An Entropy Maximization Approach to Optimal Model Selection in Gaussian Mixtures

Antonio Peñalver, Juan M. Sáez, and Francisco Escolano

Robot Vision Group
Departamento de Ciencia de la Computación e Inteligencia Artificial
Universidad de Alicante, Spain
{apenalver, jmsaez, sco}@dccia.ua.es
http://rvg.ua.es

Abstract. In this paper we address the problem of estimating the parameters of a Gaussian mixture model. Although the EM (Expectation-Maximization) algorithm yields the maximum-likelihood solution it has many problems: (i) it requires a careful initialization of the parameters; (ii) the optimal number of kernels in the mixture may be unknown beforehand. We propose a criterion based on the entropy of the pdf (probability density function) associated to each kernel to measure the quality of a given mixture model, and a modification of the classical EM algorithm to find the optimal number of kernels in the mixture. We test this method with synthetic and real data and compare the results with those obtained with the classical EM with a fixed number of kernels.

1 Introduction

Gaussian mixture models, have been widely used in the field of statistical pattern recognition. One of the most common methods for fitting mixtures to data is the EM algorithm [4]. However, this algorithm is prone to initialization errors and, in these conditions, it may converge to local maxima of the log-likelihood function. In addition, the algorithm requires that the number of elements (kernels) in the mixture is known beforehand. For a given number of kernels, the EM algorithm yields a maximum-likelihood solution but this does not ensure that pdf of the data (multi-dimensional patterns) is properly estimated. A maximum-likelihood criterion with respect to the number of kernels is not useful because it tends to use a kernel to describe each pattern.

The so called model-selection problem has been addressed in many ways. Some approaches start with a few number of kernels and add new kernels when necessary. For instance, in [14], the kurtosis is used as a measure of non-Gaussianity yielding a test for splitting a kernel in one-dimensional data. In [15] this method is extended to the multi-dimensional case. This approach has same drawbacks, because kurtosis can be very sensitive to outliers. In [16] it is proposed a greedy method, which performs a global search in combination with another local search whenever a new kernel is added.

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 432–439, 2003.

Other model-selection methods start with a high number of kernels and proceed to fuse them. In [5][6], the EM algorithm is initialized with many kernels randomly placed and then the Minimum-description length principle [9] is applied to iteratively remove some of the kernels until the optimal number of them is found. In [11], the proposed algorithm is allowed both to split and fuse kernels. Kernel fusion arises when many patterns have the same posterior probability and splitting is driven by the Kullback-Leibler divergence between a component density and empirical density in the neighborhood of the component. In this approach, the number of components remains unchanged.

In this paper we propose a method that starting with few kernels, typically one, find the maximum-likelihood solution. Then it tests whether the underlying pdf of each kernel is Gaussian and otherwise it replaces that kernel with two kernels adequately separated from each other. In order to detect non-Gaussianity we compare the entropy of the underlying pdf with the theoretical entropy of a Gaussian. After two new kernels are introduced, our method performs several steps of partial EM in order to obtain a new maximum-likelihood solution.

2 Gaussian-Mixture Models

A d-dimensional random variable $\mathbf{y}$ follows a finite-mixture distribution when its pdf $p(y|\Theta)$ can be described by a weighted sum of known pdf's named kernels. When all these kernels are Gaussian, the mixture is named in the same way:

$$p(\mathbf{y}|\Theta) = \sum_{i=1}^{K} \pi_i p(\mathbf{y}|\Theta_i), \text{ where } 0 \leq \pi_i \leq 1, \ i = 1, ..., K, \text{ and } \sum_{i=1}^{K} \pi_i = 1, \quad (1)$$

being K the number of kernels, $\pi_1, ..., \pi_k$ the a priori probabilities of each kernel, and Θ_i the parameters describing the kernel. In Gaussian mixtures, $\Theta_i = \{\mu_i, \Sigma_i\}$, that is, the average vector and the covariance matrix.

The set of parameters of a given mixture is $\Theta \equiv \{\Theta_1, ..., \Theta_k, \pi_1, ..., \pi_k\}$. Obtaining the optimal set of parameters Θ^* is usually posed in terms of maximizing the log-likelihood of the pdf to be estimated:

$$\ell(Y|\Theta) = \log p(Y|\Theta) = \log \prod_{n=1}^{N} p(y_n|\Theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(y_k|\Theta_k). \quad (2)$$

$$\Theta^* = \arg\max_{\Theta} \ell(\Theta). \quad (3)$$

where $Y = \{y_1, ...y_N\}$ is a set of N i.i.d. samples of the variable Y.

