AN EM-VDM ALGORITHM FOR GAUSSIAN MIXTURES WITH UNKNOWN NUMBER OF COMPONENTS

Nikos Vlassis, Aristidis Likas

14-2000

Preprint no. 14-00/2000

Department of Computer Science
University of Ioannina
451 10 Ioannina, Greece

An EM-VDM algorithm for Gaussian mixtures with unknown number of components

Nikos Vlassis
RWCP, Autonom. Learning Functions SNN
Computer Science Institute
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
vlassis@science.uva.nl

Aristidis Likas
Department of Computer Science
University of Ioannina
45110 Ioannina
Greece
arly@cs.uoi.gr

Abstract

We propose an algorithm for Gaussian mixture modeling with unknown number of mixing components that combines the EM and the VDM algorithm (Lindsay, 1983). In contrast to previous approaches, our method (i) applies on multivariate data, (ii) features an improved VDM gradient function based on a second-order Taylor approximation of the mixture after component insertion, (iii) uses partial EM steps and a kernel-based approach for efficient searching for the global maxima of the log-likelihood. Simulation results indicate that the method manages to estimate the true number of components in most cases, while it compares favorably to EM with fixed number of components in terms of the log-likelihood of the obtained solutions.

1 Introduction

Finite mixture distributions [1] provide a simple framework for modeling population heterogeneity. If $f(\mathbf{x};\boldsymbol{\phi}_j)$ is the j-th component model parametrized on $\boldsymbol{\phi}_j$, then a mixture density for a random vector $\mathbf{x}$ assuming k components is

$$p(\mathbf{x}) = \sum_{j=1}^{k} \pi_j f(\mathbf{x};\boldsymbol{\phi}_j) \tag{1}$$

where π_j are the mixing weights satisfying $\pi_1 + \ldots + \pi_k = 1$, $\pi_j \geq 0$. Mixtures have proven useful tools for data analysis and recent examples are mixtures of factor analyzers [2] and principal component analyzers [3].

The estimation of the parameters of the mixture, i.e., the parameter vector $\boldsymbol{\phi}_j$ of each component, is often carried out with maximum likelihood and the EM algorithm [4]. However, one of the limitations of EM is that it assumes known

number k of mixing components. For real applications, however, the number k is often unknown, and it would be desirable to have an algorithm which starts with a single component and adds components dynamically to the mixture until it reaches a solution with log-likelihood close to the global maximum.

In this paper we propose an algorithm for Gaussian mixture modeling with unknown number of components. Assuming k components, EM steps are repeated until the log-likelihood converges. Then a new component is added to the mixture according to the *vertex direction method* (VDM) [5] and a theorem that specifies the conditions that hold at the global maximum of the log-likelihood. We extend previous results [6, 7, 8] in several ways:

- Our method applies on multivariate mixtures.
- We define an improved VDM gradient function based on a second-order Taylor approximation of the new mixture after component insertion.
- We propose a search technique for the maxima of the new log-likelihood (after component insertion) using partial EM steps and a kernel-based search method.

2 Gaussian mixtures and the EM algorithm

A multivariate Gaussian mixture is defined as the weighted sum (1) with $f(\mathbf{x}; \boldsymbol{\phi}_j)$ being the d-dimensional Gaussian density

$$f(\mathbf{x}; \boldsymbol{\phi}_j) = (2\pi)^{-d/2} |\mathbf{S}_j|^{-1/2} \exp\left[-0.5(\mathbf{x} - \mathbf{m}_j)^T \mathbf{S}_j^{-1} (\mathbf{x} - \mathbf{m}_j)\right] \tag{2}$$

parametrized on the mean $\mathbf{m}_j$ and the covariance matrix $\mathbf{S}_j$, collectively denoted by the parameter vector $\boldsymbol{\phi}_j$. We assume a training set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of i.i.d. points sampled from (1) and the task is to estimate the parameters of the mixture that maximize the log-likelihood

$$\mathcal{L} = n^{-1} \sum_{i=1}^{n} \log p(\mathbf{x}_i). \tag{3}$$

The solutions to the above problem using the EM algorithm are given by the iterative update equations [9]

$$P(j|\mathbf{x}_i) = \frac{\pi_j f(\mathbf{x}_i; \boldsymbol{\phi}_j)}{p(\mathbf{x}_i)}, \tag{4}$$

$$\pi_j := \frac{1}{n} \sum_{i=1}^{n} P(j|\mathbf{x}_i), \tag{5}$$

$$\mathbf{m}_j := \frac{\sum_{i=1}^{n} P(j|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} P(j|\mathbf{x}_i)}, \tag{6}$$

$$\mathbf{S}_j := \frac{\sum_{i=1}^{n} P(j|\mathbf{x}_i)(\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T}{\sum_{i=1}^{n} P(j|\mathbf{x}_i)}. \tag{7}$$

Details concerning the convergence properties of EM can be found, e.g., in [9, 10].

