ICAPR'2003 proceedings - 5th International Conference on Advances in Pattern Recognition
Calcultta (India), 10-13 December 2003, Allied Publishers Pvt. Limited, New Delhi, pp. 81-84

Fully Nonparametric Probability Density Function Estimation with Finite Gaussian Mixture Models

Cédric Archambeau and Michel Verleysen*
Université catholique de Louvain
Machine Learning Group
Place du Levant 3
B-1348 Louvain-la-Neuve
Belgium
archambeau@dice.ucl.ac.be

Abstract

Flexible and reliable probability density estimation is fundamental in unsupervised learning and classification. Finite Gaussian mixture models are commonly used to serve this purpose. However, they fail to estimate unknown probability density functions when used for nonparametric probability density estimation, as severe numerical difficulties may occur when the number of components increases. In this paper, we propose fully nonparametric density estimation by penalizing the covariance matrices of the mixture components according to the regularized Mahalanobis distance. As a consequence, the singularities in the log-likelihood function are avoided and the quality of the estimation models is significantly improved.

1. Introduction

Probability density function (PDF) estimation is of major concern in areas such as machine learning, pattern recognition, neural networks, signal processing, computer vision and feature extraction. On the one hand, it offers a flexible way to investigate the properties of a given data set and provides a solid basis for efficient data mining tools. On the other hand, it is crucial in unsupervised learning tasks and Bayesian inference and classification.

While performing density estimation, three main alternatives may be considered. The first one, known as parametric density estimation, assumes the data is drawn from a specific density model. The model parameters are then fitted to the data. Unfortunately, in many cases an a priori choice of the PDF model is not suited since it might provide a false representation of the true PDF.

An alternative is to build nonparametric PDF estimators, as for example the Parzen window estimator [7]. For a comprehensive overview of nonparametric density estimation techniques, we refer to [4]. When using Parzen windowing, the PDF is estimated by placing a well-defined kernel function on each data point with an optimized common kernel width h. In practice, Gaussian kernels are often used. The estimated PDF is computed by averaging the Gaussian densities in each data point. By contrast to the previous method, this technique does not assume any functional form of the PDF and allows its shape to be entirely determined from the data.

A third approach consists in using semi-parametric models. As nonparametric techniques, they do not assume an a priori shape of the PDF to estimate. However, for this model family, the complexity is fixed in advance, avoiding a prohibitive increase of the number of parameters with the size of the data set. Finite mixture models, and in particular finite Gaussian mixtures (FGM), are commonly used to serve this purpose. A popular technique for approximating the maximum likelihood estimate (MLE) of the underlying PDF is the expectation-maximization (EM) algorithm, formalized by Dempster, *et al.* [3].

Finite mixture models are highly effective when an appropriate guess of the number of components in the mixture is pretty obvious. Unfortunately, when one wants to perform fully nonparametric PDF estimation by increasing the number of components arbitrarily, numerical difficulties arise due to singularities in the likelihood function. In [1], the convergence problems of the EM in conjunction with FGM have been traced and linked to the concept of isolation. Whereas isolated data points appear in sparse data sets or when the data set includes outliers, they can also occur in the the tails of the PDF.

In this paper, the regularized Mahalanobis distance is proposed in order to avoid the singularities in the likelihood function. The regularization term acts directly on the covariance matrices of the multivariate mixture components

*Michel Verleysen is a Senior Research Associate of the Belgian F.N.R.S. (National Fund for Scientific Research).

This work was partially supported by the European Commission (grant IST-2000-25145 OPTIVIP) and the Belgian F.M.S.R. (project 3.4590.02).

according to the prior belief that diagonal covariance matrices is an assumption of regularity. As a result, one can perform fully nonparametric PDF estimation by FGM without facing numerical difficulties. Meanwhile, the quality of the model estimates is improved compared to classical nonparametric techniques, such as Parzen windowing.

The paper is organized as follows. In Section 2, FGMs are recalled, as well as the computation of the MLE of the model parameters by the EM algorithm. In Section 3, regularization of the covariance matrices of the mixture components is introduced by modifying the M-step of the EM algorithm. Finally, in Section 4, experimental results are presented and discussed.

