Probability Density Function Estimation by different Methods

Vikas Chandrakant Raykar

Abstract— **The aim of the assignment was to estimate the probability density function (PDF) of any arbitrary distribution from a set of training samples. PDF estimation was done using parametric (Maximum Likelihood estimation of a Gaussian model), non-parametric (Histogram, Kernel based and K-nearest neighbor) and semi-parametric methods (EM algorithm and gradient based optimization). Application of EM algorithm for binary sequence estimation has also been discussed.**

I. INTRODUCTION

A Bayesian approach towards pattern classification consists of feature extraction and classification. Feature extraction involves the extraction of a lower dimensional feature vector from the pattern. Once the feature vector is extracted the pattern can be classified based on Bayes decision rule. Consider a C class problem. Let x be a feature vector extracted from the given input pattern. The decision rule can be stated as

$$\textit{Decide } C_k \textit{ if } p(C_k/x) > p(C_j/x)\ \forall j \neq k \qquad (1)$$

The posterior probability can be calculated using the Bayes theorem as follows

$$p(C_k/x) = p(x/C_k)p(C_k)/p(x) \qquad (2)$$

.So the important part is the evaluation of the class conditional density $p(x/C_k)$ for all the C classes. This is the training phase where we have a set of N feature vectors also called training samples $c = \{x_1, x_2 ....x_N\}$ belonging to class C_k and we estimate $p(x/C_k)$ given the N training samples. This has to be done for all the classes. To ease notation $p(x/C_k)$ is referred as $p(x)$. Rest of the discussion will be with respect to one class only.

The different methods for PDF estimation can be classified as Parametric, Non-Parametric and Semi parametric. In parametric method the PDF is assumed to be of a standard form (generally Gaussian, Raleigh or uniform). The parameters of the assumed PDF can be estimated either using ML estimation or Bayesian Estimation. The non parametric methods include histogram based, the kernel based methods and the K nearest neighbor methods. In semi parametric methods the given density can be modeled as a combination of known densities. The parameters can be estimated either using gradient descent or Expectation Maximization (EM) algorithm.

Section II discusses the example used to compare the various PDF estimation techniques and also the performance measure used. Section III, IV, V discusses the parametric, non parametric and semi parametric techniques respectively. Section VI concludes. Section VII discusses the application of EM algorithm for binary sequence estimation.

II. PROGRAM DETAILS

A 2 dimensional feature vector was used in the program. Figure 1 show the original density function used. The brightness of the pixel corresponds to the density value at that point. In our case the density function is uniform in the white region. We would like to estimate the density from a set of N training samples drawn from it. The training samples were drawn from a uniform distribution over the entire range of the image and the sample was retained if it belonged to the white region or else discarded. In this way N random training samples were drawn

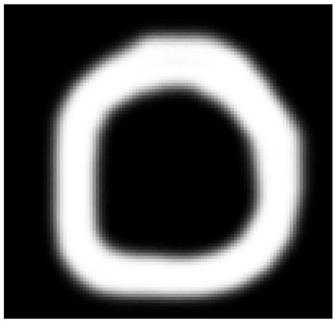

Figure 1 Plot of original pdf's used

A GUI was written in MATLAB 6.1 to estimate the PDF from these N samples using different methods. Figure 2 shows a snapshot of the GUI. Once the PDF was estimated the method was evaluated using the Kullback-Leibler distance. The performance was evaluated as follows. First we draw M samples from the image called as x_{test}. The PDF is evaluated at

This report was written for ENEE 739Q SPRING 2002 as a part of the course project. The author is a graduate student at Department of Electrical Engineering, University of Maryland, and College Park MD 40742 USA.

each of the M points $p_{eval}(x)$. Let $p(x)$ be the original PDF then the Kullback-Leibler distance. D is defined as

$$D = \sum_{x_{test}} p(x)\ \ln(\,p(x)/\,p_{eval}(x)) \qquad (3)$$

Although D does not satisfy the triangle inequality and is therefore not a true metric, it satisfies many important mathematical properties. For example, it is a convex function of $p_{eval}(x)$, is always nonnegative, and equals zero only if $p_{eval}(x) = p(x)$.For iterative algorithms, D was plotted as a function of the iteration number. Whenever $p_{eval}(x)$ was zero D was evaluated by setting $p_{eval}(x)$ to a very small value. Also it does not make sense to use this measure to compare different methods as we are choosing the test points only where the original PDF is not zero. Like in ML estimation we may get a good estimate in the region where the original PDF is not zero however where the PDF goes zero the estimate is very bad (though we are not considering regions where the original PDF is not zero). However this measure will be useful to study the effect of changing the parameters of a given method.