2.1 EM Algorithm

The EM (Expectation-Maximization) algorithm [4] is an iterative procedure that allows us to find maximum-likelihood solutions to problems involving *hidden variables*. The EM algorithm generates a sequence of estimations of parameters

$\{\Theta^*(t), t = 1, 2, ...\}$ by alternating an expectation step and the maximization step until convergence. In the case of mixtures [8], the hidden variable can be regarded as the kernel each data has been sampled from. The E-step estimates the posterior probability that the data $\mathbf{y}_n$ was sampled with the kernel k:

$$p(k|\mathbf{y}_n) = \pi_k p(\mathbf{y}^{(n)}|k)/\Sigma_{j=1}^{K} \pi_j p(\mathbf{y}^{(n)}|k) \tag{4}$$

In M-step, the new parameters $\Theta^*(t+1)$ are given by:

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} p(k|\mathbf{y}_n),\ \ \mu_k = \frac{\sum_{n=1}^{N} p(k|\mathbf{y}_n)\mathbf{y}_n}{\sum_{n=1}^{N} p(k|\mathbf{y}_n)} \text{ and } \mu_k = \frac{\sum_{n=1}^{N} p(k|\mathbf{y}_n)\mathbf{y}_n}{\sum_{n=1}^{N} p(k|\mathbf{y}_n)}. \tag{5}$$

A detailed description of this classic algorithm is given in [8]. Here we focus on the fact that if K is unknown beforehand it cannot be estimated through maximizing the log-likelihood because $\ell(\Theta)$ grows with K. Fig. 1 shows the effect of using only a kernel, in classical EM algorithm with fixed number of kernels, to describe two Gaussian distributions: density is underestimated giving a poor description of the data. In the next section we describe the use of entropy to test whether a given kernel properly describes the underlying data.

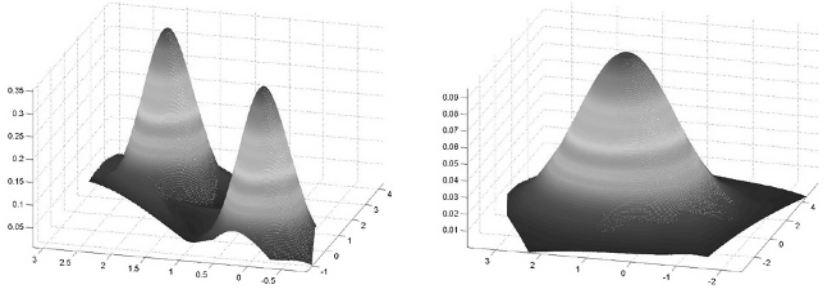

Fig. 1. Classic EM algorithm, fits erroneously data of a bimodal distribution (with averages $\mu_1 = [0, 0]$ y $\mu_2 = [3, 2]$) (left) to a Gaussian with $\mu = [1.5, 1]$ (right).

3 Entropy Estimation

Entropy is a basic concept in information theory. The entropy of a given variable Y can be interpreted in terms of information, randomness, dispersion, and so on [3][10]. For a discrete variable we have:

$$H(Y) = -E_y[\log(P(Y))] = -\sum_{i=1}^{N} P(Y = y_i) \log p(Y = y_i). \tag{6}$$

where $y_1, ..., y_N$ is the set of values of variable Y. A fundamental result of information theory is that Gaussian variables have the maximum entropy among all

the variables with equal variance. Consequently the entropy of the underlying distribution of a kernel should reach a maximum when such a distribution is Gaussian. This theoretical maximum entropy is given by:

$$H(Y) = \frac{1}{2}\log[(2\pi e)^d|\Sigma|]. \tag{7}$$

Then, in order to decide whether a given kernel is truly Gaussian or must be replaced by two other kernels, we compare the estimated entropy of the underlying data with the entropy of a Gaussian. However, one of the main problems of this approach is that we must estimate, in principle, the pdf given a few samples [12][13][17].

3.1 Entropy Estimation with Parzen's Windows

The Parzen's windows approach [7] is a non-parametric method for estimating pdf's for a finite set of patterns. The general form of these pdf's using a Gaussian kernel and assuming diagonal covariance matrix $\psi = Diag(\sigma_1^2, ...\sigma_{N_a}^2)$ is:

$$P^*(Y,a) \equiv \frac{1}{N_a}\sum_{y_a\in a}\frac{1}{\prod_{i=1}^d \sigma_i(2\pi)^{d/2}}\prod_{j=1}^d \exp\left\{-\frac{1}{2}\left(\frac{y^j - y_a^j}{\sigma_j}\right)^2\right\}, \tag{8}$$

