Component-wise Parameter Smoothing for Learning Mixture Models

Chandan K. Reddy
Department of Computer Science
Wayne State University, Detroit, MI.

Bala Rajaratnam
Department of Statistics
Stanford University, Stanford, CA.

Abstract

In this paper, we propose a novel component-wise smoothing algorithm that constructs a hierarchy (or family) of smoothened log-likelihood surfaces. Our approach first smoothens the likelihood function and then applies the EM algorithm to obtain a promising solution on this smoothened surface. Using the most promising solutions as initial guesses, the EM algorithm is applied again on the original likelihood. This effective optimization procedure eliminates extensive search in the non-promising regions of the parameter space. Empirical results on some standard datasets show the reduction of the number of local maxima and improvements in the log-likelihood values.

1. Introduction

In the field of statistical pattern recognition, finite mixtures allow a probabilistic model-based approach to unsupervised learning from multivariate data. One of the most popular methods used for fitting mixture models to the observed data is the *Expectation-Maximization* (EM) algorithm which converges locally to the *Maximum Likelihood Estimate* (MLE) of the mixture parameters [1]. One of the main disadvantages of using the EM algorithm is that it is very sensitive to the initialization. Because of its greedy nature, EM algorithm tends to get stuck at a local maximum which will correspond to erroneous set of parameters for the mixture components. The log-likelihood surface on which the EM algorithm is applied is very rugged with many local maxima. The fact that the local maxima are not uniformly distributed across the entire search space makes it important for us to develop algorithms that help in avoiding search in non-promising regions. More focus needs to be given for searching the promising subspaces by obtaining promising initial estimates. Usually, a global method which incorporates the global structure of the parameter space guides the EM algorithm to obtain a more precise set of parameters which correspond to a higher likelihood function value. Though, several approaches for exploring the non-linear log-likelihood surface were proposed [3, 1], the idea of modifying the log-likelihood surface using component-wise smoothing techniques has not been investigated in the literature.

The smoothing procedure described in this paper is a way to estimate the optimal set of parameters of the Gaussian components in an effective manner. It is a procedure that reduces the ruggedness of the log-likelihood surface and has the capability to avoid non-promising regions during the search. Smoothing the log-likelihood surface can potentially obtain the set of promising regions which can then be used to gradually trace back the promising solutions on the original log-likelihood surface. In the convolution based component-wise smoothing approach, a simplified version of the global method is applied in combination with the EM algorithm to obtain an optimal set of parameters on the smoothened log-likelihood surface which are again used as initial parameters for the EM algorithm to obtain optimal set of parameters on the original log-likelihood surface. Though applied for two levels in this paper, our algorithm can be generically applied to any number of levels.

Table 1. Description of the Notation used in this paper.

Notation	Description
Θ	parameter space
$\tilde{\Theta}$	smooth parameter space
θ_i	parameters of a single i^{th} component
θ_0	parameters of the smoothing kernel
α_i	mixing weight for i^{th} component
$\mathcal{X}$	observed data
$\mathcal{Z}$	missing data

2 Mixture Models and the EM

Table 1 gives the notation used in this paper. Let us assume that there are k Gaussian components in the mixture model. The form of the probability density function is given as follows:

$$p(x|\Theta) = \sum_{i=1}^{k} \alpha_i p(x|\theta_i) \tag{1}$$

where $x = [x_1, x_2, ..., x_d]^T$ is the feature vector of d dimensions. Θ represents the collection of parameters $(\alpha_1, \alpha_2, ...\alpha_k, \theta_1, \theta_2, ...\theta_k)$ and p is a multivariate density function parameterized by θ_i. Given a set of n i.i.d. samples $\mathcal{X} = \{x^{(1)}, x^{(2)}, .., x^{(n)}\}$, the log-likelihood corresponding to a mixture is

$$log\ p(\mathcal{X}|\Theta) = \sum_{j=1}^{n} log \sum_{i=1}^{k} \alpha_i\ p(x^{(j)}|\theta_i) \tag{2}$$

