Two Further Gradient BYY Learning Rules for Gaussian Mixture with Automated Model Selection*

Jinwen Ma, Bin Gao, Yang Wang, and Qiansheng Cheng

Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China
jwma@math.pku.edu.cn

Abstract. Under the Bayesian Ying-Yang (BYY) harmony learning theory, a harmony function has been developed for Gaussian mixture model with an important feature that, via its maximization through a gradient learning rule, model selection can be made automatically during parameter learning on a set of sample data from a Gaussian mixture. This paper proposes two further gradient learning rules, called conjugate and natural gradient learning rules, respectively, to efficiently implement the maximization of the harmony function on Gaussian mixture. It is demonstrated by simulation experiments that these two new gradient learning rules not only work well, but also converge more quickly than the general gradient ones.

1 Introduction

As a powerful statistical model, Gaussian mixture has been widely applied to data analysis and there have been several statistical methods for its modelling(e.g., the expectation-maximization(EM) algorithm [1] and k-means algorithm [2]). But it is usually assumed that the number of Gaussians in the mixture is pre-known. However, in many instances this key information is not available and the selection of an appropriate number of Gaussians must be made with the estimation of the parameters, which is rather difficult [3].

The traditional approach is to choose a best number k^* of Gaussians via some selection criterion. Actually, many heuristic criteria have been proposed in the statistical literature(e.g.,[4]-[5]). However, the process of evaluating a criterion incurs a large computational cost since we need to repeat the entire parameter estimating process at a number of different values of k.

Recently, a new approach has been developed from the Bayesian Ying-Yang (BYY) harmony learning theory [6] with the feature that model selection can be made automatically during the parameter learning. In fact, it was shown in [7] that this Gaussian mixture modelling problem is equivalent to the maximization of a harmony function on a specific architecture of the BYY system related

* This work was supported by the Natural Science Foundation of China for Project 60071004.

Z.R. Yang et al. (Eds.): IDEAL 2004, LNCS 3177, pp. 690–695, 2004.

to Gaussian mixture model and a gradient learning rule for maximization of this harmony function was also established. The simulation experiments showed that an appropriate number of Gaussians can be automatically allocated for the sample data set, with the mixing proportions of the extra Gaussians attenuating to zero. Moreover, an adaptive gradient learning rule was further proposed and analyzed for the general finite mixture model, and demonstrated well on a sample data set from Gaussian mixture [8].

In this paper, we propose two further gradient learning rules to efficiently implement the maximization of the harmony function in a Gaussian mixture setting. The first learning rule is constructed from the conjugate gradient of the harmony function, while the second learning rule is derived from Amari and Nagaoka's natural gradient theory [9]. Moreover, it has been demonstrated by simulation experiments that the two new gradient learning rules not only make model selection automatically during the parameter learning, but also converge more quickly than the general gradient ones.

In the sequel, the conjugate and natural gradient learning rules are derived in Section 2. In Section 3, they are both demonstrated by simulation experiments, and finally a brief conclusion is made in Section 4.

2 Conjugate and Natural Gradient Learning Rules

In this section, we first introduce the harmony function on the Gaussian mixture model and then derive the conjugate and natural gradient learning rules from it.

2.1 The Harmony Function

Under the BYY harmony learning principle, we can get the following harmony function on a sample data set $D_x = \{x_t\}_{t=1}^N$ from a Gaussian mixture model (Refer to [6] or [7] for the derivation):

$$J(\Theta_k) = \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{k}\frac{\alpha_j q(x_t|m_j,\Sigma_j)}{\sum_{i=1}^{k}\alpha_i q(x_t|m_i,\Sigma_i)} ln[\alpha_j q(x_t|m_j,\Sigma_j)], \tag{1}$$

where $q(x|m_j,\Sigma_j)$ is a Gaussian density given by

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

where m_j is the mean vector and Σ_j is the covariance matrix which is assumed positive definite. α_j is the mixing proportion, $\Theta_k = \{\alpha_j, m_j, \Sigma_j\}_{j=1}^k$ and $q(x,\Theta_k) = \sum_{j=1}^k \alpha_j q(x|m_j,\Sigma_j)$ is just the Gaussian mixture density.