where a is a sample of the variable Y, N_a is the size of the sample, y^j represents the j-th component of y and y_a^j represents the j-th component of kernel y_a. In [12] it is proposed a method for adjusting the widths of the kernels using maximum likelihood. Given the definition of entropy in Equation 6, we have:

$$H_b(Y) \equiv -E_b[\log(P(Y))] = -\frac{1}{N_b}\sum_{y_b\in b}\log(P(y_b)) = -\frac{1}{N_b}\log(\ell(b)), \tag{9}$$

where $\ell(b)$ is the likelihood of the data. As maximizing likelihood is equivalent to minimice entropy, this approach consists of estimating the derivative of entropy with respect to the widths of the kernels, and performs a gradient descent towards the optimal widths:

$$\frac{\partial}{\partial\sigma_d}H^*(Y) = \frac{1}{N_b}\sum_{y_b\in b}\sum_{y_a\in a}\frac{K_\psi(y_b - y_a)}{\sum_{y_a\in a}K_\psi(y_b - y_a)}\left(\frac{1}{\sigma_d}\right)\left(\frac{[y_b - y_a]_d^2}{\sigma_d^2} - 1\right), \tag{10}$$

being σ_d the standard deviation in each dimension.

Given the optimal widths of the kernel, the entropy is estimated by

$$H^*(Y) = \frac{1}{N_b}\sum_{y_b\in b}\log\left(\frac{1}{N_a}\sum_{y_a\in a}K_\psi(y_b - y_a))\right), \tag{11}$$

In Fig. 2 we show the entropy estimation obtained for a sample of a $2D$ Gaussian variable with a diagonal covariance matrix with $\sigma_1^2 = 0.36$ and $\sigma_2^2 = 0.09$, for different widths. The approximation of the maximum entropy defined in Equation 7 is 1.12307. From the shape of this function, it can be deduced that the optimal widths lay in a wide interval and consequently their choice is not so critical.

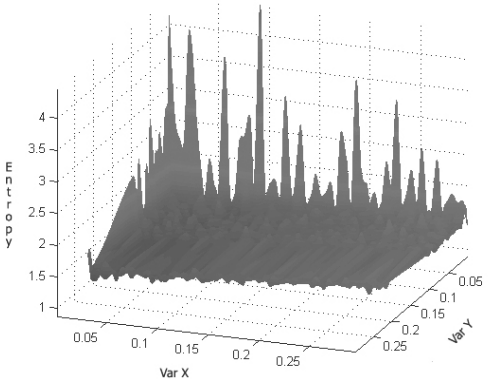

Fig. 2. Representing entropy as a function of the widths of the Parzen's kernels.

4 Optimal Model Selection with Maximum Entropy

4.1 Proposed Method

Comparing the estimations given for Equations 7 and 11, we have a way of quantifying the degree of Gaussianity of a given kernel. Given a set of kernels for the mixture (initially one kernel) we evaluate the real global entropy $H(y)$ and the theoretical maximum entropy $H_{max}(y)$ of the mixture by considering the individual pairs of entropies for each kernel, and the prior probabilities:

$$H(Y) = \sum_{k=1}^{K} \pi_k H_k(Y) \text{ and } H_{max}(Y) = \sum_{k=1}^{K} \pi_k H_{max_k}(Y)\,. \tag{12}$$

If the ratio $H(y)/H_{max}(y)$ is above a given threshold (typically 0.95) we consider that all kernels are well fitted. Otherwise, we select the kernel with the lowest individual ratio and it is replaced by two other kernels that are conveniently placed. Then, a new EM starts.

As the estimation of the entropy of a kernel requires two data sets, we select those whose distance to the average μ_k is between the limits of a Gaussian: $-3\sqrt{\lambda_i^k} \leq b_i \leq 3\sqrt{\lambda_i^k}$, with $b = P_t^T(\mu_k - y)$. λ_i^k, with $i = 1, 2..d$, are the eigenvectors associated to the kernel, and b is the projection of a data y on the eigenspace spanned by the eigenvectors of the covariance matrix collected in P_k.