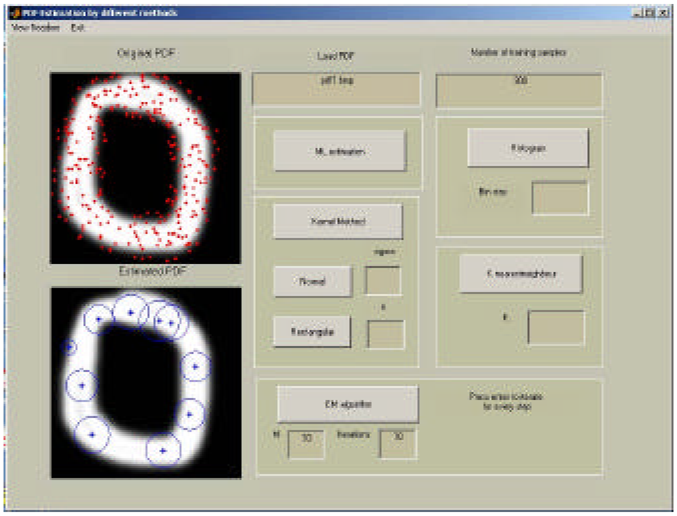

Figure 2 A snapshot of the GUI

III. PARAMETRIC ESTIMATION

In parametric estimation the the PDF is assumed to have a known distribution. In our case a standard bivariate Gaussian was used. The standard multivariate Gaussian has the following form

$$p(x) = \frac{1}{(2\,pi)^{d/2}\,|\Sigma|^{1/2}}\, e^{-1/2(x-m)^T\,\Sigma^{-1}\,(x-m)} \qquad (4)$$

For the bivariate case d=2, x is 2D vector µ is the mean vector and Σ is the 2x2 covariance matrix. The parameters µ and Σ can be estimated either using Bayesian estimation or Maximum Likelihood (ML) estimation. Using the N training samples $c = \{x_1, x_2 ....x_N\}$ randomly drawn the mean and the covariance matrix are given by ML estimation as

$$\hat{m} = \frac{1}{N}\sum_{j=1}^{N} x_j$$
$$\hat{\Sigma} = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \hat{m})^T\,(x_j - \hat{m}) \qquad (5)$$

Where $\hat{m}$ and $\hat{\Sigma}$ are the estimated mean vector and covariance matrix respectively. $\hat{m}$ is an unbiased consistent estimate of the mean vector. $\hat{\Sigma}$ is divided by N-1 and not N in order to make the covariance matrix unbiased estimate. Also the covariance matrix estimate is consistent.

Figure 3 shows the plot of the original and the estimated PDF for N=500. It can be seen that the PDF is not the same as the original PDF except in the mean and covariance sense. This is because our basic assumption of modeling the distribution as a single bivariate Gaussian is not sufficient.

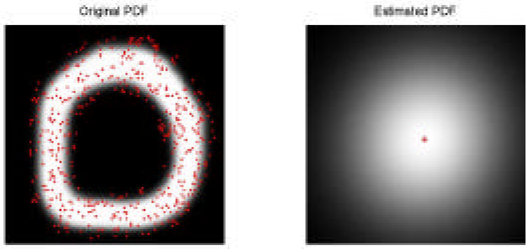

Figure3 Original and Estimated PDF using ML estimation

Figure 4 shows the Kullback-Leibler distance as a function of N for 500 test points(i.e. M=500). So increasing N beyond 300 does not help much as our model is essentially flawed.

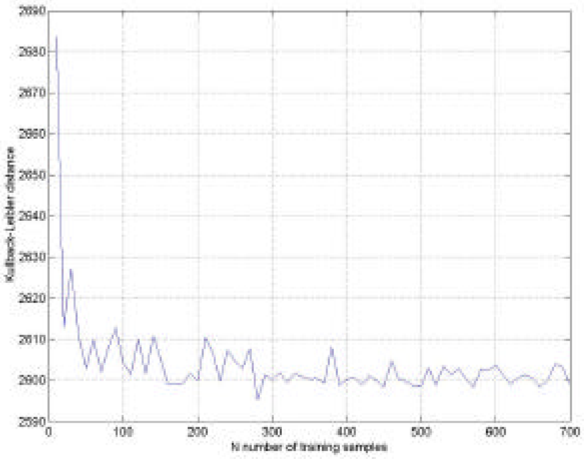

Figure 4 Plot of D Vs N for ML estimation for M=500

IV. NON PARAMETRIC METHODS

1.Histogram

In this approach the entire image is divided into a small number of bins and using the training samples $c = \{x_1, x_2 ....x_N\}$ the PDF is calculated as a histogram. This is a very direct and simple approach and once the PDF is estimated the training data can be discarded. The disadvantage is that we may lose some information and also it is computationally expensive in higher dimensions. The bin width M has to be chosen optimally. If M is too large we get a spiky PDF or if M is too small there will be significant loss in structure. Figure 5 shows D as a function of bin size for different N. Using this to decide the bin size does not make sense as we are evaluating only where the original PDF is not zero. For our case as we increase the bin size since our original distribution is uniform as bin size increases D decreases which may not be the case for any general PDF. The only thing we can conclude is that as N increases we get better estimates. Figure 6 shows the estimated PDF for M=4, N=1000.

