Density Estimation Using Mixtures of Mixtures of Gaussians*

Wael Abd-Almageed and Larry S. Davis

Institute for Advanced Computer Studies, University of Maryland,
College Park, MD 20742
{wamageed, lsd}@umiacs.umd.edu

Abstract. In this paper we present a new density estimation algorithm using mixtures of mixtures of Gaussians. The new algorithm overcomes the limitations of the popular Expectation Maximization algorithm. The paper first introduces a new model selection criterion called the Penalty-less Information Criterion, which is based on the Jensen-Shannon divergence. Mean-shift is used to automatically initialize the means and covariances of the Expectation Maximization in order to obtain better structure inference. Finally, a locally linear search is performed using the Penalty-less Information Criterion in order to infer the underlying density of the data. The validity of the algorithm is verified using real color images.

1 Introduction

The Expectation Maximization algorithm (EM) [1] perhaps is the most frequently used parametric technique for estimating probability density functions (PDF) in both univariate and multivariate cases. It has been widely applied in computer vision [2], and pattern recognition [3] applications. In all of these areas, EM is used to model the PDF of a set of feature vectors using a given parametric model. Usually a mixture of Gaussians with a finite number of components is used to approximate the density function. The main advantage of EM is that it provides a closed-form analytical representation of the PDF. However, EM suffers a few limitations that will be discussed later.

This paper introduces a new nonparametric approach based on the mean-shift algorithm for overcoming the limitations of the EM algorithm. The paper is organized as follows. Section 2 briefly discusses the limitations of both the EM and the mean-shift algorithms. In Section 3 we introduce a new model selection criterion called the Penalty-less Information Criterion (PIC) that will be used in the subsequent sections. Section 4 presents a mean-shift- and PIC-based method for nonparametrizing the EM algorithm. The results of using the proposed algorithm are introduced in Section 5. Finally, Section 6 summarizes the paper and highlights directions for future research. Throughout the paper, we kindly encourage the reader to refer to the electronic copy for the clearer color version of the figures.

* This research has been funded, in part, by the Army Research Laboratory's Robotics Collaborative Technology Alliance program, contract number DAAD 19-012-0012 ARL-CTA-DJH.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part IV, LNCS 3954, pp. 410–422, 2006.

2 Expectation Maximization and Mean-Shift

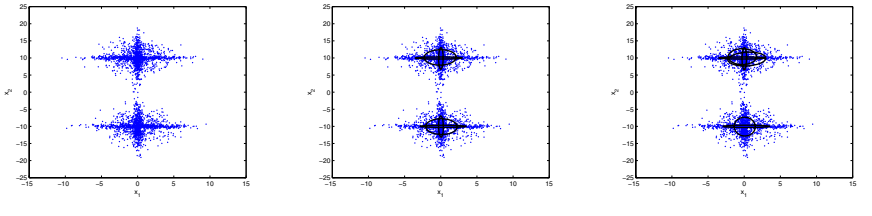
2.1 Expectation Maximization

The Expectation Maximization is popular parametric approach for estimating the underlying density function of a set of data vectors in both the univariate and multivariate cases. Usually, the EM is used to approximate the underlying PDF using a mixture of Gaussian components. EM, however, sufferers from two major limitations. The first is that the number of components in the mixture must be *a priori* specified in order to obtain a reasonable estimate of the true PDF. This number must be accurately specified in order to balance the computational cost in both training and testing phases on one hand and the estimation accuracy on the other. The second limitation is that EM is highly sensitive to the initialization of the mean vectors and covariance matrices of the mixture.

Usually, the k-means algorithm (or similar algorithms) is used to initialize the mean vectors and covariance matrices of EM. Unfortunately, this approach sometimes drives EM towards the wrong mixture values. It sometimes also leads to numerical problems when estimating the covariance matrices.

Fig. 1.a shows the scatter plot of a bivariate data set drawn from a six-component Gaussian mixture. In Fig. 1.b the k-means algorithms correctly initialized the EM which helps convergence to the correct mixture parameters. In Fig. 1.c the k-means drives EM to converge to the wrong mixture parameters even though the data set has not changed. This example illustrates the significance of the initialization problem even when we know the true number of mixture components. The advantage of using EM here is its superior ability to infer the hidden structure of the data (assuming we can initialize it correctly).

