A NEW APPROACH TO CONSTRAINED EXPECTATION-MAXIMIZATION FOR DENSITY ESTIMATION

Hunsop Hong and Dan Schonfeld

University of Illinois at Chicago
Department of Electrical and Computer Engineering
Chicago, IL 60607

ABSTRACT

In this paper, we present two density estimation methods based on constrained expectation-maximization (EM) algorithm. We propose a penalty-based maximum-entropy expectation-maximization (MEEM) algorithm to obtain a smooth estimate of the density function. We further propose an attraction-repulsion expectation-maximization (AREM) algorithm for density estimation in order to determine equilibrium between over-smoothing and over-fitting of the estimated density function. Computer simulation results are used to show the effectiveness of the proposed constrained expectation-maximization algorithms in image reconstruction and sensor field estimation from randomly scattered samples.

Index Terms— Gaussian mixture model (GMM), maximum entropy penalty, Gibbs density function, expectation-maximization (EM), image reconstruction, sensor field estimation.

1. INTRODUCTION

Estimation of the probability density function (pdf) from samples has been the topic of an intense research effort for several decades. The Parzen window method [1] is one of the most powerful techniques for density estimation. It relies on the use of narrow kernels (usually low-variance Gaussian functions) at each sample. This method has been shown to converge to the true density function as the number of samples increases. However, the computational burden of this approach also increases rapidly as the number of samples rises. Therefore, much research has been devoted to reduce the computational complexity by approximating the result of the Parzen window method. One of the popular methods used is obtained by minimizing the integrated squared-error (ISE) between the Parzen window method and an approximation represented by a linear combination of a much smaller number of kernel functions with arbitrary variance. The limitation of ISE-based solutions is that they suffer from a degeneracy problem [2] and do not fully utilize the sample information as the number of samples increases. Moreover, ISE-based methods are generally used to determine optimal weights used in the linear combination. Selection of the mean and variance of the kernel functions is accomplished by using the K-means algorithm, which can be viewed as a hard limiting case of the Expectation-Maximization algorithm (EM) [3]. The EM algorithm offers a very effective iterative method for estimating the model parameters including the weight, mean and covariance matrix of the kernel functions. In this paper, we propose a maximum-entropy expectation maximization (MEEM) algorithm. The MEEM algorithm provides a smooth estimate of the density function by using a maximum entropy penalty as a constraint and thus avoiding the well-known over-fitting problem. We therefore propose a different approach to density estimation by introducing the attraction-repulsion expectation-maximization (AREM) algorithm which aims to achieve a balance between over-fitting and over-smoothing. We use the Gibbs and inverse-Gibbs density functions to model the attraction and repulsion penalties, respectively.

The paper is organized as follows: The MEEM and AREM algorithms are introduced in Sections 2 and 3, respectively. Experimental results are presented in Section 4. In Section 5, we provide a brief summary and conclusions.

2. MAXIMUM ENTROPY EXPECTATION MAXIMIZATION ALGORITHM

A probability density function can be expressed by K Gaussian mixture,

$$\hat{f}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k G\left(\mathbf{x} - \mathbf{m}_k, \mathbf{C}_k\right), \tag{1}$$

where $\mathbf{m}_k$ is center of a Gaussian function, $\mathbf{C}_k$ is a covariance matrix of k^{th} function and α_k is the weight for each center. The conditions for weights are $\sum_{i=1}^{K} \alpha_i = 1, \alpha_i > 0$ to maintain the property of pdf. The Gaussian function is given as

$$G\left(\mathbf{x} - \mathbf{m}_k, \mathbf{C}_k\right) = \frac{\exp\left\{-\frac{(\mathbf{x}-\mathbf{m}_k)^T \mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)}{2}\right\}}{(2\pi)^{D/2}|\mathbf{C}_k|^{1/2}}. \tag{2}$$

The logarithm of the likelihood function for the given Gaussian mixture parameters that has N observations can be written as,

$$L_L\left(\theta\right) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \alpha_k G\left(\mathbf{x}_n - \mathbf{m}_k, \mathbf{C}_k\right) \tag{3}$$

using (1) and (2) where $\mathbf{x}_n$ is n^{th} sample and θ is a set of parameters (the weights, centers and covariances) to be estimated.

