Factorized EM Algorithm for Mixture Estimation

Ivan Nagy[1], Petr Nedoma and Miroslav Kárný [2]

Abstract

A classical version of the EM algorithm is considered in the paper. Its numerical properties are improved using factorized algorithms for maximization in M step of the algorithm. The results are illustrated on simulated examples.

1 Introduction

Expectation and Maximization (EM) algorithm is well known and widely spread procedure for parameter estimation in cases when some of the modeled data can and some cannot be measured [1]. The parameter estimates are computed iteratively with monotonic convergence guaranteed [2].

Estimation of mixture models [3] is a well known theoretical application of the EM algorithm. Here, the measured data are represented by the output of the system. The internal pointers, indicating the active components of the mixture at respective time moments t, are not known. They correspond to the unmeasured part of the data.

Mixture models have a wide practical use as they have universal approximation property: they can be interpreted as approximators of pdfs or approximators based on radial basis functions. In practical applications, they may serve for describing rather large systems with many output variables and rather high number of components. In such cases, however, numerical instabilities occur while manipulating estimates of covariances that are extensively used in each iterative estimation step. These estimates of covariance matrices can easily loose their theoretical positive definiteness. Then, the subsequent estimation produces unbounded values.

The aim of this paper is to utilize numerical procedures adopted from factorized on-line estimation [4, 5] and to combine them with EM algorithm to improve its numerical properties. Indeed, both the single tasks are known and successfully used for a long time. It is astonishing that they were not coupled, yet.

[1]FD ČVUT, Konviktská 20, Prague 1, Czech Republic e-mail: nagy@utia.cas.cz

[2]ÚTIA, AV ČR, Pod vodárenskou věž 4, Prague 8, Czech Republic e-mail: nedoma@utia.cas.cz, school@utia.cas.cz,
The work has been partially supported by EC IST-99-12058

The mostly used way, how to reduce the instabilities mentioned, is to add a small diagonal to the estimated covariances. This solution helps, but it adds an exogenous element to the computations, even if negligible. The solution suggested adds nothing and so it is purer and is hoped to give generally better results.

2 EM algorithm

Let us consider a system generating at time t data $d_t = [y_t, c_t]$. The part y_t of the data is measured and thus known, the rest c_t cannot be measured and thus it is unknown to us. We suppose to have a model of the system, but this model depends also on the unknown part of the data c_t. Consequently, it cannot be directly used in estimation for constructing likelihood function. Instead, the expectation of the logarithmic likelihood taken with respect to unknown data c_t is used for parameter estimation.

2.1 General task

The model of the system is supposed in the form of conditional probability density function (pdf)

$$f(d_t|d_1,\ldots,d_{t-1},\Theta) = f(d_t|\Theta) = f(y_t,c_t|\Theta), \quad \text{where}$$

d_t is the complete data pair $d_t = [y_t, c_t]'$,
y_t are measured data,
c_t are unmeasured data and
Θ is a vector of all unknown parameters of the pdf.

The data y_t are measured at discrete time instances $t = 1, 2, ..., \mathring{t}$. By Y, we denote a vector of all collected measured data $Y = [y_1, y_2, ..., y_{\mathring{t}}]$. Similarly we denote a vector of unmeasured part of data c_t by C and complete data by D. The likelihood is

$$L(\Theta, Y, C) = \prod_{t=1}^{\mathring{t}} f(y_t, c_t|\Theta).$$

As the likelihood contains unknown variables C, it is a random variable. The main trick of the EM algorithm is that it minimizes expectation of its logarithm. The expected value of the log-likelihood, with respect to C, given data Y and current estimate of the parameters Θ, denoted by $\hat{\Theta}$, is

$$\mathcal{L}(\Theta, Y) \quad = \quad E[\log L(\Theta, Y, C)|Y, \hat{\Theta}] =$$

$$= \sum_{C \in C^*} \log L(\Theta, Y, C) f(C|Y, \hat{\Theta}),$$

where the summation is performed over all possible sequences of unmeasured data $C = [c_1, c_2, ..., c_{\mathring{t}}] \in C^*$; C^* denotes the sample space of the random vector C. The estimation of parameters Θ is based on maximization of this expectation of log-likelihood. The maximum is searched iteratively. In each iteration step, the expectation is computed on the basis of old parameter estimates (E step). Then the resulting function of Θ is maximized giving new parameter estimates (M step).

2.2 Normal mixture model

Let us specify the system under consideration in the following way. It consists of several subsystems, on which the set of data $y_t, t = 1, 2, ..., \mathring{t}$ is measured. Each subsystem is modeled by Gaussian pdf

$$f_c(y_t|\theta_c) \sim \mathcal{N}_{y_t}(\mu_c, R_c), \quad c \in c^*, \text{ where}$$

$c \in c^* = \{1, 2, ..., \mathring{c}\}$ is a label of a particular submodel (mixture component),
$\theta_c = [\mu_c, R_c]'$ is a sub-model parameter with μ_c and R_c being the corresponding mean and covariance of y_t, respectively.