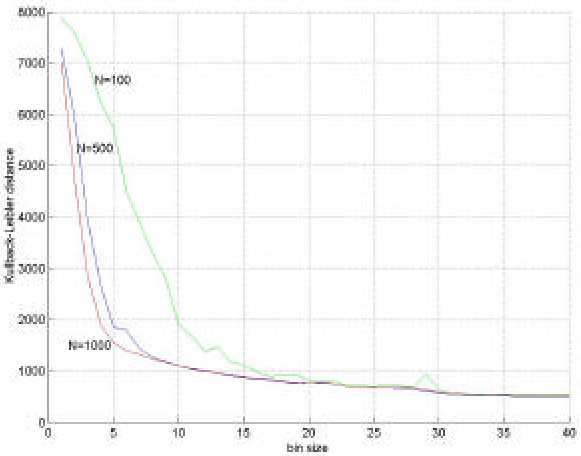

Figure 5 bin size vs. Kullback-Leibler distance for different N for a Histogram based PDF estimator (500 test points)

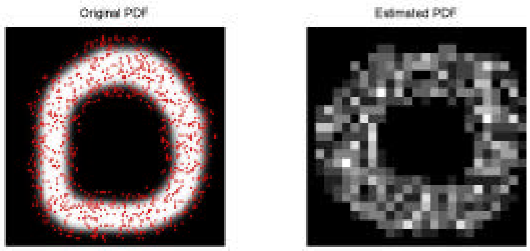

Figure6. Histogram estimation for bin size 10 and N=1000

2.Principled approach

A more principled version of the histogram can be formulated .Given N training samples $c = \{x_1, x_2 ....x_N\}$ let K samples lie inside a region R of volume V. Then the PDF at any point inside the region R is given by $p(x) = K / NV$. Kernel based methods fix V and find K. K nearest neighbor method fixes K and finds V. The advantage is that these methods do not have the 'Curse of dimensionality'. However we need to keep all data to evaluate the PDF.

2.1 Kernel based methods

In this method we fix the volume of the region R as V and vary K the estimated PDF at any point x is given by

$$\hat{p}(x) = \frac{\sum_{n=1}^{N} H(\frac{x - x_n}{h})}{Nh^d} \qquad (6)$$

where H(x) is the kernel function. In this case H(x) is a hypercube of length h centered at x_n defined as

$H(\frac{x - x_n}{h}) =$ 1 if x falls inside the hypercube centered at x_n and height h, 0 otherwise. The hypercube is basically a discontinuous kernel. Instead of the hypercube we can chose a Gaussian kernel. The variance s of the Gaussian kernel and h the height of the hypercube are the smoothing parameters. The smoothing parameters are to be optimally chosen. If the smoothing parameter is too low then the PDF is very patchy and N has to be very large to get a good estimate of the PDF. If the smoothing parameter is very large then the PDF spreads out.

Figure 7 shows the Kullback-Leibler distance for square kernel as a function of h for different N for M=500 test points. As can be seen from the plot initially D decreases up till a certain point reaches a minimum and then again increases. The part where D decreases (i.e. better estimate) is when the squares have enough width to overlap and give a better estimate. From the plot it can be seen that for N=600 the optimal value of is around 4 to 6. Figure 8 shows the estimated PDF and the original PDF for N=600 h=6 for a rectangular kernel.

Figure 9 shows the Kullback-Leibler distance for the Gaussian kernel as a function of the variance s for different N for 800 test points. As can be seen from the plot that D decreases as the smoothing parameter increases till a certain point and after that it increases again. From the plot it can be seen that the optimal value for s is 2. Also as N increases the curves shift downwards which is straightforward that as the number of training samples increases we get a better estimate of the PDF. . Figure 10 shows the estimated PDF for N=500 and s =2 for the Gaussian kernel.