The entropy term is added in order to avoid degeneracy problem. Here we use Renyi's quadratic entropy measure [4],

$$H(\theta) = -\log \sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_i \alpha_j G\left(\mathbf{m}_i - \mathbf{m}_j, \mathbf{C}_i + \mathbf{C}_j\right) \tag{4}$$

We therefore form an augmented likelihood function L_{ME} parameterized by a positive scalar ν for simultaneously maximizing the entropy and the likelihood using (3) and (4),

$$L_{ME}\left(\theta, \nu\right) = L_L\left(\theta\right) + \nu H\left(\theta\right). \tag{5}$$

We note that the EM algorithm here we use is a simple expansion of the lower bound maximization method appears in [5], which converges to the maximum point of the augmented likelihood in (5).

The expectation step of the EM algorithm can be separated into two terms, one is the expectation related with likelihood and the other is the expectation related with the entropy penalty,

$$p_L^t(k,n) = \frac{\alpha_k G(\mathbf{x}_n - \mathbf{m}_k, \mathbf{C}_k)}{\sum_{l=1}^K \alpha_l G(\mathbf{x}_n - \mathbf{m}_l, \mathbf{C}_l)} \tag{6}$$

$$p_E^t(k,l) = \frac{\alpha_k \alpha_l G(\mathbf{m}_k - \mathbf{m}_l, \mathbf{C}_k + \mathbf{C}_l)}{\sum_{m=1}^K \sum_{n=1}^K \alpha_m \alpha_n G(\mathbf{m}_m - \mathbf{m}_n, \mathbf{C_m} + \mathbf{C_n})} \tag{7}$$

where L denotes the likelihood function, E denotes the entropy penalty and t denotes the number of iteration.

The Jensen's inequality is applied to find the new lower bound $\beta_{ME}^t(\theta,\nu)$ of the likelihood functions using eqs. (6) and (7). Therefore, The lower bound function $\beta_L^t(\theta)$ for the likelihood function $L_L(\theta)$ can be derived as

$$\beta_L^t(\theta) = \sum_{n=1}^N \sum_{k=1}^K p_L^t(k,n) \log \frac{\alpha_k G(\mathbf{x}_n - \mathbf{m}_k, \mathbf{C}_k)}{p_L^t(k,n)} .$$

The lower bound for the entropy term $\beta_E^t(\theta)$

$$\begin{aligned}&\beta_E^t(\theta)\\ &= -\sum_{k=1}^K \sum_{l=1}^K p_E^t(k,l) \log \left(\frac{\alpha_k \alpha_l G(\mathbf{m}_k - \mathbf{m}_l, \mathbf{C}_k + \mathbf{C}_l)}{p_E^i(k,l)} \right) .\end{aligned}$$

Therefore, the lower bound $\beta_{ME}^i(\theta,\nu)$ which combines two lower bounds is

$$\beta_{ME}^t(\theta,\nu) = \beta_L^t(\theta) + \nu \beta_E^t(\theta) . \tag{8}$$

Now we have the lower bound function, the new estimates of the parameters are easily calculated by setting the derivatives of $\beta^t(\theta,\nu)$ with respect to each parameters to zero. The update equation for mean vector is,

$$\begin{aligned}\mathbf{m}_k^{t+1} = &\left(\sum_{n=1}^N p_L^t(k,n) \mathbf{C}_k^{-1} \right.\\ &\left. - 2\nu \sum_{l=1, l\neq k}^K p_E^t(k,l) (\mathbf{C}_k + \mathbf{C}_l)^{-1} \right)^{-1}\\ &\left(\sum_{n=1}^N p_L^t(k,n) \mathbf{C}_k^{-1} \mathbf{x}_n \right.\\ &\left. - 2\nu \sum_{l=1, l\neq k}^K p_E^t(k,l) (\mathbf{C}_k + \mathbf{C}_l)^{-1} \mathbf{m}_l \right) .\end{aligned} \tag{9}$$

we use soft-max function[6] for weight in order to consider the weight constraint. Thus α_k^{t+1} is ,

$$\alpha_k^{t+1} = \frac{\sum_{n=1}^N p_L^t(k,n) - 2\nu \sum_{l=1}^K p_E^t(k,l)}{N - 2\nu}.$$