3 The VDM algorithm

The VDM algorithm specifies an incremental strategy for the dynamic insertion of new components in a mixture until a global maximum of the log-likelihood is reached. It is based on the following result.

Theorem 1 (Lindsay 1983, Th. 4.1, 5.3) *Let $p_k(\mathbf{x})$ be a k-component mixture and $f_{k+1}(\mathbf{x};\boldsymbol{\phi})$ a new component model outside the mixture with parameter vector $\boldsymbol{\phi}$. If*

$$D(\boldsymbol{\phi}) \equiv n^{-1} \sum_{i=1}^{n} \left\{ \frac{f_{k+1}(\mathbf{x}_i;\boldsymbol{\phi})}{p_k(\mathbf{x}_i)} - 1 \right\} \tag{8}$$

is a function of the parameter vector $\boldsymbol{\phi}$ of the new component then:

1. *At the global maximum of the log-likelihood holds*

$$\sup_{\boldsymbol{\phi}} D(\boldsymbol{\phi}) = 0. \tag{9}$$

2. *If for some $\boldsymbol{\phi}^*$ holds $D(\boldsymbol{\phi}^*) > 0$, then the log-likelihood cannot decrease if we add the component $f_{k+1}(\mathbf{x};\boldsymbol{\phi}^*)$ to the mixture, with weight $a \in (0,1)$ so that the new mixture is*

$$p_{k+1}(\mathbf{x}) = a f_{k+1}(\mathbf{x};\boldsymbol{\phi}^*) + (1-a) p_k(\mathbf{x}). \tag{10}$$

The first part of the theorem specifies the conditions that hold at the global maximum of the log-likelihood, namely, that the *gradient* function $D(\boldsymbol{\phi})$ is less or equal to zero everywhere in the parameter space. The second part of the theorem is more useful in practice: it states that the addition of a new component to a mixture in the form (10) will always lead to an increase of the log-likelihood, unless we have already reached the global maximum.

Proving the first part of the theorem for the mixture model (10) gives some useful insight. The difference between the new log-likelihood $\mathcal{L}_{k+1} = n^{-1} \sum_{i=1}^{n} \log p_{k+1}(\mathbf{x}_i)$ and the old log-likelihood $\mathcal{L}_k = n^{-1} \sum_{i=1}^{n} \log p_k(\mathbf{x}_i)$ is

$$\begin{aligned} \Delta\mathcal{L} &= n^{-1} \sum_{i=1}^{n} \log \left\{ \frac{a f_{k+1}(\mathbf{x}_i;\boldsymbol{\phi}) + (1-a) p_k(\mathbf{x}_i)}{p_k(\mathbf{x}_i)} \right\} \\ &= n^{-1} \sum_{i=1}^{n} \log \left\{ a \left(\frac{f_{k+1}(\mathbf{x}_i;\boldsymbol{\phi})}{p_k(\mathbf{x}_i)} - 1 \right) + 1 \right\}. \end{aligned} \tag{11}$$

If we use the inequality $\log x \leq x - 1$ we get

$$\Delta\mathcal{L} \leq a D(\boldsymbol{\phi}) \tag{12}$$

which, since a is positive, states that the log-likelihood cannot increase (i.e., we reached the global maximum) if $D(\boldsymbol{\phi}) \leq 0$ for all $\boldsymbol{\phi}$ in the parameter space (i.e., $\sup_{\boldsymbol{\phi}} D(\boldsymbol{\phi}) = 0$).

Intuitively, the above theorem says that in order to increase the log-likelihood of our data set we must place a new component so that it fits as many input data as possible (numerator in (8)), which are at the same time inadequately fitted by the existing k-component mixture (denominator in (8)). From a practical point of view, an important consequence of the theorem is that the mixture (10) is a sufficient model for searching for the global maxima of the log-likelihood.

Based on the above theorem, an incremental algorithm called the *vertex direction method* (VDM) was proposed in [5] for searching for the global maximum likelihood solutions. The idea is to find in each step the parameter vector $\boldsymbol{\phi}^*$ that maximizes $D(\boldsymbol{\phi})$ and then to estimate a in (10) that maximizes $\Delta\mathcal{L}$. The VDM algorithm can be shown to converge to the global maximum of the log-likelihood [5].