The goal of learning mixture models is to obtain the parameters $\widehat{\Theta}$ from a set of n data points which are the samples of a distribution with density given by (1). The MLE is given as follows:

$$\widehat{\Theta}_{MLE} = arg \max_{\Theta}\ \{\ log\ p(\mathcal{X}|\Theta)\ \} \tag{3}$$

Since this MLE cannot be found analytically for mixture models, one has to rely on iterative procedures that can find the global maximum of $log\ p(\mathcal{X}|\Theta)$. The EM algorithm assumes $\mathcal{X}$ to be *observed* data. The missing part is a set of *n* labels $\mathcal{Z} = \{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, .., \mathbf{z}^{(n)}\}$ associated with n samples, indicating which component produced each sample [1].

The EM algorithm produces a sequence of estimates $\{\widehat{\Theta}(t), t = 0, 1, 2, ...\}$ by alternately applying two steps (Expectation and Maximization) until convergence. Several variants of the EM algorithm have been extensively used to solve this problem. One of the main challenges of the EM algorithm is the initialization step. In this paper, we explore the idea of convolution-based smoothing of the log-likelihood surface in order to reduce the number of local maxima thus diminishing the sensitivity to the initial parameters used.

3 Convolution Kernels and Smoothing

Let us consider a continuous mapping $p(x) : \mathbb{R}^d \rightarrow \mathbb{R}$. In scale-space theory, $p(x)$ is embedded into a continuous family $P(x, \sigma)$. Our method starts with an approximation of the entire dataset with Gaussian kernel of σ width. As the resolution (or scale) increases, the sigma value is reduced and eventually converges to zero. In simple terms, one can write the new kernel $p(x, \sigma)$ as a convolution of $p(x)$ with a Gaussian kernel $g(x, \sigma)$ as shown below:

$$P(x, \sigma) = p(x) \otimes g(x, \sigma) = \int p(x-y) \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\|y\|^2}{2\sigma^2}} dy$$

For smoothing the mixture model, any kernel can be used for convolution, if it can yield a closed form solution in each E and M step. Let us consider the following Gaussian kernel for smoothing and obtain the new density function.

$$g(x) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \tag{4}$$

$$\begin{aligned} p'(x|\theta_i) &= p(x|\theta_i) \otimes g(x) \\ &= \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \otimes \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \\ &= \frac{1}{\sqrt{2\pi(\sigma_i^2+\sigma_0^2)}} e^{-\frac{(x-(\mu_i+\mu_0))^2}{2(\sigma_i^2+\sigma_0^2)}} \end{aligned} \tag{5}$$

When a Gaussian density function with parameters μ_1 and σ_1 is convolved with a Gaussian kernel with parameters μ_0 and σ_0, then the resultant density function is also Gaussian with mean $(\mu_1 + \mu_0)$ and variance $(\sigma_1^2 + \sigma_0^2)$. Since shifting the mean is not a good choice for optimization problems and we are more interested in reducing the peaks, we chose to increase the variance parameter without shifting the mean.

Convolving the complete log-likelihood function using a Gaussian kernel directly might result in an analytic expression that may not be easy to handle. Hence, *component-wise convolution* is performed. This approach can smoothen different regions of the parameter space differently. Since the log-likelihood surface is obtained from individual densities, smoothing each component's individual function will smoothen the overall log-likelihood surface. The smooth log-likelihood function is given by:

$$f'(\mathcal{X}, \Theta) = \sum_{j=1}^{n} log \sum_{i=1}^{k} \alpha_i\ p'(x^{(j)}|\theta_i) \tag{6}$$

Fig. 1 shows the block diagram of the smoothing procedure. The original likelihood surface is obtained from the initial set of parameters and the given dataset. The kernel parameters are chosen from the initial set of parameters and the original likelihood surface. The kernel is then convolved with the original likelihood surface to obtain smooth likelihood surface. The parameters of the smoothing kernel can be chosen to be fixed so that they need not depend on the parameters of individual components. Fixed kernels will be effective when the underlying distribution comes from similar components.