However, in covariance case we cannot solve directly because of the existence of inverse matrix appears in the derivative. By Cauchy-Schwartz inequality we can get,

$$\begin{aligned}&\{G(\mathbf{m}_l - \mathbf{m}_m; \mathbf{C}_l + \mathbf{C_m})\}^2\\ &\qquad \leq G(0, 2\mathbf{C}_l) G(0, 2\mathbf{C_m}) .\end{aligned} \tag{10}$$

We can derive new lower bound $\{\beta_E^t(\theta,\nu)\}_{\mathbf{C}}$ from $\beta_E^t(\theta)$ using (10). Therefore, the covariance $\mathbf{C}_k^{t+1}$ is

$$\mathbf{C}_k^{t+1} = \frac{\sum_{n=1}^N p_L^t(k,n)(\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T}{\sum_{n=1}^N p_L^t(k,n) - \nu \sum_{l=1}^K p_E^t(k,l)} .$$

The equations for new parameter estimation show that the proposed algorithm requires slight more computational burden over conventional EM algorithm because $K << N$ and moreover most of inverse matrix used in maximization steps is also required for the conventional expectation step in (6).

However, the MEEM algorithm only considers over-fitting problem. In next section, we present an EM based density estimation algorithm which considers over-fitting and over-smooth.

3. ATTRACTION REPULSION EXPECTATION MAXIMIZATION ALGORITHM

In choosing a proper penalty for quantizing both attraction and repulsion, the Gibbs distribution provides a useful representation. A Gibbs distribution can be shown

$$g(\mathbf{x}) = \frac{1}{Z} \exp\left\{ -\frac{E(\mathbf{x})}{T} \right\} \tag{11}$$

where Z is normalizing constant and T is temperature. Since Gaussian mixture model contains covariance matrix, we can use Mahalanobis distance [7] measure for the energy function. Thus, the energy function on some distance $\mathbf{x}$ with k^{th} covariance matrix $\mathbf{C}_k$ is

$$E_k(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{C}_k^{-1} \mathbf{x} . \tag{12}$$

Therefore, we plug (12) into (11) then we get Gibbs distribution $g_k^A(\mathbf{x})$ induced by k^{th} covariance matrix

$$g_k(\mathbf{x}) = \frac{1}{C_k^0 \sqrt{T}} \exp\left\{ -\frac{E_k(\mathbf{x})}{T} \right\}$$

where $C_k^0 = (2\pi)^{D/2} |\mathbf{C}_k|^{1/2}$.

However, We want to estimate the parameters by minimizing the effect of other centers with considering bidirectional way, so we introduce the inverse Gibbs density where the energy function is,

$$E_k^R(\mathbf{x}) = -\frac{1}{2} \mathbf{x}^T \mathbf{C}_k^{-1} \mathbf{x} . \tag{13}$$

The inverse Gibbs distribution function can be expressed as using (13) and (11),

$$g_k^R(\mathbf{x}) = \begin{cases} C_k^0 C_N \sqrt{T_R} \exp\left\{ -\frac{E_k^R(\mathbf{x})}{T_R} \right\}, & E_k^R(\mathbf{x}) \geq -D_{TH} \\ 0, & E_k^R(\mathbf{x}) < -D_{TH} \end{cases}$$

which truncated according to some threshold D_{TH} and the constant C_N is a normalizing constant

$$C_N = \left(\int_{E_k^R(\mathbf{x}) \geq -D_{TH}} C_k^0 \sqrt{T_R} \exp\left\{ -\frac{E_k^R(\mathbf{x})}{T_R} \right\} d\mathbf{x} \right)^{-1} .$$

Since the energy function is defined on covariance matrix, one Gibbs density function cannot describe a system properly. We therefore introduce the Gibbs mixture model with K Gibbs and inverse Gibbs mixture

$$g^{A/R}(\mathbf{x}) = \sum_{k=1}^K \alpha_k g_k^{A/R}(\mathbf{x})$$

where the weight α_k also keeps the condition as the MEEM case to maintain the property of probability.

Now, we have formed penalty functions using Gibbs and inverse Gibbs density function defined on the distance between centers. Given a distance, we already know the two centers. The probability that a given distance $\mathbf{x}$ is originated from k^{th} center is α_k. Therefore the logarithm of the likelihood of penalties for θ can be written as

$$L_{A/R}(\theta) = -\sum_{i=1}^{K} \log \alpha_i \sum_{j=1}^{K} \alpha_j g_j^{A/R}(\mathbf{m}_i - \mathbf{m}_j) \qquad (14)$$

where minus signs are added in order to estimate the parameter by minimizing the effect of other centers.

