Digital Image Processing Laboratory: The EM Algorithm

January 7, 2007

1 Introduction

This laboratory explores the use of the expectation-maximization (EM) algorithm for the estimate of parameters. In particular, we will use the EM algorithm for two applications: clustering, and hidden Markov model training. You are encouraged to implement your solutions to this laboratory in Matlab.

For an original derivation of the monotone convergence of the likelihood for the EM algorithm, the reader is encouraged to read [1]. The publication [2] has a clear treatment of the application to the EM algorithm to Gaussian mixtures and to novel approach to the problem of order identification. The paper [3] is a widely cited tutorial reference, and the papers [3] and [4] present valuable analyses of the convergence properties of the EM algorithm.

The EM algorithm is very useful for computing maximum likelihood (ML) estimates of parameters when observations are incomplete. Consider the case when it is necessary to estimate the parameter vector $\theta \in \Omega$ from the complete data Y and X. The ML estimate of θ is given by

$$\hat{\theta} = \arg\max_{\theta\in\Omega} \log p(y, x|\theta) \ .$$

In some cases, it may not be possible to observe X. In this case, we refer to Y as the incomplete data, since it is an incomplete observation; and the ML estimate of θ given y has the form

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta\in\Omega} \log p(y|\theta) \\ &= \arg\max_{\theta\in\Omega} \log \int_x p(y, x|\theta)\, dx \ . \end{aligned} \tag{1}$$

In practice, (1) may be difficult to compute due to the required integral over x. The EM algorithm is a method for solving (1) which works through iterative optimization of a function

$$Q(\theta', \theta) = E\left[\log p(y, X|\theta')|Y = y, \theta\right] \tag{2}$$

using the recursion

$$\theta'' = \arg\max_{\theta'\in\Omega} Q(\theta', \theta) \ . \tag{3}$$

Questions or comments concerning this laboratory should be directed to Prof. Charles A. Bouman, School of Electrical and Computer Engineering, Purdue University, West Lafayette IN 47907; (765) 494-0340; bouman@ecn.purdue.edu

It can be shown that using the update formulat of (3), the following two properties hold.

$$\begin{aligned} \log p(y|\theta'') - \log p(y|\theta) &\geq Q(\theta'', \theta) - Q(\theta, \theta) \quad (4) \\ \nabla_\theta \log p(y|\theta) &= \nabla_{\theta'} Q(\theta', \theta)|_{\theta'=\theta} \quad (5) \end{aligned}$$

These two properties guarrentee that EM updates produce a monotone increasing sequence of log likelihood values, and that fixed points of the EM algorithm must correspond to points where the log likelihoods gradient is zero.

When the joint distribution of Y and X is exponential, as is often the case, the EM update has a particularly simple structure to it.

- Definition: A family of density functions $p(y, x|\theta)$ is said to be a *k-parameter exponential family* if there exist functions $g(\theta) \in I\!R^k$, $s(y, x)$, $d(\theta)$ and statistic $T(y, x) \in I\!R^k$ such that
$$p(y, x|\theta) = \exp\{< g(\theta), T(y, x) > + d(\theta) + s(y, x)\} \quad (6)$$
for all (y, x) and $\theta \in \Omega$ where $< \cdot, \cdot >$ denotes the inner product. We refer to $T(y, x)$ as the *natural sufficient statistic* or *natural statistic* for the exponential distribution.

Assume that the joint density of (Y, X) is from an exponential family. Then we know that

$$p(y, x|\theta) = \exp\{< g(\theta), T(y, x) > + d(\theta) + s(y, x)\}$$

where $T(y, x)$ is the natural statistic from the distribution. Assuming the ML estimate of θ exists and is unique, then it is given by

$$\begin{aligned} \theta_{ML} &= \arg\max_{\theta \in \Omega} \{< g(\theta), T(y, x) > + d(\theta)\} \quad (7) \\ &= f(T(y, x)) \quad (8) \end{aligned}$$

where $f(\cdot)$ is some function of the k dimensional suffient statistic for the exponential density.