4.2 Introducing a New Kernel

A low $H(y)/H_{max}(y)$ local ratio indicates that multimodality arises and thus the kernel must be replaced by two other kernels. Applying PCA (Principal Component Analysis) to the original kernel we find that the main eigenvector

indicates the direction of maximum variability and we can put the two new kernels along the opposite senses of this direction (Fig. 3). Being k the kernel with low Gaussianity, after splitting it, the two new kernels k_1 and k_2 with parameters $\Theta_{k_1} = (\mu_{k_1}, \Sigma_{k_1})$ and $\Theta_{k_2} = (\mu_{k_2}, \Sigma_{k_2})$ have the following initial averages $\mu_{k_1} = \mu_k + \sqrt{\lambda_k}\mathbf{V}$ and $\mu_{k_2} = \mu_k - \sqrt{\lambda_k}\mathbf{V}$, with λ_k the principal eigenvalue for kernel k and $\mathbf{V}$ its associated normalized eigenvector. Furthermore, the width of the two

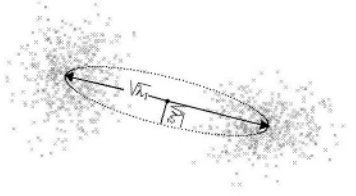

Fig. 3. The direction of maximum variability is associated to the eigenvector with highest eigenvalue autovector λ_1

new kernel is divided by two. If $\lambda_k^{'}$ is the main eigenvalue in both kernels, then $\sqrt{\lambda_k'} = \frac{\sqrt{\lambda_k}}{2}$, consecuently, $\Sigma_{k_1} = \Sigma_{k_2} = \frac{1}{4}\Sigma_k$. Finally, the new priors should also verify $\sum_{k=1}^{K} \pi_k = 1$, so we initialize them with $\pi_{k_1} = \pi_{k_2} = \frac{1}{2}\pi_k$. The proposed algorithm is described in Fig. 4.

Initialization: Start with a unique kernel. $K = 1$. $\Theta_1 = \{\mu_1, \Sigma_1\}$ with random values.
Repeat: Main loop
  Repeat: E, M Steps
    Estimate log-likelihood in iteration i: ℓ_i
  Until: $|\ell_i - \ell_{i-1}| <$ CONVERGENCE_THRESHOLD
  Evaluate $H(Y)$ and $H_{max}(Y)$ globally
  If $(H(Y)/H_{max} <$ ENTROPY_THRESHOLD)
    Select kernel k with the lowest ratio and decompose into k_1 and k_2
    Initialize parameters Θ_{k_1} and Θ_{k_2}
      Initialize new averages: $\mu_{k_1} = \mu_k + \sqrt{\lambda_k}\mathbf{V}$, $\mu_{k_2} = \mu_k - \sqrt{\lambda_k}\mathbf{V}$
      Initialize new covariance matrices: $\Sigma_{k_1} = \Sigma_{k_2} = \frac{1}{4}\Sigma_k$
      Set new a priori probabilities: $\pi_{k_1} = \pi_{k_2} = \frac{1}{2}\pi_k$
  Else
    Final = True
Until: Final = True

Fig. 4. Our maximum-entropy algorithm

4.3 Validation of the Method

In order to test our approach we have performed several experiments with synthetic and real data. In the first one we have generated 2500 samples from 5 bi-dimensional Gaussians with prior probabilities $\pi_k = 0.2\ \forall k$. Their averages are: $\mu_1 = [-1,-1]^T$, $\mu_2 = [6,3]^T$, $\mu_3 = [3,6]^T$, $\mu_4 = [2,2]^T$, $\mu_5 = [0,0]^T$ and their covariance matrices are

$$\Sigma_1 = \Sigma_5 = \begin{bmatrix} 0.20 & 0.00 \\ 0.00 & 0.30 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.60 & 0.15 \\ 0.15 & 0.60 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.40 & 0.00 \\ 0.00 & 0.25 \end{bmatrix}, \Sigma_4 = \begin{bmatrix} 0.60 & 0.00 \\ 0.00 & 0.30 \end{bmatrix}.$$

We have used a Gaussianity threshold of 0.95, and a convergence threshold of 0.001 for the EM algorithm. In order to evaluate the robustness of the proposed algorithm, several outliers were added to the data set. The sample size for estimating entropy through Parzen has been 75. We have found that despite this small size, entropy estimation is good enough. Our algorithm converges after 30 iterations finding correctly the number of kernels. In Fig. 5 we show the evolution of the algorithm. We have also applied the classical EM with 5 kernels. We

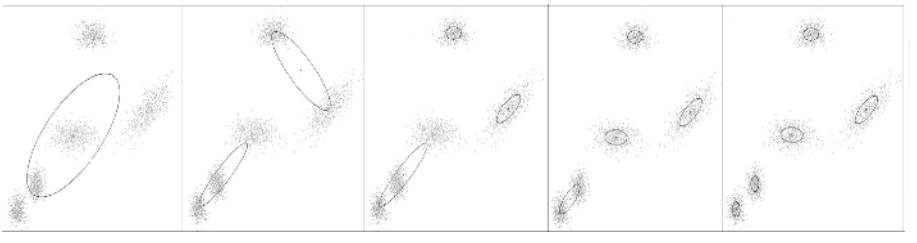

Fig. 5. Evolution of our algorithm from one initial kernel to 5 real kernels.

have performed 20 experiments with the latter data but randomly placing the kernels in each one. In 18 of the 20 experiments the classical EM finds a local maxima. The averaged number of iterations needed was 95 (being 250 the maximum and 23 the minimum). Then, only in two cases the classical EM found the global maxima using 21 and 31 iterations respectively. Thus, our approach addresses two basic problems of the classical EM: the initialization and the model selection.