Now we have the logarithm of the penalty function for the Gibbs distribution and inverse Gibbs distribution to find equilibrium between two quantities, the attraction and the repulsion in (14). Therefore an augmented likelihood function $L_{AR}(\theta)$ is

$$L_{AR}(\theta) = L_L(\theta) + L_A(\theta) + L_R(\theta)\ . \qquad (15)$$

The augmented likelihood function (15) can be solved by the EM algorithm. The expectation of t step for the EM algorithm can be separated into three terms. They are for likelihood, attraction and repulsion. The EM steps for likelihood is same as the MEEM case. The expectations for the penalties are,

$$p_{A/R}^t(i,j) = \frac{\alpha_i \alpha_j g_j^{A/R}(\mathbf{m}_i - \mathbf{m}_j)}{\sum_{m=1}^{K}\sum_{n=1}^{K} \alpha_m \alpha_n g_m^{A/R}(\mathbf{m}_n - \mathbf{m}_m)}\ .$$

The lower bound $\beta_A^t(\theta)$ for attraction and $\beta_R^t(\theta)$ for repulsion by Jensen's inequality is

$$\beta_{A/R}^t(\theta) = -\sum_{i=1}^{K}\sum_{j=1}^{K} p_{A/R}^t(i,j) \log \frac{\alpha_i \alpha_j g_j^{A/R}(\mathbf{m}_i - \mathbf{m}_j)}{p_{A/R}^t(i,j)}\ .$$

Hence, the lower bound $\beta_{AR}^t(\theta)$ for the augmented likelihood function $L_{AR}(\theta)$ is,

$$\beta_{AR}^t(\theta) = \beta_L^t(\theta) + \beta_A^t(\theta) + \beta_R^t(\theta)\ . \qquad (16)$$

The new estimates of the parameters are easily calculated by setting the derivatives of $\beta_{AR}^t(\theta)$ with respect to each parameters to zero. The update equation for mean $\mathbf{m}_k^{t+1}$ is

$$\mathbf{m}_k^{t+1} = \left[\left\{\sum_{n=1}^{N} p_L^t(k,n) + \sum_{i=1}^{K} p_t^{AR}(i,k)\right\}\mathbf{C}_k^{-1} + \sum_{i=1}^{K} p_t^{AR}(k,i)\mathbf{C}_i^{-1}\right]^{-1} \left[\mathbf{C}_k^{-1}\left\{\sum_{n=1}^{N} p_L^t(k,n)\mathbf{x}_n + \sum_{i=1}^{K} p_t^{AR}(i,k)\mathbf{m}_i\right\} + \sum_{i=1}^{K} \mathbf{C}_i^{-1} p_t^{AR}(k,i)\mathbf{m}_i\right]\ .$$

where $p_t^{AR}(i,j)$ is,

$$p_t^{AR}(i,j) = \frac{p_A^t(i,j)}{T_A} - \frac{p_R^t(i,j)}{T_R} \qquad (17)$$

Table 1. SNR comparison of algorithm for image reconstruction

Algorithm	SNR (dB)
70% sampled image	5.166dB
HMEKDE	16.026dB
Conventional EM	23.077dB
MEEM	23.198dB
AREM	25.843dB

The equilibrium between over-fitting and over-smooth can be achieved by adjust the temperature T_A and T_R in (17). The newly estimated covariance $\mathbf{C}_k^{t+1}$ is,

$$\mathbf{C}_k^{t+1} = \left\{\sum_{n=1}^{N} p_L^t(k,n)(\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T + \sum_{i=1}^{K} p_t^{AR}(i,k)(\mathbf{m}_i - \mathbf{m}_k)(\mathbf{m}_i - \mathbf{m}_k)^T\right\} \left\{\sum_{n=1}^{N} p_L^t(k,n) + \sum_{i=1}^{K} p_t^{AR}(i,k)\right\}^{-1}\ .$$