Several attempts have been made to overcome the drawbacks of the EM algorithm. Figuerideo and Jain [4] broadly classify these methods into two categories: deterministic approaches and stochastic approaches. Deterministic methods, such as [5] and [6], are based on selecting the number of components according to some model selection criterion, which usually contains an increasing



(a) Scatter plot of bivariate data set withdrawn from a 6-component mixture model

(b) The k-means helps the convergence to the correct mixture parameters

(c) The k-means biases the EM to converge to the wrong parameters

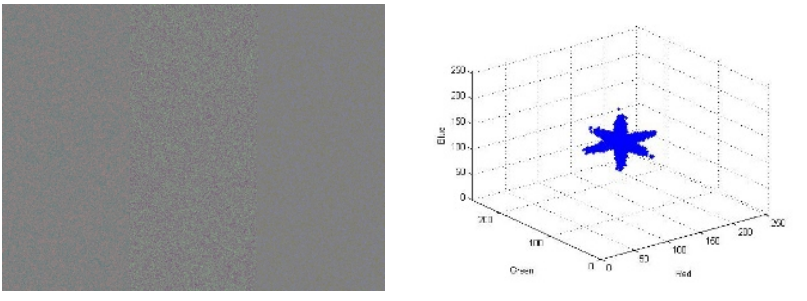
Fig. 1. Mixture parameter estimation using EM initialized by K-means

function that penalizes higher number of components. In [7] and [8] stochastic approaches based on Markov Chain Monte Carlo methods are used.

2.2 The Mean-Shift

The mode-finding algorithm introduced in [9] which is based on the mean-shift algorithm [10], compared to the k-means, consistently converges to the modes of the underlying density function. Therefore, for example, when applied to the data set in Fig. 1, the mean-shift successfully finds the two local maximum at $[0,\ -10]^T$ and $[0,\ 10]^T$ respectively.

The limitation of the mean-shift-based mode-finding algorithm is its inability to infer the hidden structure of the data. For example, Fig. 2.a shows a noisy image with a hidden structure (the reader is encouraged to refer to the electronic copy for a clearer image.) The 3D scatter plot of the RGB values of the image is shown in Fig. 2.b. When mean-shift is used to segment the image in Fig. 2, the result is a gray image with all pixels set to $[128,\ 128,\ 128]^T$ since mean-shift is technically "blind" to the structure of the data.



(a) A noisy image with a hidden structure

(b) The RGB scatter plot of the image in Fig. 2.a

Fig. 2. Noisy image with hidden structure

3 The Penalty-Less Information Criterion

Let $\mathrm{X} = \{\mathrm{x}_i\}_{i=1}^{N}$ be a set of N vectors to be modeled. We use EM to model the data by a mixture of k Gaussian components as shown in Equation 1,

$$p(\mathrm{x}|\Theta) = \sum_{i=1}^{k} \pi_i \, \mathcal{N}(\mathrm{x},\ \mu_i,\ \Sigma_i) \tag{1}$$

where μ_i, Σ_i and π_i are the mean, covariance and weight of component i, respectively and $\Theta = \{\mu_i s,\ \Sigma_i s,\ \pi_i s\}$ is the set of parameters of the k-component mixture model (such that $\sum_{i=1}^{k} \pi_i = 1$). The parameters set Θ can be estimated by applying the EM algorithm on X. The set X can now be clusterized into k subsets (i.e. clusters) using the Mahalanobis distance based on Θ_k, such that

$$\mathrm{X} = \bigcup_{i=1}^{k} \mathrm{X}^i \tag{2}$$

where X^i is the subset of vectors that belong to cluster i.