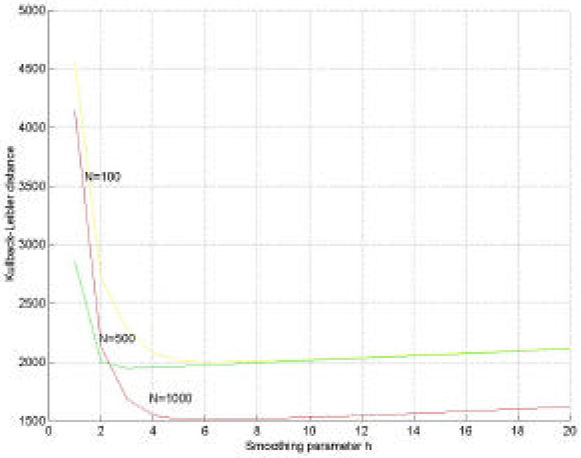

Figure 7 Plot of height vs. Kullback-Leibler distance for different N for a rectangular kernel based PDF estimator based on 500 test points

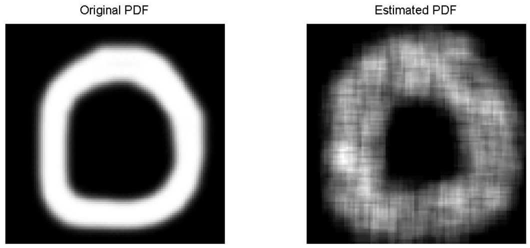

Figure 8 Estimated PDF using rectangular kernel based method for N=600 and h=6

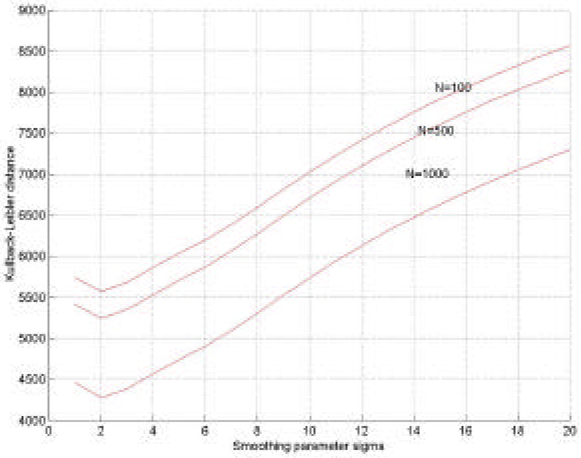

Figure 9 Plot of sigma vs. Kullback-Leibler distance for different N for a Gaussian kernel based PDF estimator based on 800 test points

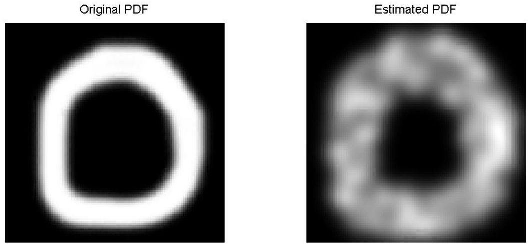

Figure 10 Estimated PDF using Gaussian kernel based method for N=500 and sigma=2

2.2 K Nearest Neighbor

In this method V is fixed and K is varied. Essentially we need to search the K nearest neighbors. In this case K has to be optimally chosen for a good estimate of the PDF. Figure 11 shows the Kullback-Leibler distance as a function of K for different N and M=200 test points. Figure 12 shows the estimated PDF for N=300 and K=12.

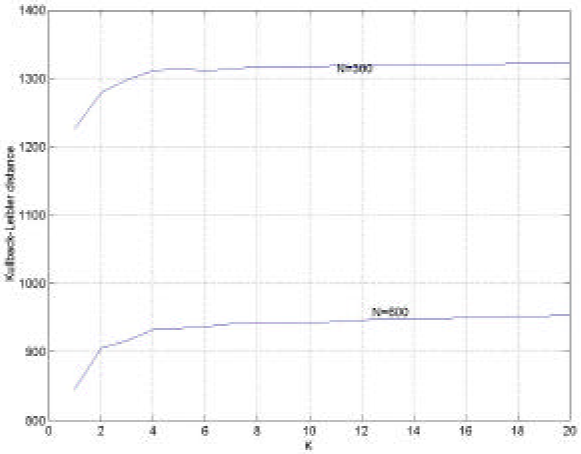

Figure 11 Plot of K vs. Kullback-Leibler distance for different N for a KNN based PDF estimator based on 500 test points

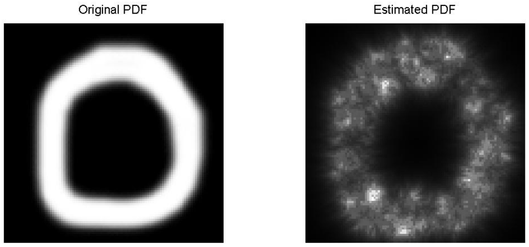

Figure 12 Estimated PDF using KNN based method for N=300 and K=12

V. Semi Parametric Methods

These methods combine the flexibility of nonparametric methods and the efficiency in evaluation of parametric methods. Here we model the PDF as a mixture of parametric PDF. The parameters have to be estimated either by some optimization technique like gradient descent or Expectation Maximization Algorithm.