The VDM algorithm can be justified by a first order Taylor expansion of $\Delta\mathcal{L}$ in (11), regarded as a function of a, about $a = 0$. This corresponds to approximating $\Delta\mathcal{L}$ by its upper bound from (12)

$$\Delta\mathcal{L} = aD(\phi). \tag{13}$$

Since a is positive, the above equation implies that we can search for the maxima of $\Delta\mathcal{L}$ by maximizing $D(\phi)$. However, a linear approximation can often be too rough and in [7] a second order Taylor expansion about $a = 0$ is proposed. However, due to the discontinuities of $\Delta\mathcal{L}$ near $a = 0$, suboptimal solutions may arise as we noticed in our implementations. Moreover, it is not clear how we can efficiently search for the maxima of $\Delta\mathcal{L}$ over the parameter space, especially in the case of multivariate mixtures. These limitations of the original VDM method motivated our approach described below.

3.1 A VDM approach based on partial EM

Our approach is based on the observation that the mixture $p_{k+1}(\mathbf{x})$ in (10) can be regarded as a two-component mixture, the first component being the new added component $f_{k+1}(\mathbf{x};\phi)$ and the second component being the old mixture $p_k(\mathbf{x})$. Thus, *partial* EM steps can be used to find the parameters a^* and ϕ^* which maximize the new log-likelihood $\mathcal{L}_{k+1}$ (and thus $\Delta\mathcal{L}$) while leaving the parameters of $p_k(\mathbf{x})$ unchanged. This means applying the EM equations (4)–(7) only on the mixing weight a, the mean $\mathbf{m}$, and the covariance matrix $\mathbf{S}$ of the newly inserted component, i.e.,

$$P(k+1|\mathbf{x}_i) = \frac{a f_{k+1}(\mathbf{x}_i;\mathbf{m},\mathbf{S})}{a f_{k+1}(\mathbf{x}_i;\mathbf{m},\mathbf{S}) + (1-a)p_k(\mathbf{x}_i)}, \tag{14}$$

$$a := \frac{1}{n}\sum_{i=1}^{n} P(k+1|\mathbf{x}_i), \tag{15}$$

$$\mathbf{m} := \frac{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)}, \tag{16}$$

$$\mathbf{S} := \frac{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T}{\sum_{i=1}^{n} P(k+1|\mathbf{x}_i)}. \tag{17}$$

Since only the parameters of the new component are updated, partial EM steps provide a simple and fast method for searching for the maxima of $\mathcal{L}_{k+1}$, without needing to resort to expensive gradient-based nonlinear optimization methods.

3.2 Initialization of partial EM

Still, for the partial EM to be effective we need a good initialization of the parameters of the new component. We propose the following method. We expand $\mathcal{L}_{k+1}$ by second order Taylor about $a_o = 0.5$ (which avoids problems of discontinuities near the boundaries of a) and then maximize the resulting quadratic function w.r.t. a. It is not difficult to check that this procedure gives the solution

$$\mathcal{L}_{k+1} = \mathcal{L}_{k+1}(a_o) - \frac{[\mathcal{L}'_{k+1}(a_o)]^2}{2\mathcal{L}''_{k+1}(a_o)} \tag{18}$$

with $\mathcal{L}'_{k+1}$ and $\mathcal{L}''_{k+1}$ the first and second derivatives of $\mathcal{L}_{k+1}$ w.r.t. a. If we define

$$\delta(\mathbf{x}_i,\phi) = 2\,\frac{f_{k+1}(\mathbf{x}_i;\phi) - p_k(\mathbf{x}_i)}{f_{k+1}(\mathbf{x}_i;\phi) + p_k(\mathbf{x}_i)} \tag{19}$$

then the local maximum of $\mathcal{L}_{k+1}$ near $a_o = 0.5$ is

$$\mathcal{L}_{k+1} = n^{-1} \sum_{i=1}^{n} \log \frac{f_{k+1}(\mathbf{x}_i; \boldsymbol{\phi}) + p_k(\mathbf{x}_i)}{2} + \frac{\left[n^{-1} \sum_{i=1}^{n} \delta(\mathbf{x}_i, \boldsymbol{\phi})\right]^2}{2n^{-1} \sum_{i=1}^{n} \delta^2(\mathbf{x}_i, \boldsymbol{\phi})} \tag{20}$$

and is obtained for a equal to

$$a = \frac{1}{2} + \frac{n^{-1} \sum_{i=1}^{n} \delta(\mathbf{x}_i, \boldsymbol{\phi})}{n^{-1} \sum_{i=1}^{n} \delta^2(\mathbf{x}_i, \boldsymbol{\phi})}. \tag{21}$$