Since our objective is to maximize Q with respect to θ', we only need to know the function Q within a constant that is not dependent on θ'. Therefore, we have

$$\begin{aligned} Q(\theta', \theta) &= E\left[\log p(y, X|\theta')|Y = y, \theta\right] \\ &= E\left[< g(\theta), T(y, X) > + d(\theta) + s(y, X)|Y = y, \theta\right] \\ &= < g(\theta'), \bar{T} > + d(\theta') + constant \end{aligned}$$

were

$$\bar{T} = E\left[T(y, X)|Y = y, \theta\right]$$

is the conditional expectation of the statistic $T(y, x)$. A single update of the EM algorithm is then given by the recursion

$$\begin{aligned} \theta'' &= \arg\max_{\theta' \in \Omega} Q(\theta', \theta) \quad (9) \\ &= \arg\max_{\theta' \in \Omega} < g(\theta'), \bar{T} > + d(\theta') \\ &= f(\bar{T}) \end{aligned}$$

Intuitively, we see that the EM update has the same form as the computation of the ML estimate, but with the expected value of the statistic replacing the actual statistic.

2 Parameter Estimation for Multivariate Gaussian Distributions

In this section, we will apply the EM algorithm to estimate the parameters of a multivariate Gaussian mixture distribution. This results in an algorithm that can be used for multivariate clustering of data.

Let X_n be a sequence of N discrete i.i.d. random variables with probability mass function

$$P\{X_n = k\} = \pi_k$$

for $k = 0, \cdots, K-1$ where $\sum_{k=0}^{K-1} \pi_k = 1$. Furthermore, let $Y_n \in I\!R^M$ be a sequence of N multivariate Gaussian random vectors which are conditionally i.i.d. given X, and let the distribution of Y_n be given by $N(\mu_{X_n}, R_{X_n})$ where $\mu_k \in I\!R^M$ and $R_k \in I\!R^M \times I\!R^M$ for $k = 0, \cdots, K-1$. Then the joint density of (Y, X) is given by

$$p(y, x|\theta) = \prod_{n=0}^{N-1} \frac{\pi_{x_n}}{(2\pi)^{M/2}} |R_{x_n}|^{-1/2} \exp\left\{-\frac{1}{2}(y_n - \mu_{x_n})^t R_{x_n}^{-1}(y_n - \mu_{x_n})\right\} \tag{10}$$

where the parameter vector of the distribution is given by

$$\theta = [\pi_1, \mu_1, R_1, \cdots, \pi_{K-1}, \mu_{K-1}, R_{K-1}] \ . \tag{11}$$

It may be shown that this is an exponential distribution with natural sufficient statistics

$$N_k = \sum_{n=0}^{N-1} \delta(x_n - k) \tag{12}$$

$$t_{1,k} = \sum_{n=0}^{N-1} y_n \delta(x_n - k) \tag{13}$$

$$t_{2,k} = \sum_{n=0}^{N-1} y_n y_n^t \delta(x_n - k) \tag{14}$$

where $k \in \{0, \cdot, K-1\}$ and $\delta(\cdot)$ is a Kroniker delta function. It may also be shown that the ML parameter estimate of θ given (Y, X) is

$$\hat{\mu}_k = \frac{t_{1,k}}{N_k} \tag{15}$$

$$\hat{R}_k = \frac{t_{2,k}}{N_k} - \frac{t_{1,k} t_{1,k}^t}{N_k^2} \tag{16}$$

$$\hat{\pi}_k = \frac{N_k}{N} \ . \tag{17}$$

Section Problems:

1. Show that the density function of (10) with parameters θ of (11) forms an exponential family with natural sufficient statistics given in (12), (13), and (14).
2. Show that the ML estimate of θ given (Y, X) is formed by the expressions in (15), (16), and (17). (Hint: This is not easy. You will probably need some matrix trace identities from a good text book.)

3 Parameter Estimation for Gaussian Mixture Distributions

In Section (2), we derived the joint distribution for (Y, X) where Y_n is distributed as $N(\mu_{X_n}, R_{X_n})$ and $P\{X_n = k\} = \pi_k$. However, if we do not know the values X_n (i.e. the state or class of each observation) then the distribution of Y_n is known as a Gaussian mixture distribution and may be computed by summing over the variables x_n. We first derive the density function for the entire sequence of observations Y.

$$p(y|\theta) = \sum_{x_0=0}^{K-1}\sum_{x_1=0}^{K-1}\cdot\sum_{x_{N-1}=0}^{K-1}\prod_{n=0}^{N-1}\frac{\pi_{x_n}}{(2\pi)^{M/2}}|R_{x_n}|^{-1/2}\exp\left\{-\frac{1}{2}(y_n-\mu_{x_n})^t R_{x_n}^{-1}(y_n-\mu_{x_n})\right\} \quad (18)$$

$$= \prod_{n=0}^{N-1}\sum_{x_n=0}^{K-1}\frac{\pi_{x_n}}{(2\pi)^{M/2}}|R_{x_n}|^{-1/2}\exp\left\{-\frac{1}{2}(y_n-\mu_{x_n})^t R_{x_n}^{-1}(y_n-\mu_{x_n})\right\} \quad (19)$$

$$= \prod_{n=0}^{N-1}\sum_{k=0}^{K-1}\frac{\pi_k}{(2\pi)^{M/2}}|R_k|^{-1/2}\exp\left\{-\frac{1}{2}(y_n-\mu_k)^t R_k^{-1}(y_n-\mu_k)\right\} \ . \quad (20)$$