For each cluster, i, we compute two estimates of the probability density function (PDF) underlying the data. First, we compute a parametric estimate of the PDF as shown in Equation 3

$$p^i_{EM}(\mathrm{x}) = \pi_i \, \mathcal{N}(\mathrm{x}, \ \mu_i, \Sigma_i) \tag{3}$$

where p^i_{EM} indicates the EM-based estimate of the PDF of subset X^i. The second PDF is a kernel density estimate (KDE) of the PDF of the cluster data given by

$$p^i_{KDE}(\mathrm{x}) = \frac{1}{N^i} \sum_{j=1}^{N^i} \frac{1}{|\mathrm{H}_j|} \mathcal{K}\left(\frac{\mathrm{x} - \mathrm{x}^i_j}{\mathrm{H}_j}\right) \tag{4}$$

where N^i is the number of vectors in cluster i, H_j is the adaptive bandwidth matrix of vector j, x^i_j is vector number j of cluster i and $\mathcal{K}(.)$ is the kernel function. Since computing the kernel-based estimate of the PDF is computationally prohibitive in higher dimensions, we use the Improved Fast Guass Transform [11], which significantly reduces the complexity of the problem. The adaptive bandwidth is computed using the sample-point estimator of [12]. In this paper, we use the standard multivariate Gaussian as the kernel function.

We define the Penalty-less Information Criterion ($\mathcal{PIC}$) of a model with k components as the sum of weighted Jensen-Shannon divergence [13] between p^i_{EM} and p^i_{KDE} for all clusters as follows

$$\mathcal{PIC}_k = \sum_{i=1}^{k} \pi_i \ \mathcal{JSD}(p^i_{EM}, \ p^i_{KDE}) \tag{5}$$

where

$$\mathcal{JSD}(p^i_{EM}, \ p^i_{KDE}) = \frac{1}{2}\left(\mathcal{KLD}(p^i_{EM}, p^i_{Avg}) + \mathcal{KLD}(p^i_{KDE}, p^i_{Avg})\right), \tag{6}$$

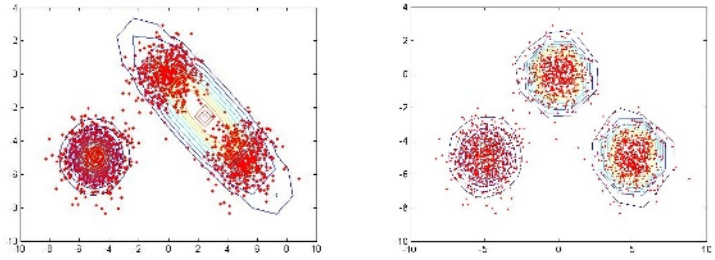
$$p^i_{Avg} = \frac{1}{2}\left(p^i_{EM} + p^i_{KDE}\right) \tag{7}$$

and

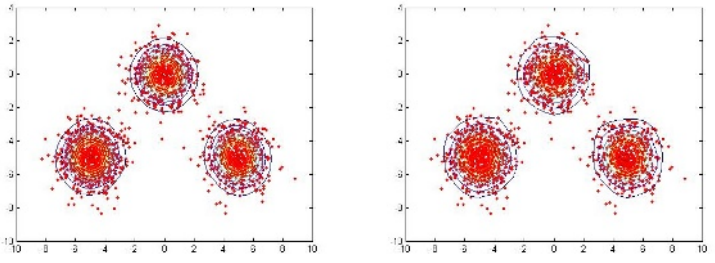
$$\mathcal{KLD}(p_1, \ p_2) = \int_{\forall \mathrm{x}} p_1(\mathrm{x}) \ \log\left(\frac{p_1(\mathrm{x})}{p_2(\mathrm{x})}\right) \ d\mathrm{x} \tag{8}$$

Jensen-Shannon divergence is used here because it is a symmetric version of Kullback-Leibler $\mathcal{KLD}$ divergence. Symmetry is important to equally emphasize both estimates of the PDF; i.e. p^i_{KDE} and p^i_{EM}. To determine the model $\hat{k}$ that best represents the data set X, $\mathcal{PIC}$ is computed for a range of possible mixture components and the mixture with a minimum $\mathcal{PIC}$ is selected as shown in Equation 9.

$$\hat{k} = \arg_k \min \mathcal{PIC}_k \quad \text{and} \quad k = k_{min}, \ \ldots, \ k_{max} \tag{9}$$



(a) p^1_{EM} and p^2_{EM} (b) p^1_{KDE} and p^2_{KDE}



(c) p^1_{EM}, p^2_{EM} and p^3_{EM} (d) p^1_{KDE}, p^2_{KDE} and p^3_{KDE}

Fig. 3. Data generated from three bivariate normal distributions. a and b:) fitting with a 2-component mixture, c and d:) fitting with 3-component mixture.