1.EM Algorithm

The EM algorithm convergence properties are studied as a function of the number of iterations Figure 13 shows the Kullback-Leibler distance for 500 training samples and 500 test points for different M(number of mixture components) as a function of the iteration number. It can be seen that the EM algorithm converges in three to six iterations. Also D decreases as M increases.

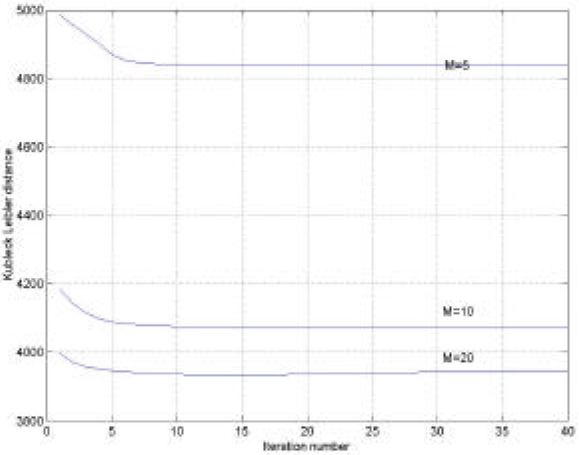

Figure 13 Kullback-Leibler distance for N=500 and 500 test points for different M (number of mixture components) as a function of the iteration number.

Figure 14 shows the Kullback-Leibler distance for M=10(component densities) and 500 test points for different N as a function of the iteration number. Increasing N has no effect on the speed of convergence however as N increases the Kullback-Leibler distance decreases.

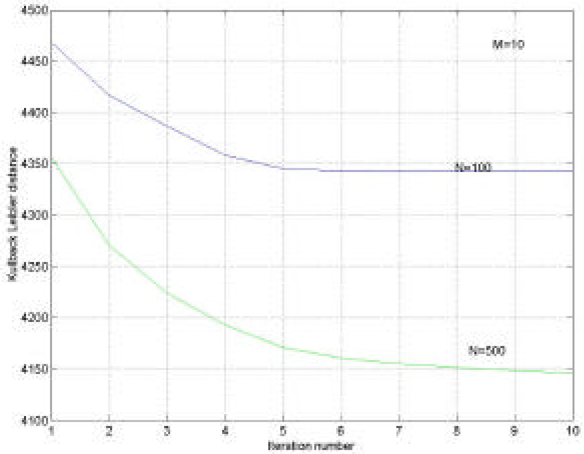

Figure 14 Kullback-Leibler distance for M=10 and 500 test points for different N as a function of the iteration number

Figure 15 shows the log likelihood for N=500 and 500 test points as a function of the iteration number for different M. The log likelihood increases from the point where it starts to converge.

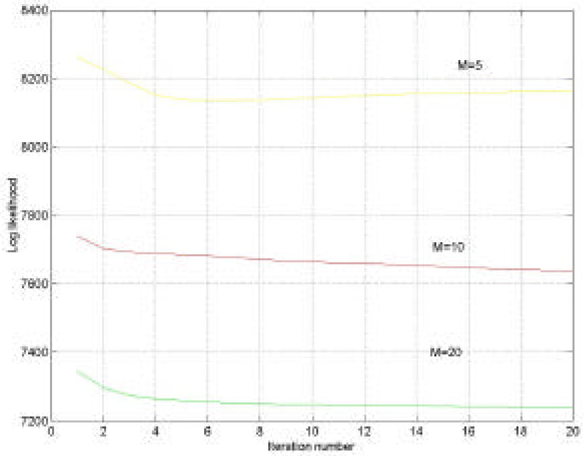

Figure 15 log likelihood for N=500 and 500 test points as a function of the iteration number for different M.

The EM algorithm also depends on the initialization strategies which in turn affect the number of iterations required to converge. If the initial points are within the uniform region of the PDF them the EM algorithm will converge very fast. Mostly it was observed that the EM algorithm invariant to initialization strategies converged in 5 to 10 steps. Figure 16 shows the initial and the final position of the Gaussians.

