional architecture (BI-architecture) of the BYY system via a gradient learning rule to solve the Gaussian mixture modelling problem.

2. Gradient learning rule

A BYY system describes each observation $x \in \mathcal{X} \subset R^n$ and its corresponding inner representation $y \in \mathcal{Y} \subset R^m$ via the two types of Bayesian decomposition of the joint density $p(x, y) = p(x)\,p(y|x)$ and $q(x, y) = q(x|y)q(y)$, being called Yang and Ying machine, respectively. In this paper, y is only limited to be an integer variable, i.e., $y \in \mathcal{Y} = \{1, 2, \ldots, k\} \subset R$ with $m = 1$. Given a data set $D_x = \{x_t\}_{t=1}^N$, the task of learning on a BYY system consists of specifying all the aspects of $p(y|x), p(x), q(x|y), q(y)$ with a harmony learning principle implemented by maximizing the functional

$$H(p\|q) = \int p(y|x)\,p(x)\ln[q(x|y)q(y)]\,\mathrm{d}x\,\mathrm{d}y - \ln z_q, \tag{1}$$

where z_q is a regularization term. The details are referred to [3].

If both $p(y|x)$ and $q(x|y)$ are parametric, i.e., from a family of probability densities with a parameter $\theta \in R^d$, the BYY system is called to have a Bi-directional Architecture (BI-Architecture). For Gaussian mixture modelling, we use the following specific BI-architecture of the BYY system. $q(j) = \alpha_j$ with $\alpha_j \geqslant 0$ and $\sum_{j=1}^k \alpha_j = 1$. Also, we ignore the regularization term z_q (i.e., set $z_q = 1$) and let $p(x)$ be the empirical density $p_0(x) = (1/N)\sum_{t=1}^N \delta(x - x_t)$, where $x \in \mathcal{X} = R^n$. Moreover, the BI-architecture is constructed with the following parametric form:

$$p(y = j|x) = \frac{\alpha_j q(x|\theta_j)}{q(x|\Theta_k)}, \quad q(x|\Theta_k) = \sum_{j=1}^k \alpha_j q(x|\theta_j), \tag{2}$$

where $q(x|\theta_j) = q(x|y = j)$ with θ_j consisting of all its parameters and $\Theta_k = \{\alpha_j, \theta_j\}_{j=1}^k$. Substituting these component densities into Eq. (1), we have

$$H(p\|q) = J(\Theta_k) = \frac{1}{N}\sum_{t=1}^N \sum_{j=1}^k \frac{\alpha_j q(x_t|\theta_j)}{\sum_{i=1}^k \alpha_i q(x_t|\theta_i)} \ln[\alpha_j q(x_t|\theta_j)]. \tag{3}$$

That is, $H(p\|q)$ becomes a harmony function $J(\Theta_k)$ on the parameters Θ_k of a finite mixture model, which was originally introduced in [2] as $J(k)$ and developed into this form in [3] using as a selection criterion of the number k. Letting $q(x|\theta_j)$ be a Gaussian density given by

$$q(x|\theta_j) = q(x|m_j, \Sigma_j) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \mathrm{e}^{-(1/2)(x-m_j)^{\mathrm{T}}\Sigma_j^{-1}(x-m_j)}, \tag{4}$$

where m_j is the mean vector and Σ_j is the covariance matrix, and $\alpha_j = e^{\beta_j}/\sum_{i=1}^{k} e^{\beta_i}$ for $j = 1, 2, \ldots, k$ with $-\infty < \beta_1, \ldots, \beta_k < +\infty$. By the derivatives of $J(\Theta_k)$ with respect to β_j, m_j and Σ_j, we have the following gradient learning rule:

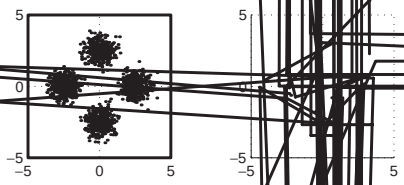
$$\Delta\beta_j = \eta \frac{\alpha_j}{N} \sum_{i=1}^{k} \sum_{t=1}^{N} h(i|x_t) U(i|x_t)(\delta_{ij} - \alpha_i), \tag{5}$$