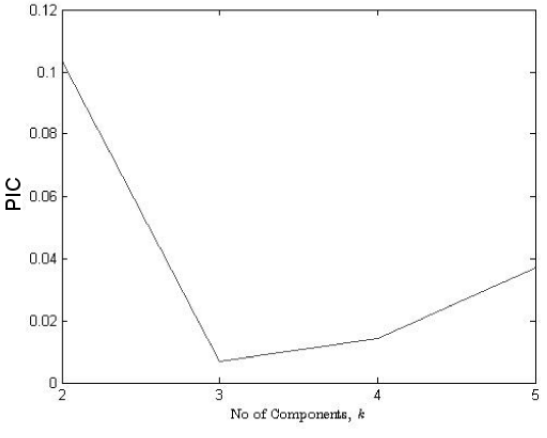


Fig. 4. The $\mathcal{PIC}$ for fitting the bivariate data of Fig. 3 using $k_{min} = 2$ and $k_{min} = 5$. The $\mathcal{PIC}$ is minimum at the correct number of components.

The search procedure is typical for many model selection criteria such as Bayesian Information Criterion (BIC). In that sense, $\mathcal{PIC}$ can be used alone as a model selection criterion. However, we will show in the next sections that within the proposed algorithm k_{min} and k_{max} are indeed constants.

Fig. 3 shows a simple example where bivariate data is generated using three normal distributions. The result of fitting the data using a two-component Gaussian mixture is shown in Fig. 3.a. The corresponding kernel density estimates of the two clusters is shown in Fig. 3.b. Because of the clear mis-modeling, the $\mathcal{PIC}$ value becomes relatively large. Fig. 3.c shows the result of fitting the data using a three-component mixture, which is similar to the kernel density estimates of the three clusters. As a result, the $\mathcal{PIC}$ value produced is relatively small. Repeating the same procedure for the range $k = \{2,\ 3,\ 4,\ 5\}$ results in the $\mathcal{PIC}$ values of Fig. 4, which has a clear minimum at the correct number of components.

4 Nonparametric EM Using Mixtures of Mixtures

Let $\mathrm{Y} = \{\mathrm{x}_i\}_{i=1}^{M}$ be a set of M vectors to be modeled. Here we use Y instead of X to denote the entire data set for reasons that will become clear later on. If we apply the mean-shift mode finding algorithm, proposed in [9], and only retain the modes with positive definite Hessian, we will obtain a set of m modes $\mathrm{Y}_c = \{\mathrm{x}_{c_j}\}_{j=1}^{m}$ which represent the local maxima points of the density function, where $m \ll M$. For details on computing the Hessian, the reader is referred to Han et al.'s method [14].

To infer the structure of the data, we start by partitioning Y into m partitions each of which corresponds to one of the detected modes. For all vectors of Y we compute a Mahalanobis-like distance δ defined by:

$$\begin{aligned} &\delta(\mathrm{x}_i|j) = (\mathrm{x}_i - \mathrm{x}_{c_j})^T\ \mathrm{P}_j\ (\mathrm{x}_i - \mathrm{x}_{c_j})^T, \\ &i = 1,\ 2,\ \ldots,\ M \quad \text{and} \\ &j = 1,\ 2,\ \ldots,\ m \end{aligned} \tag{10}$$

where P_j is the Hessian of mode j. The rationale here, as explained in [14] is to replace the covariance matrix, which may not be accurate at this point, by the Hessian which represents the local curvature around the mode x_{c_j}. Each vector is then assigned to a specific mode according to Equation 11.

$$\mathcal{C}(i) = \arg_j \min \delta(\mathrm{x}_i|j) \text{ and } j = 1,\ 2,\ \ldots,\ m \tag{11}$$

The data set can now be partitioned as

$$\mathrm{Y} = \bigcup_{j=1}^{m} \mathrm{Y}^j \tag{12}$$

where

$$\mathrm{Y}^j = \{\forall \mathrm{x}_i \in \mathrm{Y};\ \mathcal{C}(i) \equiv j\} \tag{13}$$

It is important to note here that the partitioning of Equation 12 is different than that of Equation 2.