3.3 Kernel-based search

The above procedure makes the new log-likelihood (20) independent of the mixing weight a, while the value of a from (21) that maximizes $\mathcal{L}_{k+1}$ can be used in the initialization of the partial EM. The next task is to find a sensible initialization of the center $\mathbf{m}$ and covariance matrix $\mathbf{S}$ of the new component so that $\mathcal{L}_{k+1}$ is maximized.

We observe that $\mathcal{L}_{k+1}$ in (20) depends only on $f_{k+1}(\mathbf{x}_i; \mathbf{m}, \mathbf{S})$ which, for constant spherical covariance $\mathbf{S} = \sigma^2 \mathbf{I}$, is a function of the Euclidean distance between the input $\mathbf{x}_i$ and the new center $\mathbf{m}$. If we restrict our search for $\mathbf{m}$ over all training points $\mathbf{x}_j$, evaluation of (20) for all possible data implies estimation of $O(n^2)$ Euclidean distances for each pair $(\mathbf{x}_i, \mathbf{x}_j)$ of training points. However, this computation may be carried out only once at the beginning of the algorithm by storing a *kernel* matrix $\mathbf{K}$ with elements

$$\mathbf{K}_{ij} = (2\pi\sigma^2)^{-d/2} \exp(-0.5||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma^2) \tag{22}$$

and then use $\mathbf{K}_{ij}$ for computing the $f_{k+1}(\mathbf{x}_i; \mathbf{x}_j, \mathbf{S})$ values in (20). This gives rise to an algorithm similar to *kernel feature analysis* [11] where a kernel matrix is defined and a search over all inputs is used in each step to optimize a contrast function acting on a 'feature' space of the original data. In our case, we can define many kernel matrices, one for each σ, and then compute the maximum of (20) among all of them.

3.4 The algorithm

Summarizing the above ideas we have the following algorithm for Gaussian mixture modeling with unknown number of components.

1. Initialize using one component. Set σ^2 to a fraction (e.g., 0.1) of the minimum eigenvalue of the covariance matrix of $\mathbf{x}_i$. Compute the kernel matrix from (22).
2. Perform EM steps until convergence.
3. Search over all $\mathbf{x}_j$ for candidate locations for the new component. Set $\mathbf{m}$ to the $\mathbf{x}_j$ that maximizes (20) using the precomputed kernel values $\mathbf{K}_{ij}$ in place of $f_{k+1}(\mathbf{x}_i; \mathbf{x}_j, \sigma^2 \mathbf{I})$.
4. Initialize the partial EM with the estimated $\mathbf{m}$, $\mathbf{S} = \sigma^2 \mathbf{I}$, and a from (21).
5. Apply partial EM steps (14)–(17) until convergence.
6. If $\Delta\mathcal{L} \leq$ VDMTHRESHOLD (e.g., 0.05) then terminate, otherwise allocate the new component and go to 2.

Since EM cannot lead to decrease of the log-likelihood and the partial EM solutions are accepted only if $\Delta\mathcal{L} > 0$, the algorithm ensures the monotone increase of the log-likelihood.

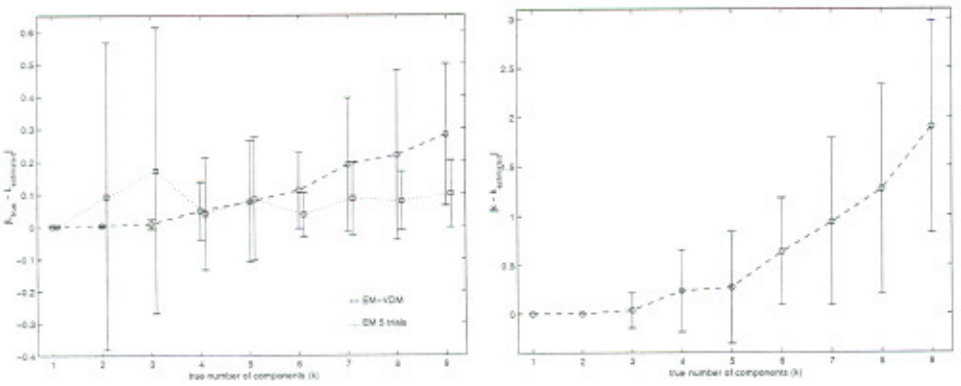

Figure 1: Performance of EM-VDM: log-likelihood (left), estimated k (right).