According to the best harmony learning principle of the BYY system [6] as well as the experimental results obtained in [7]-[8], the maximization of $J(\Theta_k)$ can realize the parameter learning with automated model selection on a sample data set from a Gaussian mixture. For convenience of analysis, we

let $\alpha_j = e^{\beta_j}/\sum_{i=1}^{k} e^{\beta_i}$ and $\Sigma_j = B_j B_j^T$ for $j = 1, 2, \cdots, k$, where $-\infty < \beta_1, \cdots, \beta_k < +\infty$, and B_j is a nonsingular square matrix. By these transformations, the parameters in $J(\Theta_k)$ turn into $\{\beta_j, m_j, B_j\}_{j=1}^{k}$.

2.2 Conjugate Gradient Learning Rule

We begin to give the derivatives of $J(\Theta_k)$ with respect to β_j, m_j and B_j as follows. (Refer to [7] for the derivation.)

$$\frac{\partial J(\Theta_k)}{\partial \beta_j} = \frac{\alpha_j}{N} \sum_{i=1}^{k} \sum_{t=1}^{N} h(i|x_t) U(i|x_t)(\delta_{ij} - \alpha_i), \tag{3}$$

$$\frac{\partial J(\Theta_k)}{\partial m_j} = \frac{\alpha_j}{N} \sum_{t=1}^{N} h(j|x_t) U(j|x_t) \Sigma_j^{-1}(x_t - m_j), \tag{4}$$

$$\frac{\partial J(\Theta_k)}{\partial B_j} = \frac{\partial (B_j B_j^T)}{\partial B_j} \frac{\partial J(\Theta_k)}{\partial \Sigma_j}, \tag{5}$$

where δ_{ij} is the Kronecker function, and

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

$$h(i|x_t) = \frac{q(x_t|m_i, \Sigma_i)}{\sum_{r=1}^{k} \alpha_r q(x_t|m_r, \Sigma_r)}, \qquad p(i|x_t) = \alpha_i h(i|x_t),$$

$$\frac{\partial J(\Theta_k)}{\partial \Sigma_j} = \frac{\alpha_j}{N} \sum_{t=1}^{N} h(j|x_t) U(j|x_t) \Sigma_j^{-1}[(x_t - m_j)(x_t - m_j)^T - \Sigma_j] \Sigma_j^{-1}.$$

In Eq. (5), B_j and $\frac{\partial J(\Theta_k)}{\partial \Sigma_j}$ are considered as their vector forms, i.e., $vec[B_j]$ and $vec[\frac{J(\Theta_k)}{\partial \Sigma_j}]$, respectively. $\frac{\partial (B_j B_j^T)}{\partial B_j}$ is an n^2-order square matrix which can be easily computed.

Combining these β_j, m_j, and B_j into a vector θ_k, we have the conjugate gradient learning rule as follows:

$$\theta_k^{i+1} = \theta_k^i + \eta \widehat{s_i}, \tag{6}$$

where η is the learning rate, and the searching direction $\widehat{s_i}$ is computed from the following iterations of conjugate vectors:

$$s_1 = -\nabla J(\theta_k^1), \quad \widehat{s_1} = \frac{-\nabla J(\theta_k^1)}{\|\nabla J(\theta_k^1)\|}$$

$$s_i = -\nabla J(\theta_k^i) + v_{i-1} s_{i-1}, \quad \widehat{s_i} = \frac{s_i}{\|s_i\|}, \quad v_{i-1} = \frac{\|\nabla J(\theta_k^i)\|^2}{\|\nabla J(\theta_k^{i-1})\|^2},$$

where $\nabla J(\theta_k)$ is the general gradient vector of $J(\theta_k) = J(\Theta_k)$ and $\|\cdot\|$ is the Euclidean norm.