Each of the detected modes corresponds to either a single Gaussian, such as those of Fig. 3.a, or a mixture of more than one Gaussian such as that in Fig. 1.a.

To determine the complexity of density around a given mode x_{c_j}, we model the partition data Y_j using a mixture of Gaussians specific to partition j. In other words,

$$p(\mathrm{x}|\Theta^j) = \sum_{i=1}^{k} \pi_i \, \mathcal{N}(\mathrm{x},\ \mu_i,\ \Sigma_i) \tag{14}$$

where Θ^j is the parameter set of the mixture associated with mode x_{c_j}. The initial values for the mean vectors are all set to x_{c_j}. The initial values for the covariance matrices are all set to P_j.

Since the structure of the data around x_{c_j} is unknown, we repeat the process for a search range of mixture complexities $[k_{min},\ k_{max}]$ and compute $\mathcal{PIC}$ for each complexity. The mixture that minimizes the $\mathcal{PIC}$ is chosen to represent the given partition.

Applying the Penalty-less Information Criterion to all partitions results in m mixtures of Gaussians with different complexities. The underlying density of the entire data set Y is now modeled as a *mixture of mixtures of Gaussians* as follows

$$p(\mathrm{x}|\Theta) = \sum_{j=1}^{m} \omega_j \ p(\mathrm{x}|\Theta^j) \tag{15}$$

where $\Theta = \{\Theta^j,\ \omega_j; j = 1,\ 2,\ \ldots,\ ,\ m\}$ is the set of all parameters. (Note that we extend the notation Θ here.) Finally, the weights of the mixtures ω_js are computed according to Equation 16.

$$\omega_j = \frac{\sum_{i=1}^{M} p(\mathrm{x}_i|\Theta^j)}{\sum_{j=1}^{m}\sum_{i=1}^{M} p(\mathrm{x}_i|\Theta^j)} \tag{16}$$

Algorithm 1 summarizes the proposed algorithm.

Algorithm 1. Nonparametric EM

Data: $\mathrm{Y} = \{\mathrm{x}_1,\ \mathrm{x}_2,\ \ldots,\ \mathrm{x}_i,\ \ldots,\ \mathrm{x}_M\}$
Result: $\Theta^j s$ and $\omega^j s$
begin
  $modes,\ m \longleftarrow MeanShift(\mathrm{Y})$
  $\mathrm{Y}^i s \longleftarrow PartitionFeatureSpace(\mathrm{Y},\ modes)$
  for $j \leftarrow 1$ **to** m **do**
    $\mathrm{X} \longleftarrow \mathrm{Y}^j$
    $N \longleftarrow M^j$
    for $k \longleftarrow k_{min}$ **to** k_{max} **do**
      $InitializeAllMeansAtTheModeLocation()$
      $InitializeAllCovariancesAtTheModeLocation()$
      $\mathrm{PIC}_k,\ \Theta_k \longleftarrow ComputePIC(X,k)$
    $\hat{k}^j \longleftarrow \arg_k \min \mathrm{PIC}_k$
    $\Theta^j \longleftarrow \Theta_{\hat{k}^j}$
  $\omega_i s \longleftarrow EstimateMixtureWeights(\mathrm{Y})$
end

5 Experimental Results

5.1 Synthetic Data Example

Fig. 5 shows a set of bivariate vectors generated from a four-component Gaussian mixture. Three of the Gaussian components are co-centered at $[.5,\ .75]^T$ but with different covariance matrices. The fourth component is a simple component centered at $[.5,\ .2]^T$. The modes detected using mean-shift [9] are overlaid and marked by crosses. The result of the partitioning procedure of Equation 11 is also shown where the green points indicate Y^1 and the blue points indicate Y^2. The mode-based partitioning results in two partitions with different hidden structures. For each partition separately, the PIC is computed in the range $[k_{min} = 1,\ k_{max} = 4]$. In our experiments we use $k_{max} = 2^d$, where d is the dimensionality of the data. Fig. 6.a shows that the correct number of mixture components, $\hat{k}$ for the first partition Y^1 is $\hat{k} = 1$. On the other hand, the correct