2. Finite Gaussian Mixture Models

Let us consider a d-dimensional continuous random vector $X \in \Re^d$. Its PDF can be approximated by a Gaussian mixture model [6], defined as a linear combination of M Gaussian component densities:

$$p(\mathbf{x}) = \sum_{j=1}^{M} P(j) p(\mathbf{x}|j), \tag{1}$$

where the mixing proportions $P(j)$ are non-negative and must sum to one. The Gaussian components are characterized by their centers $\mathbf{c}_j$ and their covariance matrices Σ_j:

$$\begin{aligned} p(\mathbf{x}|j) &= (2\pi)^{-\frac{d}{2}} |\Sigma_j|^{-1/2} \\ &\cdot \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{c}_j)^T \Sigma_j^{-1} (\mathbf{x}-\mathbf{c}_j)\right]. \end{aligned} \tag{2}$$

Consider an i.i.d. realization $\chi = \{\mathbf{x}_n\}_{n=1}^N$ of X. Based on the mixture model, we may define the log-likelihood function:

$$L(\theta) = \log \prod_{n=1}^{N} p(\mathbf{x}_n). \tag{3}$$

In this equation θ summarizes the model parameters $P(j)$, $\mathbf{c}_j$ and Σ_j. By means of the the EM algorithm, the MLE of θ can be computed iteratively, avoiding the intricacy of non-linear optimization schemes. Maximizing the log-likelihood function is then equivalent to finding the most probable PDF estimate provided the data set χ.

In order to compute the MLE of the log-likelihood function the EM operates in two stages. First, in the E-step, the expected value of some “unobserved” data is computed, using the current parameter estimates and the observed data. Here the “unobserved” data indicates which data sample was generated by which component in the mixture. Subsequently, during the M-step, the expected values computed in the E-step are used to compute the MLE and the model parameters are updated. Each iteration step t can be summarized as follows [6]:

E-step:

$$P^{(t)}(j|\mathbf{x}_n) = \frac{p^{(t)}(\mathbf{x}_n|j) P^{(t)}(j)}{p^{(t)}(\mathbf{x}_n)}. \tag{4}$$

M-step:

$$\mathbf{c}_j^{(t+1)} = \frac{\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n)}, \tag{5}$$

$$\begin{aligned} \Sigma_j^{(t+1)} &= \\ &\frac{\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n)\left(\mathbf{x_n}-\mathbf{c_j^{(t+1)}}\right)\left(\mathbf{x}_n-\mathbf{c}_j^{(t+1)}\right)^T}{\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n)}, \end{aligned} \tag{6}$$

$$P^{(t+1)}(j) = \frac{1}{N}\sum_{n=1}^{N} P^{(t)}(j|\mathbf{x}_n). \tag{7}$$

In this equation set, $P^{(t)}(j|\mathbf{x}_n)$ corresponds to the posterior probability that $\mathbf{x}_n$ is generated by component j provided that the data point $\mathbf{x}_n$ is known.

3. Regularized Mahalanobis Distance

Finite mixture models can approximate any continuous PDF, provided the model has a sufficient number of components and provided the parameters of the model are chosen correctly [2]. In addition, if sufficient data samples are available and the singularities of the likelihood function can be avoided, we may approximate the true PDF arbitrarily well by increasing the number of components. In order to recover from singular sample covariance matrices, the regularized Mahalanobis distance is proposed.

The Mahalanobis distance D_M is defined as:

$$D_M(\mathbf{x}_n, \mathbf{c}_j) = (\mathbf{x}_n - \mathbf{c}_j)^T \Sigma_j^{-1} (\mathbf{x}_n - \mathbf{c}_j). \tag{8}$$

From (2), one can easily see that the multivariate Gaussian disribution uses D_M to determine its shape. When the number of data samples contributing to the computation of the covariance matrix of a component is small with respect to the dimension d of the data samples, it may be singular. Moreover, as discussed in [5], D_M tends to produce hyperellipsoidal components, leading to unusually large and elongated densities. By contrast, when one considers the Euclidean distance D_E, large data clusters tend to split unnecessarily, as the component densities are hyperspherical. The Euclidean distance is defined as:

$$D_E(\mathbf{x}_n, \mathbf{c}_j) = (\mathbf{x}_n - \mathbf{c}_j)^T (\mathbf{x}_n - \mathbf{c}_j). \tag{9}$$

Based on the hyperspherical character of D_E and the hyperellipsiodal character of D_M, one can construct a regularized Mahalonbis distance D_{ME}, which is a linear combination of both distances [5]:

$$D_{ME}(\mathbf{x}_n, \mathbf{c}_j) = (\mathbf{x}_n - \mathbf{c}_j)^T \left[(1-\lambda)\left(\Sigma_j + \epsilon I\right)^{-1} + \lambda I\right] \cdot \left(\mathbf{x}_n - \mathbf{c}_j\right), \tag{10}$$

where ϵ and λ are learning parameters, and I is the $d \times d$ identity matrix. Note that parameter λ is included in the interval $[0, 1]$. It controls the tradeoff between hyperspherical and hyperellipsoidal components. Therefore, when the covariance matrices cannot be estimated reliably, a large value of λ should be used.