2.3 Natural Gradient Learning Rule

In order to get the natural gradient of $J(\Theta_k)$, we let $\theta_j = (m_j, B_j) = (m_j, \Sigma_j)$ so that $\Theta_k = \{\alpha_j, \theta_j\}_{j=1}^k$ which can be considered as a point in the Riemann space. Then, we can construct a $k(n^2+n+1)$-dimensional statistical model $S = \{q(x, \Theta_k) : \Theta_k \in \Xi\}$. The Fisher information matrix of S at a point Θ_k is $G(\Theta_k) = [g_{ij}(\Theta_k)]$, where $g_{ij}(\Theta_k)$ is given by

$$g_{ij}(\Theta_k) = \int \partial_i l(x, \Theta_k) \partial_j l(x, \Theta_k) q(x, \Theta_k) dx, \tag{7}$$

where $\partial_i = \frac{\partial}{\partial \Theta_k^i}$ and $l(x, \Theta_k) = lnq(x, \Theta_k)$. According to the following derivatives:

$$\frac{\partial q(x_t|\Theta_k)}{\partial \beta_j} = \alpha_j q(x_t|\theta_j) \sum_{i=1}^{k} (\delta_{ij} - \alpha_i), \tag{8}$$

$$\frac{\partial q(x_t|\Theta_k)}{\partial m_j} = \alpha_j q(x_t|\theta_j) \Sigma_j^{-1}(x_t - m_j), \tag{9}$$

$$\frac{\partial q(x_t|\Theta_k)}{\partial B_j} = \frac{\partial B_j^T B_j}{\partial B_j} \alpha_j q(x_t|\theta_j) \Sigma_j^{-1}[(x_t - m_j)(x_t - m_j)^T - \Sigma_j]\Sigma_j^{-1}, \tag{10}$$

we can easily estimate $G(\Theta_k)$ on a sample data set through the law of large number. According to Amari and Nagaoka's natural gradient theory [9], we have the following natural gradient learning rule:

$$\Theta_k(m+1) = \Theta_k(m) - \eta G^{-1}(\Theta_k(m)) \frac{\partial J(\Theta_k(m))}{\partial \Theta_k}, \tag{11}$$

where η is the learning rate.

3 Simulation Experiments

We conducted experiments on seven sets (a)-(g) of samples drawn from a mixture of four or three bivariate Gaussians densities (i.e., $n = 2$). As shown in Figure 1, each data set of samples consists three or four Gaussians with certain degree of overlap. Using k^* to denote the number of Gaussians in the original mixture, we implemented the conjugate and natural gradient learning rules on those seven sample data sets always with $k^* \leq k \leq 3k^*$ and $\eta = 0.1$. Moreover, the other parameters were initialized randomly within certain intervals. In all the experiments, the learning was stopped when $|J(\Theta_k^{new}) - J(\Theta_k^{old})| < 10^{-5}$.

The experimental results of the conjugate and natural gradient learning rules on the data sets (c) and (d) are given in Figures 2 & 3, respectively, with case $k = 8$ and $k^* = 4$. We can observe that four Gaussians are finally located accurately, while the mixing proportions of the other four Gaussians were reduced to below 0.01, i.e, these Gaussians are extra and can be discarded. That is, the correct number of the clusters have been detected on these data sets. Moreover, the

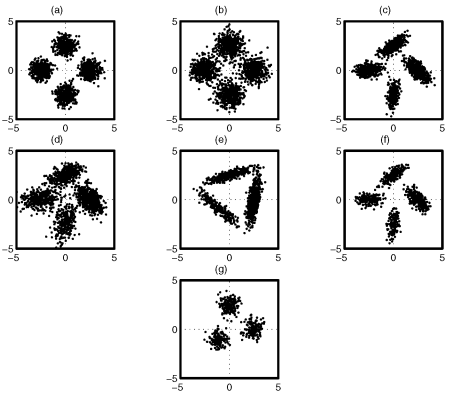

Fig. 1. Seven sets of sample data used in the experiments.

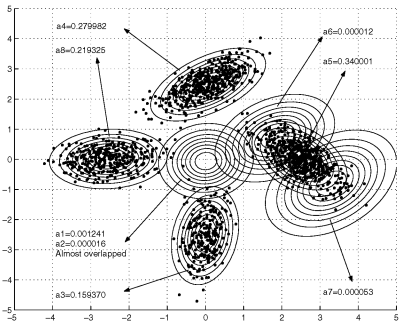

Fig. 2. The experimental result of the conjugate gradient learning rule on the data set (c) (stopped after 63 iterations).