We suppose that at each time instant, only one of the models is valid and we take the sequence $c_t \in c^*$ for $t = 1, 2, ..., \mathring{t}$ as pointers to the active subsystem at each time instant t. As we do not know which model is currently active, the individual terms of the sequence are random. They correspond to the unmeasured system output from the general EM algorithm formulation. We suppose them to be independent of $y_\tau, \tau \neq t$ at time t and of the component parameters $\theta_c, c \in c^*$. Their probabilities are

$$f(c_t|\Theta) = \alpha_{c_t}, \quad c_t \in c^*, \quad t = 1, 2, ..., \mathring{t}, \text{ where}$$

$\Theta = [\theta', \alpha']'$,
$\theta = [\theta'_1, \ldots, \theta'_{\mathring{c}}]'$ are parameters of the components,
$\alpha = [\alpha_1, \ldots, \alpha_{\mathring{c}}]$ are component probabilities.

With this assumption, the system model is equivalent to a mixture with $\mathring{c}$ static Gaussian components, each component having the probability $\alpha_c, \quad c \in c^*$, to be active at time $t = 1, 2, ..., \mathring{t}$. The complete description of such mixture model at time t is given by the joint pdf

$$\begin{aligned} & f(y_t, c_t|d_1, \ldots, d_{t-1}, c_1, \ldots, c_{t-1}, \Theta) = \\ = \; & f(y_t|c_t, \Theta) f(c_t|\Theta) = f_{c_t}(y_t|\theta)\, \alpha_{c_t}. \end{aligned}$$

The pdf $f_{c_t}(y_t|\theta)$ models the c_t-th component of the mixture. We have used its assumed independence on α and independence of pdfs $f(c_t|\Theta)$ on θ.

3 EM algorithm for normal mixtures

Our task is to identify the parameters of the mixture μ_c, R_c, α_c for $c \in c^*$. For the solution of the task, we will follow the standard EM algorithm with the only exception that for the numerically sensitive parts of the data covariance estimation we will use factorized algorithms, worked up in connection with the Bayesian approach to identification [5].

3.1 Distribution of unmeasured data

To be able to take the expectation of the log-likelihood, we need the probability $f(C|Y, \Theta)$ of all unknown pointers conditioned not only on parameters Θ but also on measured data $y_t, \ t = 1, 2, ...\mathring{t}$. This joint probability is the product

$$f(C|Y, \Theta) = \prod_{t=1}^{\mathring{t}} f(c_t|y_t, \Theta)$$

in which we have used the independency of the pointer c_t of other pointers c_τ and data y_τ for $t \neq \tau$. The particular probability $f(c_t|y_t, \Theta)$ can be viewed as a prediction of the activity of a component at time t based on the measured data. This probability is determined by Bayes rule

$$f(c_t|y_t, \Theta) \propto f(y_t|c_t, \Theta) f(c_t|\Theta) = f_{c_t}(y_t|\theta)\alpha_{c_t}.$$

It uses the models of components and their probabilities. In accordance with EM algorithm, the expectation defined by this pdf is evaluated with the current parameter estimates substituted for the unknown parameter values. Using the measured data Y we can compute values of the needed pdf $f(c_t|y_t, \Theta = \hat{\Theta})$ for all t and at each time t for all possible values $c \in c^*$ of random variable c_t. To stress that the results are just numbers – probabilistic weights for expectation and consequently for weighted data accumulation, we denote them

$$w_{c,t} = f(c_t = c|y_t, \hat{\Theta}),$$

which is the probability, that the component labeled by c is active at time t.

3.2 E step

The expectation of log-likelihood $\mathcal{L}(\Theta, Y) =$

$$= \sum_{C \in C^*} \left[\sum_{i=1}^{\mathring{t}} \log[\alpha_{c_i} f_{c_i}(y_i|\theta)] \prod_{j=1}^{\mathring{t}} f(c_j|y_j, \hat{\Theta}) \right] =$$

$$= \sum_{c=1}^{\mathring{c}} \sum_{t=1}^{\mathring{t}} \log[\alpha_c f_c(y_t|\theta)] w_{c,t} =$$

$$= \underbrace{\sum_{c=1}^{\mathring{c}}\sum_{t=1}^{\mathring{t}} \log[\alpha_c] w_{c,t}}_{\mathcal{L}_1} + \underbrace{\sum_{c=1}^{\mathring{c}}\sum_{t=1}^{\mathring{t}} \log[f_c(y_t|\theta)] w_{c,t}}_{\mathcal{L}_2} .$$

The first equality is implied by the fact, that probabilities sum to unity.