Finally, we have applied the proposed method to the well known *Iris* [2] data set, that contains 3 classes of 50 (4-dimensional) instances referred to a type of iris plant: *Versicolor*, *Virginica* and *Setosa*. Because the problem is 4-dimensional, 50 samples are insufficient to construct the pdf using Parzen. In order to test our method, we have generated 300 training samples from the averages and covariances of the original classes and we have checked the performance in a classification problem with the original 150 samples. Starting with $K = 1$, the method correctly selected $K = 3$. Then, a maximum a posteriori classifier was built, with classification performance of 98% (only three *Versicolor* were classified like *Virginica*).

5 Conclusions and Future Work

In this paper we have presented a method for finding the optimal number of kernels in a Gaussian mixture based on maximum entropy. We start the algorithm with only one kernel and then we decide to split it on the basis of the entropy of the underlying pdf. The algorithm converges in few iterations and is suitable for density estimation and classification problems. We are currently validating this algorithm in real image classification problems and also exploring new methods of estimating entropy directly, bypassing the estimation of the pdf.

References

1. Bishop, C.: Neural Networks for Pattern Recognition. Oxford Univ. Press (1995).
2. Blake, C.L, Merz, C.J.:UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences (1998).
3. Cover, T, Thomas, J: Elements of Information Theory. J. Wiley and Sons (1991).
4. Dempster, A, Laird, N, Rubin, D.: Maximum Likelihood estimation from incomplete data via the EM Algorithm. Journal of The Royal Statistical Society B, vol. 39 pp. 1 38 (1977).
5. Figueiredo, M.A.T, Leitao, J.M.N, Jain, A.K.: On Fitting Mixture Models. In: Hancock, E.R., Pelillo, M. (eds.): Energy Minimization Methods in Computer Vision and Pattern Recognition. Lecture Notes in Computer Science, Vol. 1654. Springer-Verlag, Berlin Heidelberg New York (1999) 54-69.
6. Figueiredo, J.M.N, Jain, A.K.: Unsupervised Selection and Estimation of Finite Mixture Models. In Proceedings of the International Conference on Pattern Recognition. ICPR2000 (Barcelona), 2000.
7. Parzen, E. On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33:1065–1076, 1962.
8. Redner, R.A, Walker, H.F.: Mixture Densities, Maximum Likelihood, and the EM Algorithm. SIAM Review, 26(2):195-239, (1984).
9. Rissanen, J.: Stochastic Complexity in Statistical Inquiry. World Scientific, (1989).
10. Shannon, C.: A Mathematical Theory of Comunication. The Bell System Technical Journal, Vol. 27, 379-423, 623-656, (1948).
11. Ueda, N, Nakano, R, Ghahramani, Z., G. E. Hinton: SMEM Algorithm for Mixture Models. Neural Computation, 12:2109-2128, (2000).
12. Viola, P, Wells III, W. M.: Alignment by Maximization of Mutual Information. In 5th Intern. Conf. on Computer Vision, pages 16-23, Cambridge, MA. IEEE. (1995).
13. Viola, P, Schraudolph, N.N, Sejnowski, T.J.: Empirical Entropy Manipulation for Real-World Problems. Adv. in Neural Infor. Proces. Systems 8, MIT Press (1996).
14. Vlassis, N, Likas, A.: A Kurtosis-Based Dynamic Approach to Gaussian Mixture Modeling. IEEE Trans. Systems, Man, and Cybernetics, 29(4):393-399 (1999).
15. Vlassis, N, Likas, A, Krose, B.: A Multivariate Kurtosis-Based Dynamic Approach to Gaussian Mixture Modeling.Inteligent Autonomous Systems Tech. Report (2000).
16. Vlassis, N, Likas, A.: A Greedy EM Algorithm for Gaussian Mixture Learning. Neural Processing Letters. To appear.
17. Wolpert, D, Wolf, D.: Estimating Function of Probability Distribution from a Finite Set of Samples. Phisical Review E, Volume 52, No 6 (1995).