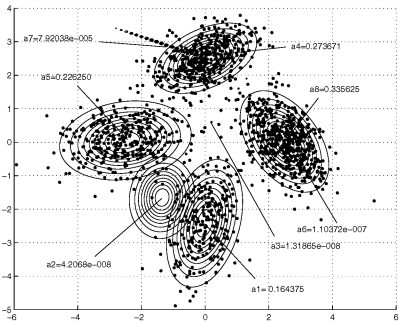

Fig. 3. The experimental result of the natural gradient learning rule on the data set (d) (stopped after 149 iterations).

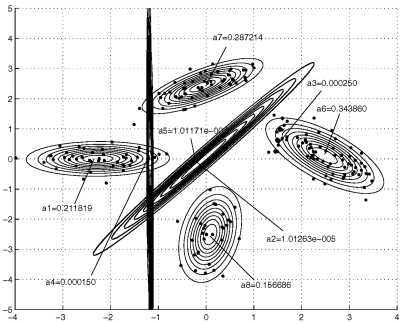

Fig. 4. The experimental result of the natural gradient learning rule on the data set (f) (stopped after 173 iterations).

experiments of the two learning rules have been made on the other sample data sets in different cases and show the similar results on automated model selection. For example, the natural gradient learning rule was implemented on the data set (f) with $k = 8, k^* = 4$. As shown in Figure 4, even each cluster has a small number of samples, the correct number of clusters can still be detected, with the mixing proportions of other four extra Gaussians reduced below 0.01.

In addition to the correct number detection, we further compared the converged values of parameters (discarding the extra Gaussians) with those parameters in the mixture from which the samples come from. We checked the results in these experiments and found that the conjugate and natural gradient learning rules converge with a lower average error between the estimated parameters and the true parameters.

In comparison with the simulation results of the batch and adaptive gradient learning rules [7]-[8] on these seven sets of sample data, we have found that the conjugate and natural gradient learning rules converge more quickly than the two general gradient ones. Actually, for the most cases it had been demonstrated by simulation experiments that the number of iterations required by each of these two rules is only about one tenth to a quarter of the number of iterations required by either the batch or adaptive gradient learning rule.

As compared with each other, the conjugate gradient learning rule converges more quickly than the natural gradient learning rule, but the natural gradient learning rule obtains a more accurate solution on the parameter estimation.

4 Conclusion

We have proposed the conjugate and natural gradient learning rules for the BYY harmony learning on Gaussian mixture with automated model selection. They are derived from the conjugate gradient method and Amari and Nagaoka's natural gradient theory for the maximization of the harmony function defined on Gaussian mixture model. The simulation experiments have demonstrated that both the conjugate and natural learning rules lead to the correct selection of the number of actual Gaussians as well as a good estimate for the parameters of the original Gaussian mixture. Moreover, they converge more quickly than the general gradient ones.

References

1. R. A. Render and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol.26, no.2, pp. 195-239, 1984.
2. A. K. Jain and R. C. Dubes, *Algorithm for Clustering Data,*, Englewood Cliffs, N. J.: Prentice Hall, 1988.
3. J. A. Hartigan, "Distribution problems in clustering," *Classification and clustering*, J. Van Ryzin Eds., pp. 45-72, New York: Academic press, 1977.
4. H. Akaike. "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control,* vol. AC-19, pp. 716-723, 1974.
5. G. W. Millgan and M. C. Copper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol.46, pp: 187-199, 1985.
6. 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," *International Journal of Neural Systems*, vol.11, no.1, pp. 43-69, February, 2001.
7. J. Ma, T. Wang and L. Xu, "A gradient BYY harmony learning rule on Gaussian mixture with automated model selection," *Neurocomputing*, vol.56, pp: 481-487, 2004.
8. J. Ma, T. Wang and L. Xu, "An adaptive BYY harmony learning algorithm and its relation to rewarding and penalizing competitive learning mechanism," *Proc. of ICSP'02*, 26-30 Aug. 2002, vol.2, pp: 1154 - 1158.
9. S. Amari and H. Nagaoka, *Methods of Information Geometry*, AMS, Oxford University Press, 2000.