3.3 M step

This step consists in maximization of the expectation of log-likelihood $\mathcal{L}(\Theta, Y)$ constructed in the previous step. The maximization can be performed independently in two steps. The first term $\mathcal{L}_1$ gives the estimate of α. For explicit maximization of the second term, we suppose the component models in the form of static and normally distributed pdfs

$$f_c(y_t|\theta) \propto |R_c|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_t-\mu_c)' R^{-1}(y_t-\mu_c)\right\}$$

for $c = 1, 2, ...\mathring{c}$, where
μ_c is the expectation of the c^{th} component,
R_c is the covariance matrix of the c^{th} component.

Then the second term $\mathcal{L}_2$ gives estimates of μ_c and R_c for $c = 1, 2, ..., \mathring{c}$. The classical results can be derived by the method of Lagrange multipliers. In their simplified description, the symbols without the sign ˆ mean estimation results obtained at previous iteration step. For $c = 1, 2, ..., \mathring{c}$, the formulae are

$$w_{c,t} =$$

$$= \frac{|R_c|^{-0.5} \exp\left\{-0.5(y_t-\mu_c)' R_c^{-1}(y_t-\mu_c)\right\}}{\sum_{\tilde{c}} |R_{\tilde{c}}|^{-0.5} \exp\left\{-0.5(y_t-\mu_{\tilde{c}})' R_{\tilde{c}}^{-1}(y_t-\mu_{\tilde{c}})\right\}}$$

$$\hat{\alpha}_c = \frac{1}{\mathring{t}}\sum_{t=1}^{\mathring{t}} w_{c,t}, \qquad \hat{\mu}_c = \frac{1}{\mathring{t}\hat{\alpha}_c}\sum_{t=1}^{\mathring{t}} y_t w_{c,t},$$

$$\hat{R}_c = \frac{1}{\mathring{t}\hat{\alpha}_c}\sum_{t=1}^{\mathring{t}} (y_t-\mu_c)(y_t-\mu_c)' w_{c,t}.$$

These formulas are simple and clearly interpretable. The explicit form of the weights $w_{c,t}$ shows the source of numerical troubles. If an iteration fails to give numerically a positive definite $\hat{R}_c$ then the next iteration breaks down. It reflects that the maximized function $\mathcal{L}_2$ is not concave any more and its maximizer has infinite-valued entries.

4 Factorized estimation

The evaluation of $\hat{R}_c$ whose inversion and determinant are used in weighting is bottle neck of the classical formulas. To avoid difficulties mentioned, we will update not the whole matrix but its LD factorization. The formula for computing weights shows, that the inverse of the matrix $\hat{R}_c$ is needed. That is why we will compute directly the LD factorization of the inverse of $\hat{R}_c$.

The unnormalized version of the estimate

$$\tilde{R}_c = \sum_{t=1}^{\mathring{t}} (y_t-\mu_c)(y_t-\mu_c)' w_{c,t}$$

can be simply updated recursively

$$\tilde{R}_{c,t} = \tilde{R}_{c,t-1} + (y_t-\mu_c)(y_t-\mu_c)' w_{c,t} = \tilde{R}_{c,t-1} + z_{c,t} z_{c,t}'$$

where $z_{c,t} = (y_t-\mu_c)\sqrt{w_{c,t}}$, $\tilde{R}_{c,0} = 0$, $\hat{R}_c = \frac{\tilde{R}_{c,\mathring{t}}}{\mathring{t}\hat{\alpha}_c}$.

As we have stated, we need to update LD factorization of the inverse of $\tilde{R}_c$. We drop the fixed subscript $_c$ and introduce factorizations of inversions of $\tilde{R}_{t-1}$ and $\tilde{R}_t$

$$\tilde{R}_{t-1}^{-1} = L\,D\,L', \ \tilde{R}_t^{-1} = \bar{L}\bar{D}\bar{L}', \quad \text{where}$$

$L, \bar{L}$ are lower triangular matrices with unit diagonals and
$D, \bar{D}$ are diagonal matrices with positive diagonals.

All involved square matrices have dimension n.