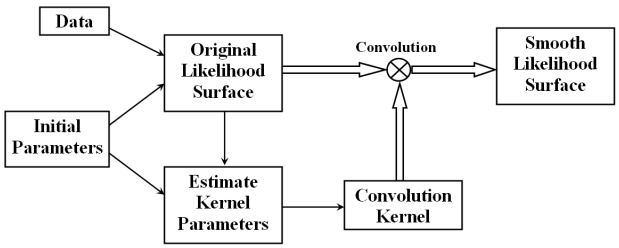


Figure 1. Block Diagram of our approach.

The main disadvantage of choosing a fixed kernel is that some of the components might not be smoothened while others might be over smoothened. Since, the Gaussian kernel has the property that the convolution sums up the parameters, this can also be treated as *Additive smoothing*. To avoid the problems of fixed kernel smoothing, we introduce the concept of *variable kernel smoothing* in this paper. In other words, σ_0 must be chosen individually for different components and it must be a function of σ_i for the i^{th} component. Since, the kernel parameters are effectively multiplied, this smoothing can be considered as *Multiplicative Smoothing*.

Proposition 1 *(Parameter Smoothing): Convolution of a Gaussian function with respect to its parameters is equivalent to convolving it with a Gaussian density.*

Convolution of Gaussian density with respect to the mean is shown below :

$$c(x,\breve{\theta_1}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-(\mu_1-\tau)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{\tau^2}{2\sigma_0^2}} d\tau$$

Replacing τ with $-\tau$ and rearranging, we get

$$\begin{aligned} c(x,\breve{\theta_1}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x+\tau-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{\tau^2}{2\sigma_0^2}} d\tau \\ &= c(\breve{x},\theta_1) \end{aligned}$$

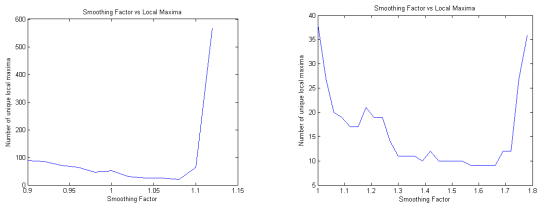
where $c(\breve{x},\theta_1)$ indicates smoothing with respect to density and $c(x,\breve{\theta_1})$ indicates smoothing with respect to parameter. Hence, we can see that convolution of a Gaussian density function with a Gaussian density with zero mean is equivalent to convolving the function with respect to its parameters.

4 Algorithm and its implementation

The basic advantage of the smoothing approach is that a simplified global method can be used to explore fewer promising local maxima on the smoothened surface. These solutions are used as initial guesses for the EM algorithm which is again applied to the next level of smoothing. Smoothing will help to avoid search in unwanted non-promising areas of the parameter space. The likelihood surface (defined by L) depends on the parameters and the available data. The smoothing factor ($sfac$) determines the extent to which the likelihood surface needs to be smoothened. If the smoothing factor exceeds certain threshold, the number of local maxima will increase tremendously. ns denotes the number of solutions that will be traced. nl determines number of levels in the smoothing hierarchy. There is a trade-off between the number of levels and the accuracy of the proposed method. Having many levels might increase the accuracy of the set of solutions, but is computationally expensive. On the other hand, having fewer number of levels is computationally cheaper, but one might have to forgo the accuracy of the solution. Choosing these parameters is not only user-specific but also depends significantly on the data that is being modeled. Algorithm 1 describes the smoothing approach. Initially, a simple global method is used to identify promising solutions (ns) on the smooth likelihood surface which are stored in Sol. With these solutions as initial estimates, EM algorithm is applied on the likelihood surface corresponding to the next level smooth surface. Smooth function returns the likelihood surface corresponding to smoothing factor at each level. The EM algorithm also returns ns number of solutions corresponding to the initial estimates. At every iteration, new likelihood surface is constructed with a reduced smoothing factor. This process is repeated until the amount of smoothing becomes zero which corresponds to the original likelihood surface. The main difference between the standard multi-level methods and our smoothing approach is that the dimensionality of the parameter space is not