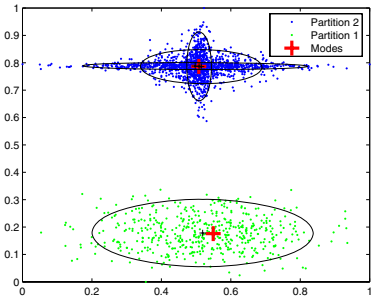
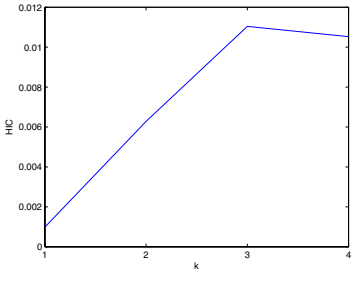
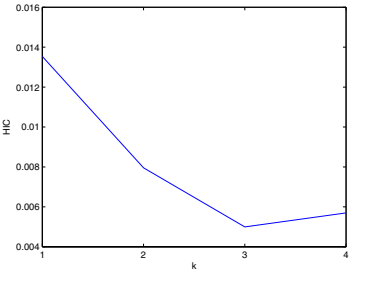


Fig. 5. Bivariate data generated from a 4-Gaussian mixture. The modes detected by the mean-shift are overlaied on the scatter plot.



(a) The PIC values for differnet mixture complexities for the first partition, Y^1, of Fig. 5. The minimum PIC value corresponds to $k = 1$ mixture, which is the correct mixture



(b) The PIC values for differnet mixture complexities for the second partition, Y^2, of Fig. 5. The minimum PIC value corresponds to $k = 3$ mixture, which is the correct mixture

Fig. 6. The PIC values for different partitions Y^j

number of components of the second partition Y^2 is $\hat{k} = 3$, due to the apparent complex structure of the data, as shown in Fig. 6.b.

5.2 Real Data Examples

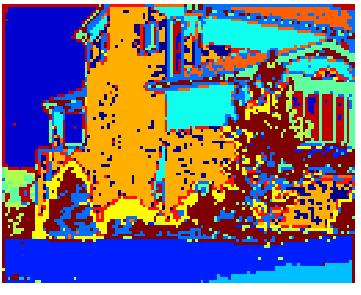
To verify the performance of the proposed algorithm, it has been applied in an image segmentation setting. The results are compared to the those of standard model selection methods. The Luv color space was used throughout the following experiments.

Fig. 7.b shows the result of segmenting Fig. 7.a using the standard EM algorithm. The number of mixture components are selected using BIC as given by Equation 17.

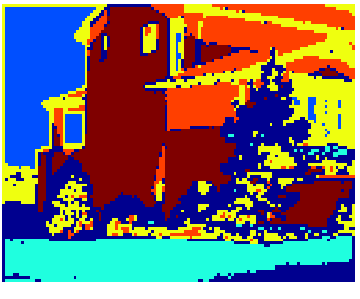
$$\hat{k} = \arg_k \min -2 \log p(Y|\theta_k) + v_k \ \ln M \tag{17}$$

where θ_k and v_k are mixture model with k components and number of free parameters in θ_k, respectively. The initial values of the means and covariance matrices are selected using K-means. It is clear that the model over-segments the

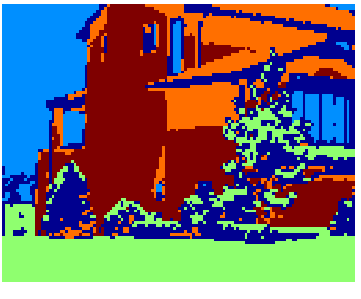
(a) Original Image



(b) Segmented image using EM initialized using K-Means. BIC was used to select the best model

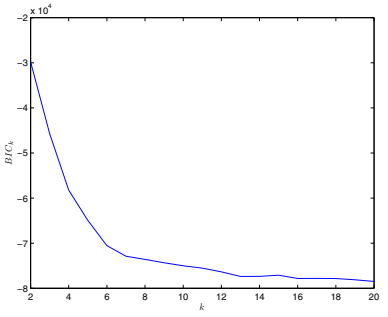


(c) Segmented Image using EM, with manually set number of mixture components

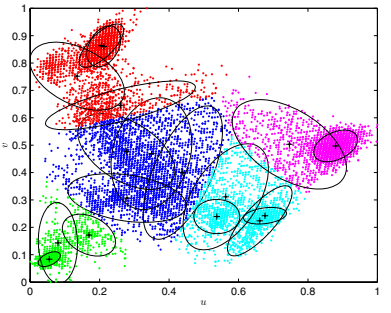


(d) Segmented Image using the proposed algorithm

Fig. 7. Image segmentations comparing the proposed algorithm against other traditional model selection methods