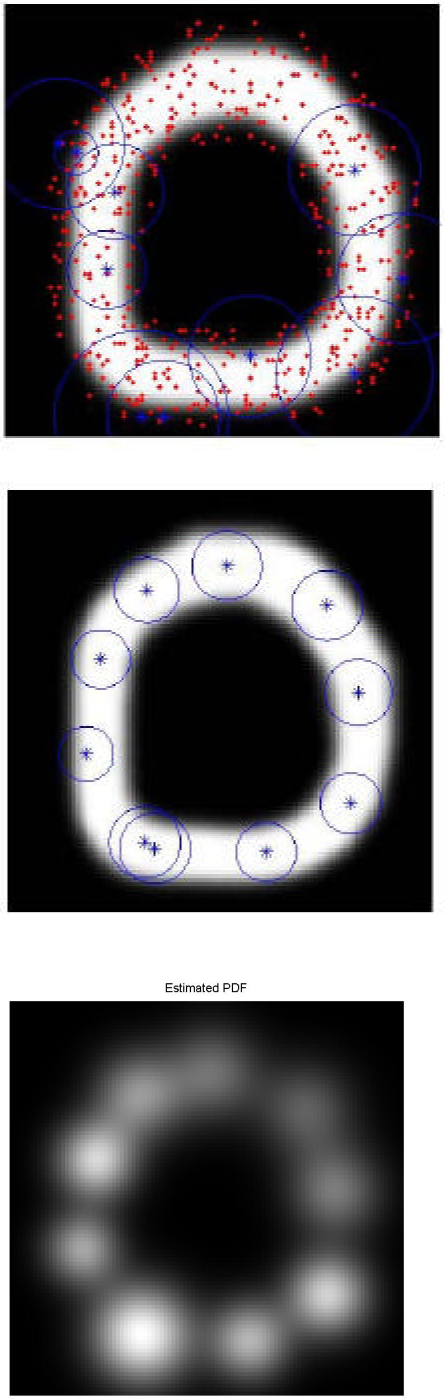

Figure 16 initial Gaussians and the final positions of the Gaussians after 5 iterations and the Estimated PDF for M=10 components N=500

2. Gradient Descent Optimization

The negative log likelihood function can be minimized by gradient descent method. The minimization is with respect to the parameters the mean, variance of the initial Gaussians and the mixing parameters. The descent rate for each of the parameters is got after trial and error approach. In this case alpha for mean was used as 0.9 for sigma 0.3 and 0.1 for the mixing parameters.

Figure 17 shows Kullback-Leibler distance for N=500 and 500 test points for different M (number of mixture components) as a function of the iteration number. As can be seen from the plot the PDF converges after around 25 iterations. The convergence is very slow as compared to the EM algorithm. Also convergence is very sensitive to the descent rate. The descent rate for mean, variance and mixing parameters were chosen by trail and error. By properly choosing the descent rate I guess we can get a faster convergence.

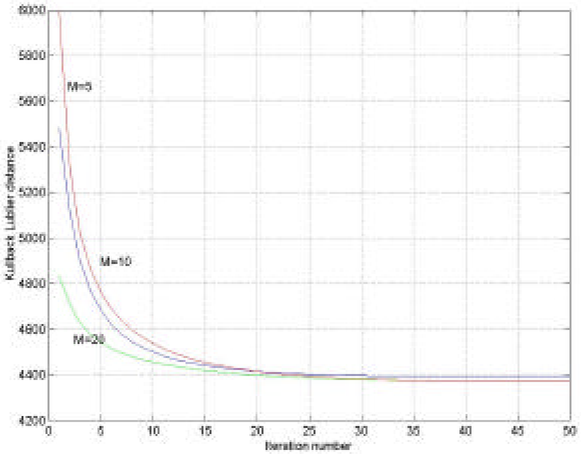

Figure 17 Kullback-Leibler distance for N=500 and 500 test points for different M (number of mixture components) as a function of the iteration number

Figure 18 shows the Kullback-Leibler distance for M=10 and 500 test points for different N as a function of the iteration number

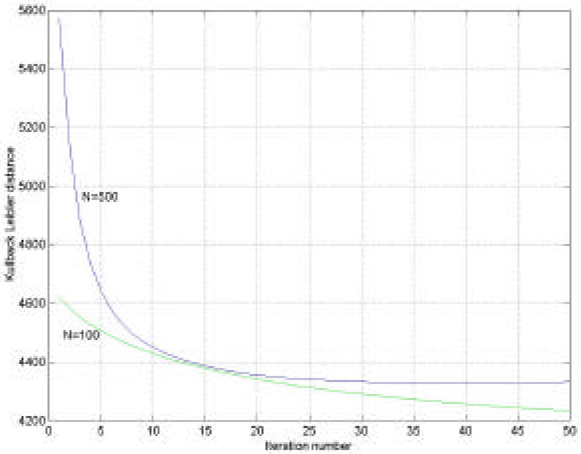

Figure 18 Kullback-Leibler distance for M=10 and 500 test points for different N as a function of the iteration number.