Notice that distribution of Y_n is then given by a single term of (20).

$$p(y_n|\theta) = \sum_{k=0}^{K-1}\frac{\pi_k}{(2\pi)^{M/2}}|R_k|^{-1/2}\exp\left\{-\frac{1}{2}(y_n-\mu_k)^t R_k^{-1}(y_n-\mu_k)\right\} \ . \quad (21)$$

Each of the k terms in a Gaussian mixture distribution represent an individual cluster with mean μ_k and covariance R_k. The weight π_k determines the probability that that cluster occurs.

Our objective is to estimate the parameters of this Gaussian mixture; however the ML estimate is difficult to compute due to the structure of the density function. In order to simplify the estimation process, we may use the EM algorithm. We know from Section 2 that the statistics given in (12), (13), and (14) are the natural sufficient statistics for the exponential distribution $p(y, x|\theta)$. Therefore, we may use the update relationship of (9). That is, we may apply the standard ML estimate equations of (15), (16), and (17), but with the expected values of the statics replacing the unknown values.

Using this approach results in the following EM algorithm updates.

- E-step: Compute expected values of statistics

$$\bar{N}_k \leftarrow \sum_{n=0}^{N-1} P\{X_n = k|Y = y, \hat{\theta}\} \quad (22)$$

$$\bar{t}_{1,k} \leftarrow \sum_{n=0}^{N-1} y_n P\{X_n = k|Y = y, \hat{\theta}\} \quad (23)$$

$$\bar{t}_{2,k} \leftarrow \sum_{n=0}^{N-1} y_n y_n^t P\{X_n = k|Y = y, \hat{\theta}\} \quad (24)$$

- M-step: Update parameter vector $\hat{\theta}$

$$\hat{\mu}_k \leftarrow \frac{t_{1,k}}{N_k} \tag{25}$$

$$\hat{R}_k \leftarrow \frac{t_{2,k}}{N_k} - \frac{t_{1,k}t_{1,k}^t}{N_k^2} \tag{26}$$

$$\hat{\pi}_k \leftarrow \frac{N_k}{N} \; . \tag{27}$$

Section Problems:

1. Use the Matlab program *mk_data.m* to create 500 samples from a Gaussian mixture with $M = 2$, $K = 3$, $\pi = [0.4, 0.4, 0.2]$,

$$\mu_0 = [2, 2]^t \tag{28}$$

$$\mu_1 = [-2, -2]^t \tag{29}$$

$$\mu_2 = [5.5, 2]^t \tag{30}$$

   and

$$R_0 = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix} \tag{31}$$

$$R_1 = \begin{bmatrix} 1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \tag{32}$$

$$R_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 0.5 \end{bmatrix} \; . \tag{33}$$

2. Print out a scatter plot of the samples generated in step 1. Circle and label each of the three clusters in the mixture distribution.

3. Derive an explicit expression for $P\{X_n = k|Y = y, \hat{\theta}\}$ used in the E-step.

4. Implement the EM algorithm for computing the ML estimate of θ. Use the initial value for θ of $\pi \leftarrow [1/3, 1/3, 1/3]$,

$$R_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ for } k = 0, 1, 2$$

   and select μ_1, μ_2, and μ_3 as the first three sample vectors produced in step 1 above.

5. Run 20 iterations of the EM algorithm, and print out the values of the estimated parameters. Do the indices k of the estimated and true parameters correspond?

6. Find the best correspondance of the true and estimated parameters, and print them out in a tabular form comparing their values.

4 Order Identification for Gaussian Mixture Distributions

In this section, you will implement a method for automaticaly determining the number of clusters in a Gaussian mixture model. Software for implementing this algorithm in C is available from [5], but you should implement your solution in Matlab.

In the previous section, we considered the problem of estimating the parameters of a mixture distribution of the form

$$\begin{aligned}\log p_y(y|K,\theta) &= \sum_{n=0}^{N-1} \log p(y_n|\theta) \\ &= \sum_{n=0}^{N-1} \log \left(\sum_{i=0}^{K-1} p_{y_n|x_n}(y_n|\theta_k)\pi_k \right)\end{aligned}$$

were $\theta_k = (\pi_k, \mu_k, R_k)$ is the parameter vector associated with the k^{th} cluster.