4 Demonstration and discussion

In order to assess the effectiveness of the proposed EM-VDM method in approximating the correct number k of components of a Gaussian mixture, and also to compare the solutions of EM-VDM with those of EM with fixed and known k, we generated 30 random two-dimensional mixtures with centers uniformly sampled within a square rectangle, and mixing weights and covariances also random (overlapping components were allowed which made the experiment difficult). From each such mixture, 500 points were sampled and the theoretical log-likelihood was computed. Then EM-VDM was applied starting with one component, and also five trials of EM using the true k and keeping the best solution among the five. This experiment was repeated for values of k from one to nine.

In Fig. 1 (left) we plot the average (with 1σ error bars) of the absolute difference between the theoretical log-likelihood and the estimated one for EM-VDM and EM, as a function of the true k. In Fig. 1 (right) we plot average and error bars of the difference between k and the estimated k for EM-VDM. We see that for small number of components our algorithm can estimate very accurately the true number of components, while its solutions outperform EM in terms of log-likelihood. The latter can be attributed, on the one hand, to the local maxima of EM due to improper initialization, and on the other, to the very good performance of EM-VDM when the components of the mixture are well-separated. This property of EM-VDM is more noticeable in dimensions higher than two, where the sparsity of the data and the sampling of the mixture centers within a hypercube give rise to well-separated mixtures. We should emphasize here that the above results are unoptimized, in the sense that a single kernel matrix was used and no tuning was tried over the VDMTHRESHOLD or the value of σ in the kernel matrix (22).

In the past we have also proposed a different dynamic approach which splits components of the mixture based on a statistical test involving the kurtosis of each component [12, 13]. However, the current approach is more robust and avoids problems of the kurtosis related to outliers. An analogous method that deals with the problem of local maxima of EM has also been recently proposed in [14]. In this method split-and-merge operations are applied on the components of the mixture using a criterion based on the Kullback divergence between a component density and the empirical density in the vicinity of the component.

The current algorithm furthers existing VDM approaches [6, 7, 8] which are either

one-dimensional, are based on Taylor approximations of the mixture log-likelihood near the boundaries of the a range, or use expensive global optimization routines for optimizing the VDM gradient function. As future work we want to study the effect of the free parameters VDMTHRESHOLD and σ in (22) on the behavior of the algorithm and how we can automate their tuning. Also it would be interesting to check to which extend our method can be used for estimating the number of components in latent mixture models [2, 3].

Acknowledgments

The first author is supported by the Real World Computing program.

References

[1] D. M. Titterington, A. F. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, 1985.

[2] Z. Ghahramani and G.E. Hinton, "The EM algorithm for mixtures of factor analyzers," Tech. Rep., University of Toronto, 1997, CRG-TR-96-1.

[3] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, pp. 1–38, 1977.

[5] B. G. Lindsay, "The geometry of mixture likelihoods: a general theory," *Ann. Statist.*, vol. 11, no. 1, pp. 86–94, 1983.

[6] R. DerSimonian, "Maximum likelihood estimation of a mixing distribution," *J. Roy. Statist. Soc. C*, vol. 35, pp. 302–309, 1986.

[7] D. Böhning, "A review of reliable maximum likelihood algorithms for semiparametric mixture models," *J. Statist. Plann. Inference*, vol. 47, pp. 5–28, 1995.

[8] B. G. Lindsay and M. L. Lesperance, "A review of semiparametric mixture models," *J. Statist. Plann. Inference*, vol. 47, pp. 29–39, 1995.

[9] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

[10] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, New York, 1997.

[11] A. J. Smola, O. L. Mangasarian, and B. Schölkopf, "Sparse kernel feature analysis," Tech. Rep., Data Mining Institute, University of Wisconsin, Madison, 1999, 99-04.

[12] N. Vlassis and A. Likas, "A kurtosis-based dynamic approach to Gaussian mixture modeling," *IEEE Trans. on Systems, Man, and Cybernetics, Part A*, vol. 29, no. 4, pp. 393–399, July 1999.

[13] N. Vlassis, A. Likas, and B. Kröse, "Multivariate Gaussian mixture modeling with unknown number of components," Tech. Rep., Computer Science Institute, University of Amsterdam, The Netherlands, Apr. 2000, IAS-UVA-00-04.

[14] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "SMEM algorithm for mixture models," *Neural Computation*, 2000, (to appear).