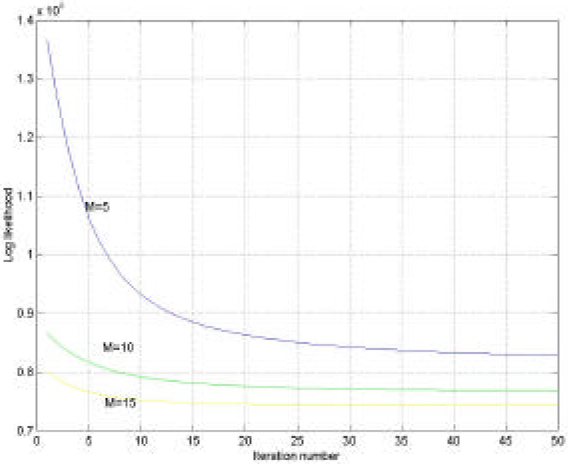

Figure 19 shows the log likelihood for N=500 and 500 test points as a function of the iteration number for different M. The log likelihood increases from the point where it starts to converge.

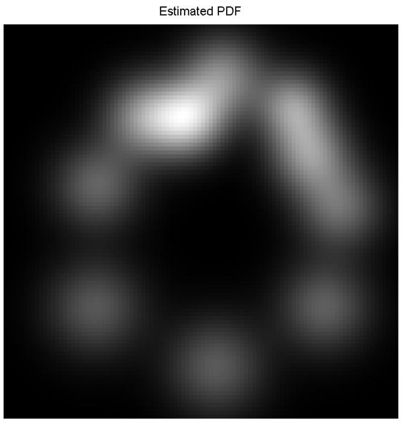

Figure 20 Estimated PDF for M=10 components N=500

Note the EM algorithm gives a better PDF than gradient descent for the same number of components.

VI. CONCLUSION

Use kernel based method for minimizing the computational requirements (but we have to keep the data) and use EM algorithm. For minimizing both memory and computational requirements.

VII. EXTRA CREDIT II

The following section discusses the application of the EM algorithm for binary sequence estimation. Consider a system shown in the Figure 21. B is a binary sequence of length N. B=[b_1,b_2,…..b_N] where each b_i could be a one or zero. A typical realization for N=5 could be [1 0 0 1 0 1]. The binary sequence is scaled by a fixed unknown non zero scalar c. it is then corrupted by additive white Gaussian noise.

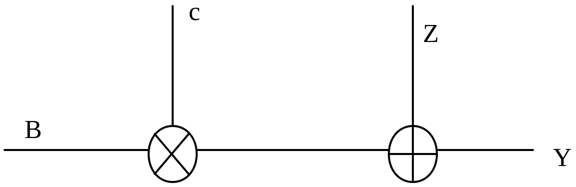

Figure 21. A simple channel with scaling and noise added.

So we have

$$Y = cB + Z \ \ where\ Y = \begin{bmatrix} y_1 \\ . \\ y_N \end{bmatrix} B = \begin{bmatrix} b_1 \\ . \\ b_N \end{bmatrix} Z = \begin{bmatrix} z_1 \\ . \\ z_N \end{bmatrix}$$

Each of the z_i are i.i.d. Gaussian random with zero mean and variance s i.e. $z_i \sim N\left(z_i;0,s^2\right)$. The problem is to get an ML estimate of B. Note that c is unknown. The ML estimation problem can be formulated as follows.

$$p_{y_i/b}(y_i/b_i) = \begin{cases} N\left(y_i;0,s^2\right) when\ b_i = 0 \\ N\left(y_i;c,s^2\right) when\ b_i = 1 \end{cases}$$

So the ML estimator is

$$if\ \ p_{y_i/b}(y_i/b_i=1) \geq p_{y_i/b}(y_i/b_i=0)\ b_i = 1 else\ b_i = 0$$

Simplifying we get the following.

$if\ \ y \geq c/2 \quad b_i = 1\ else\ b_i = 0$ Here the value of c is unknown even though if it is a fixed quantity.

We can use the EM algorithm by defining a new completes data X=(Y,C). The E step gives estimates of C which can be used in the M step.

*E STEP:*Let D be some estimate of B

$$Q(B/D) = E\left[\log\ p_{Y,C/B}(y,c/B)/Y,D\right]$$

$$p_{Y,C/B}(y,c/B) = p_{Y/C,B}(y/c,B)$$

$$c\ is\ not\ a\ random\ \ \mathrm{var}\ iable$$

$$p_{Y/C,B}(y/c,B) = \prod_{k=1}^{N} N(y_k;cb_k,s^2).$$

$$Simplifyin\ \ g\ ,$$

$$Q(B/D) = E[-\sum_{j=1}^{N}(y_j - cb_j)^2/Y,D]$$

$$Q(B/D) = -\sum_{j=1}^{N}(y_j - E[c/Y,D]b_j)^2$$

$D_{j=0}$ provides no information about c. Let $D^{|}$ the subset of D which are 1 and let $Y^{|}$ be the corresponding Y. So the E step can be summarized as follows

$$let\ \ E[C/Y,D]=E[Y^{|}]=a$$

$Y^{|}$ is the subset of Y values asssciated with the current estimates of B which are 1's.