With these notations, the recursion for $\tilde{R}$ becomes

$$\begin{aligned} \bar{L}\bar{D}\bar{L}' &= \left[(L\,D\,L')^{-1} + z_t z_t'\right]^{-1} = \\ &= L - \left[D\frac{Dff'D}{1+f'Df}\right] L', \text{ with } f = L' z_t. \end{aligned}$$

The definition of LD decomposition implies that the expression in brackets [ ] can be decomposed analytically into the LD form, say $H\bar{D}H'$, see [4]. The following algorithm provides it

for $i = n, n-1, ..., 1$

$$\begin{aligned} \sigma^2 &= 1 + \sum_{k=i}^{n} D_k f_k^2 \\ \bar{D}_{i,i} &= D_{i,i}\frac{\sigma_{i+1}^2}{\sigma_i^2} \\ H_{i,i} &= 1 \end{aligned}$$

for $j = i+1, i+2, ..., n$

$$H_{j,i} = -\frac{f_i f_j}{\sigma_{i+1}^2} D_{j,j}$$

end for j

end for i

We attach to this algorithm the multiplications $f = L' z_t$ and $\bar{L} = LH$ and organize carefully manipulations with individual entries. It gives the overall very compact algorithm that replaces L, D by $\bar{L}, \bar{D}$ after including new data vector z_t:

$s = 1$
for $j = N$ to 1
  $h = z_j$
  for $i = j + 1$ to N
    $h = h + L_{ij}$
  end j
  $a = h/s$
  $v_j = D_j * h$
  $r = s + v_j * h$
  $D_j = s/r * D_j$
  $s = r$
  for $i = j + 1$ to N
    $p = L_{ij} - v_i * a$
    $v_i = v_i + L_{ij} * v_j$
    $L_{ij} = p$
  end i
end j

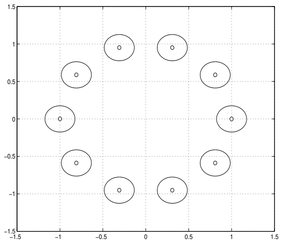
5 Illustrative example

The EM algorithm and its factorized version have been thoroughly tested. The following example illustrates advantages of the proposed factorized modification.

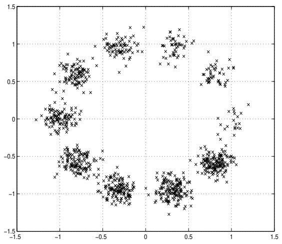
A simulated data sample of the length of 1000 data items was generated from a static normal mixture with 10 components. The mean values of the components together with iso-probability ellipsoids encircling probability 0.75 are shown on the left-hand side of the 1st figure. The data sample is plotted on its right hand side. The weights of the components grow linearly from the first one to the last as it can be seen from the plot.

The estimation setup is on left-hand side of the 2nd figure. The centers of the initial ellipsoid differ from the simulation ones by a random value. Uniform initial estimates of component weights were chosen.

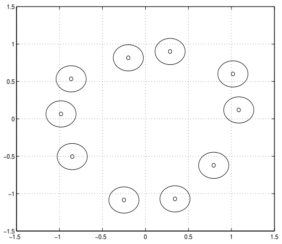
The standard EM algorithm broke down after the 1st iteration step when at least one estimate of R_c lost its definiteness. When running the square root version of EM algorithm, no problems were encountered. The estimated mixture after 20 iterations is shown in the right-hand side of the figure 2.



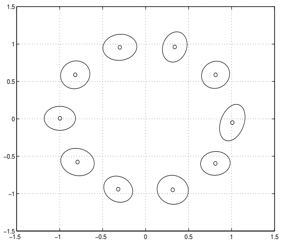
Simulaton setup



Data sample



Estimation setup



Estimation result

6 Conclusions

A new numerical solution to the estimation of normal mixture by EM algorithm has been proposed. The algorithm accepts original estimation of component probabilities α and means θ. Instead of the estimate of the noise covariances R, it updates LD decomposition of its inversion. It simplifies evaluation of weights and, first of all, it prevents these estimates from loosing their positive definiteness. The numerical stability of the proposed algorithm in comparison with the "rough" solution has been demonstrated on simulated examples.

The approach can be extended to the case of dynamic components when the expected value of the predicted data is a linear combination of estimated regression coefficients and known regression vector. The overall description is made in probabilistic terms. It opens a way of including a non-trivial prior probability on estimated parameters giving maximum a posterior probability estimates. The interpretation can be also suppressed and the algorithm used for estimating normal radial basis neural network. These options will be elaborated in detail in foreseen future.

References

[1] A.P.Dempster, N.M.Lair, and D.B.Rubin, "Maximum-likelihood from incomplete data via the em algorithm", *J.Royal Statist. Soc. Ser. B.*, vol. 39, 1977.

[2] C.F.J.Wu, "On the convergence properties of the em algorithm", *The Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[3] R.Redner and H.Walker, "Mixture densities, maximum likelihood and the em algorithm", *SIAM Review*, vol. 26, 1984.

[4] G.J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, Academic Press, New York, 1977.

[5] V. Peterka, "Bayesian approach to system identification", in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.