However, in many applications the number of clusters may be unknown; so it is also necesary to estimate K. One might consider estimating the model order using ML estimation.

$$\hat{K} = \arg \max_{K>0} \max_{\theta \in \Omega^{(K)}} \log p_y(y|K,\theta)$$

However, this strategy is fundamentally flawed because larger values of $\hat{K}$ can only increase the value of the likelihood. This results from the fact that for every model specified by $\theta \in \Omega^{(K)}$ there exists a model $\theta' \in \Omega^{(K+1)}$ which specifies the same distribution. In fact, θ' can be constructed from θ in many ways. For example, it can be formed by adding an additional $(N+1)^{th}$ cluster with $\pi_(N+1) = 0$. Therefore, we can see that

$$\max_{\theta \in \Omega^{(K+1)}} \log p_y(y|K+1,\theta) \geq \max_{\theta \in \Omega^{(K)}} \log p_y(y|K,\theta)$$

which in turn leads to divergence of the ML estimate of K.

This problem of estimating the order of a model is known as order identification, and has been studied by a variety of researchers. Methods for estimating model order generally tend to require the addition of a penalty term in the log likelihood to account for the over-fitting of high order models. One of the earliest approaches to order identification was suggested by Akaike [6], and requires the minimization of the so called AIC information criteria. The AIC criterion is given by

$$AIC(K,\theta) = -2 \log p_y(y|K,\theta) + 2L$$

where L is the number of continuously valued real numbers required to specify the parameter θ. For each cluster, there is 1 parameter for the specification of π_k, M parameters for the specification of μ_k, and $(M+1)M/2$ parameters for the specification of the symetric matrix R_k. In addition, one parameter is redundent because of the fact that the $1 = \sum_{k=0}^{K-1} \pi_k$. Therefore, we have that

$$L = K\left(1 + M + \frac{(M+1)M}{2}\right) - 1 \ .$$

However, an important disadvantage of the AIC criteria is that for a wide class of problems the AIC does not lead to a consistent estimator [7]. This means that as the number of observations tends to infinity, the estimated value for K does not converge to the true value.

Alternatively, another criterion was suggested by Rissanen [8] called the minimum description length (MDL) estimator. This estimator works by attempting to find the model order which minimizes the number of bits that would be required to code both the data samples y_n and the parameter vector θ. While a direct implementation of the MDL estimator may depend on the particular coding method used, Rissanen develop an approximate expression for the estimate based on some assumptions and the minimization of the expression

$$MDL(K,\theta) = -\log p_y(y|K,\theta) + \frac{1}{2}L\log(NM) \ . \tag{34}$$

Notice that the major difference between the AIC and MDL criteria is the dependence of the penalty term on the total number of data values NM. In practice, this is important since otherwise more data will tend to result in over fitting of the model. In fact, it has been shown that for a large number of problems, the MDL criteria is a consistent estimator of model order [9, 10]. Unfortunately, the estimation of model order for mixture models does not fall into the class of problems for which the MDL criteria is known to be consistent. This is due to the fact that the solution to the mixture model problem always falls on a boundary of the constraint space, so the normal results on the asymptotic distribution of the ML estimate are no longer valid. An alternative method for order identification which is known to be consistent for mixture models is presented in [11]. However, this method is computationally expensive when the dimensionality of the data is high.

Nonetheless, our objective in this laboratory will be to estimate the model order by minimizing the MDL criteria given by

$$MDL(K,\theta) = -\sum_{n=0}^{N-1}\log\left(\sum_{i=0}^{K-1} p_{y_n|x_n}(y_n|\theta_k)\pi_k\right) + \frac{1}{2}L\log(NM) \ . \tag{35}$$

From the previous sections, we know that for fixed K

$$MDL(K,\theta) - MDL(K,\theta^{(old)}) < Q(\theta^{(old)};\theta^{(old)}) - Q(\theta;\theta^{(old)}) \ .$$

Therefore, we can use the EM algorithm to minimize the MDL for a fixed model order.

We use the following algorithmic approach to find a value of $\hat{K}$ which approximately minimizes the MDL value.