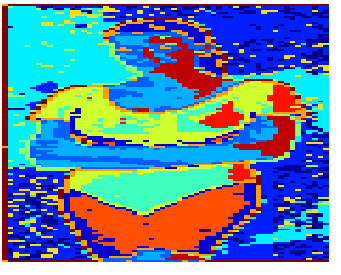
(a) BIC values for Fig. 7



(b) Hidden structures discovered by the proposed algorithm. Different colors correspond to different modes of the density function

Fig. 8. BIC values and hidden structure of Fig. 7

(a) Original Image

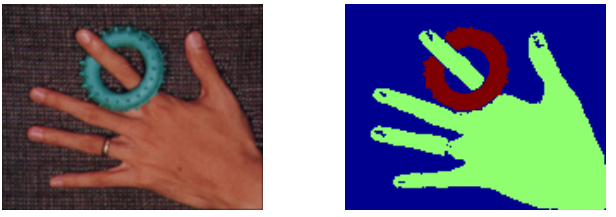


(b) Segmented image using EM initialized using K-Means. BIC was used to select the best model

(c) Segmented Image using EM, with manually set number of mixture components

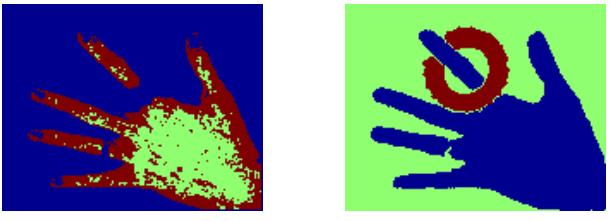
(d) Segmented Image using the proposed algorithm

Fig. 9. Image segmentations comparing the proposed algorithm against other traditional model selection methods



(a) Original Image

(b) Segmented Image using EM, with manually set number of mixture components. Good convergence of K-means



(c) Segmented Image using EM, with manually set number of mixture components. Bad convergence of K-means

(d) Segmented Image using the proposed algorithm

Fig. 10. Instability of the classical methods. Two different runs of the EM give different segmentation results.

image. The reason is that BIC only depends on the log likelihood of the feature vectors penalized by an increasing function of the number of free parameters.

The segmentation result in Fig. 7.c is obtained by visually inspecting the image and manually setting the number of mixture components to the correct value and using K-means to set the initial values of model parameters. The segmentation result are better than Fig. 7.c but not as accurate as desirable. Also, when the same experiment is repeated, the segmentation result will be different because of the random nature of the K-means.

Finally, Fig. 7.d shows the segmentation result using the proposed algorithm. The result is more accurate than the previous ones. Also, we are guaranteed to obtain the same result when repeating the experiment because the mean-shift must converge to the same stationary points every time it is applied.

Fig. 5.1.a shows the BIC values for Fig. 7.a. The curve does not have a local minima that suggests an appropriate model complexity. The ellipses in Fig. 5.1.b illustrate the hidden structures discovered by the proposed algorithm. Different modes of the density function require mixtures of variable complexities.

The same experiments was repeated for Fig. 9.a. Fig. 9.b shows the segmentation result of the EM with the BIC used to find the best model and the K-means used to initialize the model. The result of a manually selecting the number of mixture components is shown in Fig. 9.c. Finally, the result of the

Table 1. Number of data points and normalized run-time for model selection using both BIC search and proposed algorithm

	No. of points	BIC search run-time	Our run-time
Synthetic	2000	0.1583	0.0031
Woman	7598	0.0132	0.0041
House	12288	0.0525	0.0038
Hand	18544	0.2089	0.0040

proposed algorithm is shown in Fig. 9.d. The proposed algorithm yields better segmentation.