Parameter λ should be learnt properly, as it achieves the tradeoff between D_M and D_E. By contrast, a careful estimation of ϵ is not required. Indeed, its role is to stabilize the learning process by converting a singular matrix to a nonsingular one. Therefore different values of ϵ do not make much difference as long as they are significantly smaller than the average variance of the data samples. This observation is illustrated in section .

Consider again the E- and M-step for computing the model parameters of FGM (see (4) to (7)). Introducing the regularized Mahalanobis distance consists in adapting, at each iteration step t, the covariance matrix of each component density according to (10). Therefore, the following adaptation rule is inserted in the M-step:

$$\left(\Sigma_j^{(t+1)}\right)_{ME} = \left[(1-\lambda)\left(\Sigma_j^{(t+1)} + \epsilon I\right)^{-1} + \lambda I\right]^{-1}, \tag{11}$$

where $\Sigma_j^{(t+1)}$ is still computed according to (6).

4. Experimental Results

In this section, we investigate the quality of the estimated PDF with respect to the true PDF in a simple toy problem. Furthermore, it is shown that the regularized FGM avoids singularities and improves the quality of the estimation models.

Let's consider the random variable X, which is a mixture of two overlapping normal densities:

$$X \sim N(\mathbf{m}_1, S_1) + N(\mathbf{m}_2, S_2). \tag{12}$$

Assume we do not know this density. Next, consider an i.i.d. realization $\chi = \{\mathbf{x}_n\}_{n=1}^N$ of X and suppose it is corrupted by an additive Gaussian noise $n \sim N(0, \sigma_n^2)$.

A convenient quality measure when comparing PDFs is the Kullback-Leibler divergence D_{KL}. It is defined as follows:

$$D_{KL}(p^*, p) = \int_X p^*(\mathbf{x}) \log \frac{p^*(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}, \tag{13}$$

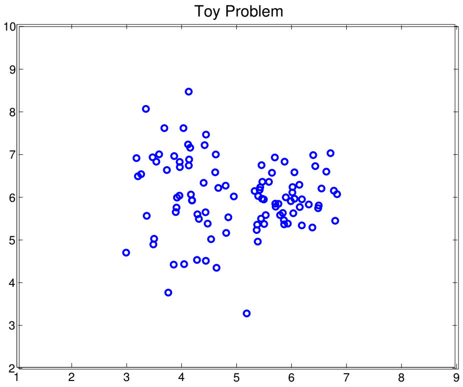

Figure 1: Mixture of two overlapping Gaussian distributions, corrupted by additive Gaussian noise ($\sigma_n = 0.05$)

Table 1: Quality of the PDF estimation models with respect to the true PDF of X. The symbol † means that numerical problems have occurred when computing the model parameters by EM. ($Q = 25$, $t_{end} = 150$)

	M	λ_{opt}	ϵ	$E[D_{KL}]$	$S[D_{KL}]$
Parzen	100	–	–	0.165	0.020
FGM	3	–	–	0.134	0.069
	> 3	–	–	†	†
Reg. FGM	3	0.2	10^{-5}	0.117	0.036
	5	0.3	10^{-5}	0.088	0.029
	7	0.3	10^{-5}	0.109	0.035
	10	0.4	10^{-5}	0.107	0.048
	15	0.4	10^{-5}	0.115	0.038

where $p^*(\mathbf{x})$ is the true PDF. When D_{KL} is zero, both densities are identical.

In the toy problem, the following numerical values are used:

$$\mathbf{m}_1 = [4, 6], \quad S_1 = \begin{bmatrix} 0.5^2 & 0 \\ 0 & 1.5^2 \end{bmatrix},$$
$$\mathbf{m}_2 = [6, 6], \quad S_2 = \begin{bmatrix} 0.5^2 & 0 \\ 0 & 0.5^2 \end{bmatrix}.$$

A realization of X is represented in Fig. 1. The standard deviation of the additive Gaussian noise is $\sigma_n = 0.05$. The data set contains 100 data samples.

In Table 1, one finds the model performance, that is $E[D_{KL}]$ and $S[D_{KL}]$, of the Parzen window estimator, the classical FGM and the regularized FGM, for a varying model complexity M. In this experiment, it was found that the optimal kernel width for Parzen windowing is $h = 0.5$. Both for FGM and regularized FGM, a sufficient number of iterations before stopping to ensure convergence of the EM is $t_{end} = 150$. We have considered 25 realizations of X.