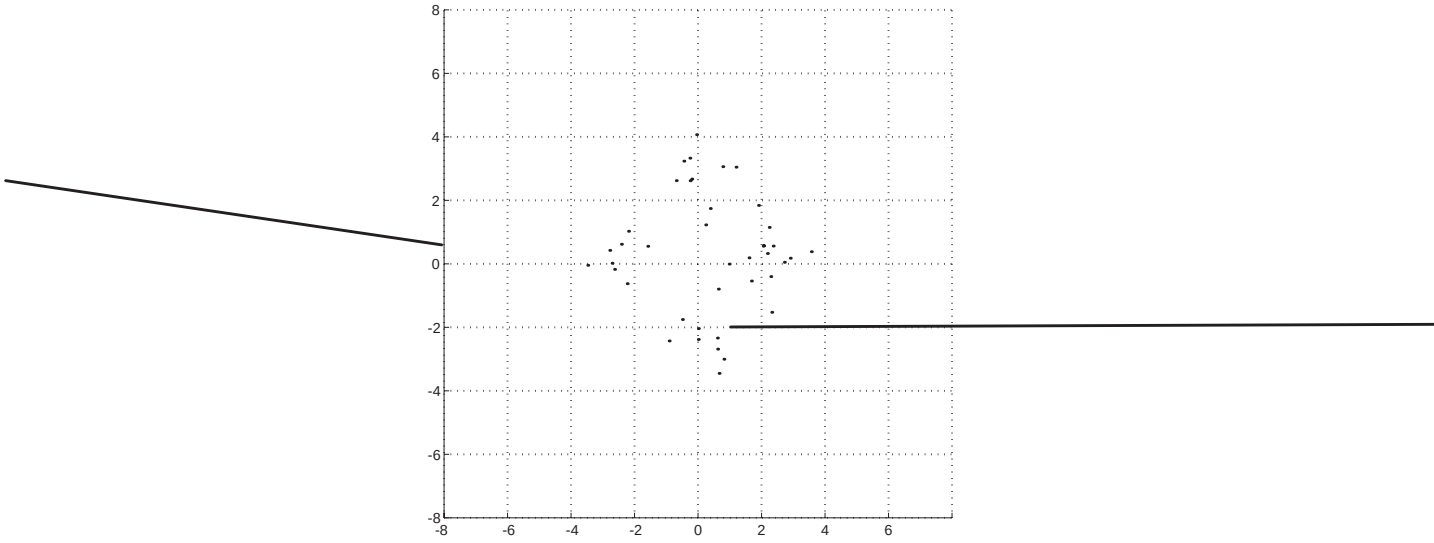
$$\Delta m_j = \eta \frac{\alpha_j}{N} \sum_{t=1}^{N} h(j|x_t) U(j|x_t) \Sigma_j^{-1}(x_t - m_j), \tag{6}$$

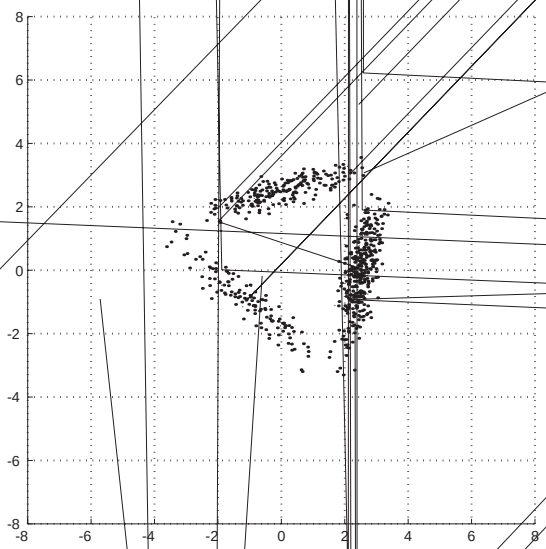
$$\Delta \Sigma_j = \eta \frac{\alpha_j}{2N} \sum_{t=1}^{N} h(j|x_t) U(j|x_t) \Sigma_j^{-1}[(x_t - m_j)(x_t - m_j)^T - I]\Sigma_j^{-1}, \tag{7}$$

where

$$U(i|x_t) = \sum_{r=1}^{k} (\delta_{ri} - p(r|x_t)) \ln \alpha_r q(x_t|\theta_r) + 1$$


5
0
−5
−5
0
5
5
0
−5
−5
5


8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6


8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8

average error between the estimated parameters and the true parameters being less than 0.1.

Furthermore, we tested the gradient learning rule for clustering on some sample data sets in which each cluster is not subject to a Gaussian. The experiment results have shown that the correct number of clusters can be still detected when those clusters can be separated in the similar degree as above. Also, under the principle of the maximum posteriori probability $p(j|x_t)$ of the converged parameters Θ_k, the clustering result is generally as good as the k-means algorithm with $k = k^*$. However, when two or more clusters are joined together like iris data, the gradient learning rule can only find out the separated clusters in the sample data set.

4. Conclusions

The automatic model selection feature of BYY harmony learning has been demonstrated on Guassian mixture modelling with a BI-architecture of the BYY system. In help of the gradient learning rule derived, a number of experiments have demonstrated that as long as the overlap among the Gaussians or clusters in a data set is not too serious, the number of Gaussians can be correctly detected automatically during learning with a good estimation on parameters of each Gaussian component density, even on a data set of a small sample size.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control AC-19 (6) (1974) 716–723.

[2] L. Xu, Ying–Yang machine: a Bayesian–Kullback scheme for unified learnings and new results on vector quantization, Proceedings of the 1995 International Conference on Neural Information Processing, ICONIP'95, Vol. 2, Beijing, China, 30 October–3 November 1995, pp. 977–988.

[3] L. Xu, Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, three-layer nets and ME-RBF-SVM models, Internat. J. Neur. Syst. 11 (1) (2001) 43–69.

[4] L. Xu, Ying–Yang learning, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, 2nd Edition, The MIT Press, Cambridge, MA, 2002, pp. 1231–1237.

[5] L. Xu, BYY harmony learning, structural RPCL, and topological self-organizing on mixture modes, Neur. Networks 15 (8–9) (2002) 1231–1237.

[6] L. Xu, A. Krzyzak, E. Oja, Rival penalized competitive learning for clustering analysis, RBF net, and curve detection, IEEE Trans. Neur. Networks 4 (4) (1993) 636–648.