1. Select an initial value of K and $\theta \in \Omega^{(K)}$
2. Repeat for $k = K$ to 1
   (a) Run the EM algorithm to convergence to produce θ^*
   (b) Record the value of $\text{MDL}(K) \leftarrow \text{MDL}(K,\theta^*)$
   (c) Merge two clusters in θ^* to form a new initial parameter θ
3. Select the value $\hat{K}$ such that $\text{MDL}(\hat{K})$ is minimum.

Typically, it is good to start with a substantially larger value of K than one expects for the final estimate. But the problem remains of how to choose the clusters to merge. Our approach is to select the two clusters which are "closest" to each other. More specifically, we select the clusters l and m which minimize the distance $d(l,m)$ as given below.

$$d(l,m) = \frac{N\hat{\pi}_l}{2}\log\left(\frac{|R_{(l,m)}|}{|\hat{R}_l|}\right) + \frac{N\hat{\pi}_m}{2}\log\left(\frac{|R_{(l,m)}|}{|\hat{R}_m|}\right)$$

where

$$\pi_{(l,m)} = \hat{\pi}_l + \hat{\pi}_m \tag{36}$$

$$\mu_{(l,m)} = \frac{\hat{\pi}_l\hat{\mu}_l + \hat{\pi}_m\hat{\mu}_m}{\hat{\pi}_l + \hat{\pi}_m} \tag{37}$$

$$R_{(l,m)} = \frac{\hat{\pi}_l\left(\hat{R}_l + (\hat{\mu}_l - \mu_{(l,m)})(\hat{\mu}_l - \mu_{(l,m)})^t\right) + \hat{\pi}_m\left(\hat{R}_m + (\hat{\mu}_m - \mu_{(l,m)})(\hat{\mu}_m - \mu_{(l,m)})^t\right)}{\hat{\pi}_l + \hat{\pi}_m} \tag{38}$$

It can be shown that this cluster distance measure has the property that it upper bounds the initial increase in the MDL value that occurs when two clusters are merged [5]. After two clusters are merged, the values of $\pi_{(l,m)}$, $\mu_{(l,m)}$, and $R_{(l,m)}$ defined in equations (36), (37), and (38) are adopted as the parameters for the newly formed cluster.

Section Problems:

1. Use the data generated in problem 1 of the previous Section 3.

2. Implement the MDL order estimation method as described above using and initial value of $K = 9$. Use the initial value for θ of $\pi \leftarrow [1/9, \cdots, 1/9]$,

$$R_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ for } k = 0, \cdots, 9$$

   and select $\mu_1, \mu_2, \cdots, \mu_9$ as the first nine sample vectors produced in step 1 above. For each value of K, run 20 iterations of the EM algorithm.

3. Implement a matlab subroutine that computes the MDL criterium for specified values of K, θ, and y.

4. Plot the value of MDL given in (35) as a function of the total number of EM iterations. For each EM iteration, you should plot the value of the MDL criteria after the iteration is complete. In addition, if the EM iteration is followed by a cluster merging, then you should also plot the value of the MDL criteria after the clusters are merged, and the parameters are updated according to (36), (37), and (38. Noticed that for cluster merging iterations, there will be two ordinate points plotted. Mark the locations on the plot that correspond to the merging of clusters.

5. Plot the MDL value versus K. For each value of K ranging from $K = 1$ to 9, plot the minimum value of the MDL obtained for that number of clusters. Label the value of K corresponding to the minimum observed value of MDL.

6. Does the estimated value of $\hat{K}$ correspond to the true value of $K = 3$?

References

[1] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[2] M. Aitkin and D. B. Rubin, "Estimation and hypothesis testing in finite mixture models," *Journal of the Royal Statistical Society B*, vol. 47, no. 1, pp. 67–75, 1985.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.

[4] C. Wu, "On the convergence properties of the EM algorithm," *Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

[5] C. A. Bouman, "Cluster: An unsupervised algorithm for modeling Gaussian mixtures." Available from http://www.ece.purdue.edu/~bouman, April 1997.

[6] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. AC-19, no. 6, pp. 716–723, December 1974.

[7] R. L. Kashyap, "Inconsistency of the AIC rule for estimating the order of autoregressive models," *IEEE Trans. Automatic Control*, vol. 25, no. 5, pp. 996–998, October 1980.

[8] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, pp. 417–431, September 1983.

[9] R. L. Kashyap, "Optimal choice of ar and ma parts in autoregressive moving average models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, no. 2, pp. 99–104, March 1982.

[10] M. Wax and R. L. Kashyap, "Detection of signals by information theoretic criteria," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. ASSP–33, no. 2, pp. 387–392, April 1985.

[11] E. Redner and H. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, April 1984.