Fig. 10 illustrates the instability of using EM for density estimation and modeling. In both Fig. 10.b and Fig. 10.c the image is segmented using the EM with manually set number of mixture components and the K-means for initializing the mixture parameters. In Fig. 10.b the K-means helps the EM to converge to a good mixture, which yields a good segmentation. However, with the same number of components, K-means biases the EM to converge to a bad mixture, which results in a very poor segmentation result.

Table 1 compares the run-times of finding the best mixture model using BIC search ($k_{min} = 1$ and $k_{max} = 10$ and using our mixture of mixtures algorithm. The run-times are normalized by the number of data points. The large difference is due to two main reasons. The first is that BIC search is performed on the entire data set while our algorithm is based on partitioning the data set before applying the $\mathcal{PIC}$ search. The second reason is that the BIC search is applied on a wide range of potential complexities (because it uses the entire data set) while the $\mathcal{PIC}$ search is applied on a smaller range because of its local nature.

6 Conclusions and Future Research

This paper introduces a new method addressing the limitations of the popular Expectation Maximization algorithm; namely the *a priori* knowledge regarding the complexity of the mixture and difficulty of accurate initialization of mixture parameters.

The paper uses the mean-shift-based mode finding algorithm developed by Comaniciu and Meer in [9] to estimate the number of Gaussian mixtures that must be used to model the data. Then, a partitioning algorithm is performed to cluster the data into subsets. For each subset the regular EM is used to infer the hidden structure of the underlying density function. The mean vectors of the mixture is initialized at the mode location found by the mean-shift.

The paper also introduces a model selection criterion, $\mathcal{PIC}$, that is used to find the mixture that best fits the density of the mixture. The $\mathcal{PIC}$ compares the parametric representation of the density against a nonparametric estimate of PDF using the Jensen-Shannon divergence, without a penalty term.

Applying the proposed algorithm on 2D and 3D data sets shows the advantages of using the algorithm to obtaining a parametric representation of the

underlying density without manually initializing the model. In the future, we plan to apply the proposed algorithm to other computer vision problems and compare its performance against other popular image segmentation algorithms.

References

1. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood form incomplete data via the em algorithm. Journal of Royal Satistical Society, Series B **39** (1977)
2. Belongie, S., Carson, C., Greenspan, H., Malik, J.: Color- and texture-based image segmentation using em and its applications to content-based image retreval. In: Sixth International Conference on Computer Vision. (1998) 675–682
3. Abu-Naser, A., Galatsanos, N., Wernick, M., Schonfeld, D.: Object recognition based on impulse restoration with use of the expectation-maximization algorithm. Journal of the Optical Society of America A (Optics, Image Science and Vision) **15** (1998) 2327 – 40
4. Figueirdeo, M., Jain, A.: Unsupervised Learning of Finite Mixture Models. IEEE Trans. on Pattern Analysis and Machine Intelligence **24** (2002) 381–396
5. Dasgupta, A., Rafetry, A.: Detecting Features ni Spatial Point Patterns with Clutter Via Model-Based Clustering. J. of American Statistical Association (1998) 294–302
6. Campbell, J., Fraley, C., Murtagh, F., Raftery, A.: Linear Flaw Detection in Woven Textiles Using Model-Based clustering. Pattern Recognition Letters **18** (1997) 1539–1548
7. Neal, R.: Bayesian Mixture Modeling. In: Proc. of 11th Int'l Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis. (1992) 197–211
8. Bensmail, H., Celeus, G., Rafetry, A., Robert, C.: Inference in Model-Based Cluster Analysis. Statistics and Computing **7** (1997) 1–10
9. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence **24** (2002)
10. Fukunaga, K., Hostetler, L.D.: The estimation of a gradient of a density function, with applications in pattern recognition. IEEE Trans. on Information Theory **21** (1975) 32–40
11. Yang, C., Duraiswami, R., Gumerov, N., Davis, L.: Improved Fast Gauss Transform and Efficient Kernel Density Estimation. In: Proc. IEEE International Conference on Computer Vision. (2003)
12. Wand, M., Jones, M.: Kernel Smoothing. Chapman and Hall (1995)
13. Lin, J.: Divergence measuers based on the shannon entropy. IEEE Trans. Information Theory **37** (1991) 145–151
14. Han, H., and, D.C., Zhu, Y., Davis, L.: Incremental density approximation and kernel-based baesian filtering for object tracking. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2004)