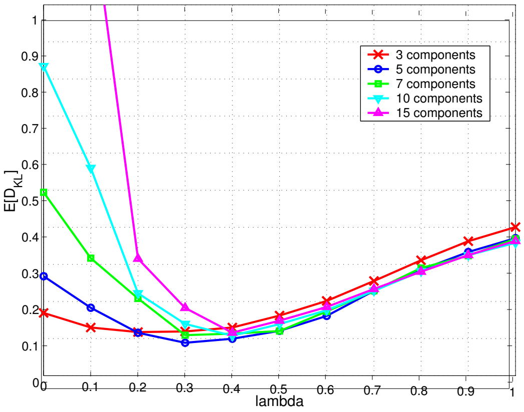

Figure 2: Expected Kullback-Leibler divergence with respect to the learning parameter λ ($\epsilon = 10^{-5}$). Each curve corresponds to a different model complexity. ($Q = 25$, $t_{end} = 150$)

In order to reliably estimate D_{KL}, Q realizations of X have been considered, therefore leading to Q estimation models. As a consequence, one may compute the estimated expected value $E\left[D_{KL}\right]$ of the Kullback-Leibler divergence, as well as its estimated standard deviation $S\left[D_{KL}\right]$. Using an expected value reduces the bias on the performance estimation D_{KL}, whereas its standard deviation gives an indication of the reliability we may put in the estimated models.

The quality of FGM containing 3 components is greater than the optimal Parzen window estimator. Unfortunately, when one wants to enhance the accuracy by increasing the number of components in the mixture (see > 3 in Table 1), numerical difficulties appear due to singularities in the log-likelihood function. As discussed in [1], this leads to the width of a component to tend to zero, and therefore the component to collapse (denoted by † in Table 1).

By contrast, when using regularized FGM the numerical difficulties are avoided. As a result, fully nonparametric PDF estimation is possible; furthermore, one can optimize the model complexity by an exhaustive search on the number of components. In addition, the consistency of the estimation model is enhanced.

In Fig. 2, the performance of regularized FGM is drawn with respect to parameter λ, for different model complexities. Although they all perform similarly, the optimal choice for the number of components M is 5.

Finally, note that in our experiments we have fixed ϵ in advance. This learning parameter has little influence on the quality of the models. Indeed, as illustrated in Fig. 3, the iso-contours of D_{KL} are independent from ϵ, provided it is chosen sufficiently small. We have found that $\epsilon = 10^{-5}$ performs well for all the considered model complexities.

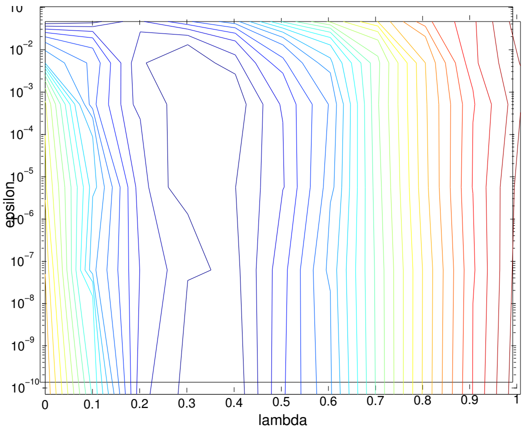

Figure 3: Iso-contour plot of the expected Kullback-Leibler divergence with respect to λ and ϵ. The PDF estimation model has a complexity of 5 components. ($Q = 25$, $t_{end} = 150$)

5. Conclusion

From a theoretical point of view, finite Gaussian mixture models can be used to perform nonparametric PDF estimation, by increasing the number of components in the mixture arbitrarily. In practice, however, one has to face numerical difficulties, as a component may collapse. Nevertheless, by introducing the regularized Mahalanobis distance in the classical FGM model and stating that diagonal covariance matrices is an assumption of regularization, we have demonstrated that fully nonparametric FGM can be performed. Furthermore, the regularized FGM shows a greater model quality compared to the Parzen window estimator, which is currently one of the most popular and practical nonparametric PDF estimation techniques.

References

[1] C. Archambeau, J. A. Lee, and M. Verleysen. On the convergence problems of the *EM* algorithm for finite gaussian mixtures. In *Proc. 11th European Symposium on Artificial Neural Networks*, pages 99–106. Bruges, Belgium, April 2003.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *J Roy. Stat. Soc. (B)*, 39:1–38, 1977.

[4] A. J. Izenman. Recent developments in nonparametric density estimation. *J Am. Stat. Assoc.*, 86(413):205–224, 1991.

[5] J. Mao and A. K. Jain. A self-orgainizing network for hyperellipsoidal clustering (hec). *IEEE Trans. Neural Networks*, 7(1):16–29, 1996.

[6] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[7] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33:1065–1076, 1962.