$$Q(B/D)=-\sum_{j=1}^{N}(y_j-ab_j)^2$$

M STEP: Find B to maximize this . We can maximize each individual term. Consider a typical term we have

$$-(y_j-a)^2\geq -y^2\ \ if\ \ b_j^{new}=1$$

So the M step is,

For each b_j^{old}, y_j^{old} belonging to B and Y ,

$$if\ \ (\ a>0\ \ and\ \ y_j>a/2)\ \ or\ \ (\ a<0\ \ and\ \ y_j<a/2)$$

$$set\ \ b_j^{new}=1\ otherwise\ set\ b_j^{new}=0$$

ALGORITHM:

```
1.Initialize $B^{old}$ = [ 1 1 ………..1]
2.E step:  a=E[$Y^{|}$ ] where $Y^{|}$ is the subset of Y associated with the current estimates of B which are 1's.
3.M step:  For each $b_j, y_j$ belonging to $B^{old}$ and Y ,
    $if\ \ (\ a>0\ \ and\ \ y_j>a/2)\ \ or\ \ (\ a<0\ \ and\ \ y_j<a/2)$
    $set\ \ b_j^{new}=1\ otherwise\ set\ b_j^{new}=0$
4.Iterate till convergence.
```

SIMULATION: Simulation was done for the case for N=1000. c=3. The algorithm converged in about 2 to 3 iterations. Convergence was decided when there was no further improvement in the value of estimated c. Figure 22 shows the error in the estimation of B as a function of iteration number for different s. It can be seen that it converges in about 2 to 3 iterations. Figure 23. shows the estimated value of c as a function of iteration number for s=1.

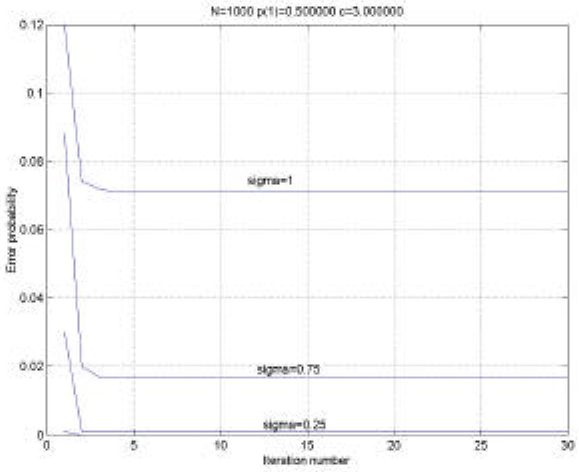

Figure 22. Error in the estimation of B as a function of iteration number for different s.

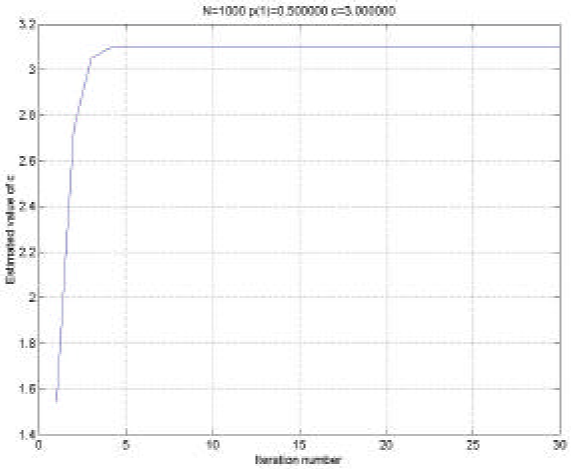

Figure 23. Estimated value of c for sigma =1 as a function of iteration number

REFERENCES

[1] Ghahramani, Z & **Jordan**, **MI** (1994). *Supervised learning from incomplete data via an EM approach*. In JD Cowan, G Tesauro and J Alspector, editors, Advances in Neural Information Processing Systems 6. San Mateo, CA: Morgan Kaufmann, 120-127. (http://citeseer.nj.nec.com/ghahramani94supervised.html)

[2] **Bilmes**, J (1998) *A Gentle Tutorial of the EM Algorithm*, UC-Berkeley TR-97-021. (http://citeseer.nj.nec.com/bilmes98gentle.html .

[3] C.N. Georghiades and J.C. Han, *Sequence Estimation in the Presence of Random Parameters Via the EM Algorithm*, IEEE Transactions on Communications, vol. 45, pp. 300-308, March 1997

Vikas Chandrakant Raykar is a graduate student at the University of Maryland College Park, MD 40742 USA (telephone: 301-405-1208, e-mail: vikas@umiacs.umd.edu).