Algorithm 1 Smooth

Input: Parameters Θ, Data $\mathcal{X}$, Tolerance τ, Smooth factor $Sfac$, number of levels nl, number of solutions ns
Output: $\widehat{\Theta}_{MLE}$
Algorithm:
L=Smooth($\mathcal{X}, \Theta$,nl*$Sfac$)
Sol=Global($\mathcal{X}, \Theta$, L,ns)
while $nl \geq 0$ **do**
  nl=nl-1
  L=Smooth($\mathcal{X}, \Theta$,nl*$Sfac$)
  for i=1:ns **do**
    Sol(i)=EM(Sol(i),$\mathcal{X}$,L,τ)
  end for
end while
$\widehat{\Theta}_{MLE}$ =max{Sol}



(a) Elliptical Dataset (b) Iris Dataset

Figure 2. Reduction in the number of local maxima.

changed during the smoothing (or coarsening) process.

5 Results and Discussion

Our algorithm is tested on four different datasets. Due to the space limitations, the readers are referred to [2] for more details about these datasets. Three different kinds of Gaussian components (spherical, elliptical and full covariance) with varying complexity (depending on the covariance matrix) are chosen. One real-world (iris) dataset is also used. One of the main advantages of the proposed smoothing algorithm is to ensure that the number of local maxima on the likelihood surface has been reduced. To the best of our knowledge, there is no theoretical way of estimating the amount of reduction in the number of unique local maxima on the likelihood surface. Hence, we use empirical simulations to justify the fact that the procedure indeed reduces the number of local maxima. Fig. 2 shows the reduction in the number of local maxima with respect to the smoothing factor for different datasets. Experiments were conducted using 100 random starts and the number of unique local maximum are reported. The same set of initial parameters were used for the smoothened surfaces. There is a gradual reduction in the number of local maxima as the smooth factor is increased. One can see that if the smoothing factor is increased beyond a certain threshold value, the number of local maxima increase rapidly. This might be due to the fact that over-smoothing the surface will make the surface flat, thus making it difficult for the EM to converge.

Smoothing the likelihood surface also helps in the optimization procedure. Table 2 summarizes the results obtained directly with the original likelihood and the smoothened likelihood. Mean and standard deviations across 100 random starts are reported. We have used only two level and tracked three solutions for each level.

Table 2. Performance results for our algorithm.

Dataset	RS+EM	Smooth+EM
Spherical	36.3 ± 2.33	41.22 ± 0.79
Elliptical	-3219 ± 0.7	-3106 ± 12
Full covariance	-2391.3 ± 35.3	-2164.3 ± 18.56
Iris	-196.34 ± 15.43	-183.51 ± 2.12

The average across all the starts is reported (RS+EM). The surface is then smoothened and the some promising solutions are used to trace the local optimal solutions and the average across all these starts are reported (Smooth+EM). For the smoothened version, only two levels were used. In other words, the optimal smoothing parameter is chosen and the EM algorithm is applied on the smoothened likelihood surface which were later used as initial guesses for the EM algorithm on the original likelihood surface.

6 Conclusion

This paper introduces a smoothing approach for learning mixture models from multivariate data. Our algorithm is based on the conventional EM algorithm applied to a smoothened likelihood surface. A hierarchy of smooth surfaces is constructed and optimal set of parameters is obtained by applying a discrete version of continuation method to the promising solutions of the smooth surface. This is an effective optimization-based smoothing procedure reduces the number of local maxima and thus it eliminates extensive search in some non-promising regions of the parameter space. Empirical results on standard datasets demonstrate a significant improvement of the proposed algorithm compared to other existing methods.

Acknowledgments

This work is partially funded by the Wayne State University faculty research award. We would like to thank Dr. Hsiao-Dong Chiang for some discussions.

References

[1] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley and Sons, New York, 1997.

[2] C. K. Reddy, H. D. Chiang, and B. Rajaratnam. TRUST-TECH based expectation maximization for learning finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1146–1157, 2008.

[3] N. Ueda, R. Nakano, Z. Ghahramani, and G. Hinton. SMEM algorithm for mixture models. *Neural Computation*, 12(9):2109–2128, 2000.