The update equation for α_k^{t+1} using soft-max function is,

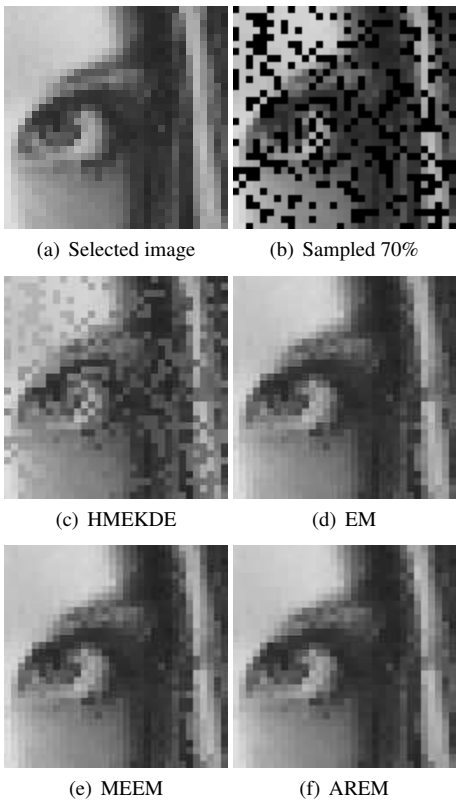
$$\alpha_k^{t+1} = \frac{1}{N-4}\left\{\sum_{n=1}^{N} p_L^t(k,n) - \sum_{i=1}^{K}\left\{p_A^t(i,k) + p_A^t(k,i) + p_R^t(i,k) + p_R^t(k,i)\right\}\right\}\ .$$

Those equations forms the closed form update equation for EM algorithm, therefore the parameters can be estimated iteratively.

4. EXPERIMENTAL RESULTS

We apply our algorithms and other conventional methods to a image reconstruction problem from samples. For experiment, we use 32×32 image selected from original 256×256 gray Lena image which shown in fig. 1(a). We use 70% samples and 49 centers. The sampled image is shown in Fig. 1(b). We use density model (L, i, j) [8] where L is intensity value of the given location (i, j). Here we compare the result of the proposed algorithm with conventional density estimation algorithm appears in [2]. We can estimate the intensity value of given location (i, j) using expectation operator of marginal density distribution function. The reconstructed images by various algorithm are shown in fig 1(c)$\sim$(f) and the signal to noise ratio of the results is given in table 1. We can observe that the bright hair (right side of the image) is properly smoothed than other algorithm.

Another application for the proposed algorithms is sensor field estimation. The density function of given field can be estimated from randomly scattered sensors. For the experiment, we generated a 256×256 polynomial surfaces and use 2% of the original field by random sampling. The original field is shown in Fig. 2 (a) and the sampled field is shown in Fig. 2(b). We also use 49 centers for experiment except HMEKDE which cannot determine the number of centers. The estimation results are given in figures 2(c)$\sim$(f). The SNR results are given in table 2.

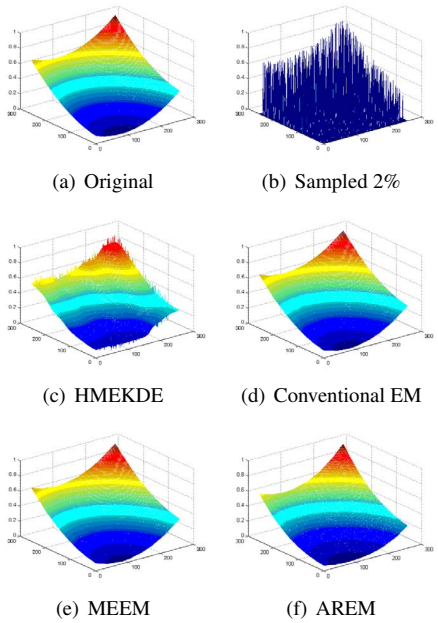


(a) Selected image (b) Sampled 70%

(c) HMEKDE (d) EM

(e) MEEM (f) AREM

Fig. 1. Comparison of density estimation for image reconstruction from randomly sampled image.



(a) Original (b) Sampled 2%

(c) HMEKDE (d) Conventional EM

(e) MEEM (f) AREM

Fig. 2. Comparison of the sensor field estimation from randomly sampled scattered sensor networks.

Table 2. SNR comparison of algorithm for sensor field estimation

Algorithm	SNR (dB)
2% sampled image	0.092dB
HMEKDE	23.053dB
Conventional EM	40.921dB
MEEM	40.958dB
AREM	41.206dB

5. CONCLUSION

In this paper, we develop two new algorithms for density estimation using a penalty-based solution to the expectation-maximization (EM) algorithm with maximum entropy and Gibbs, inverse Gibbs penalties. The proposed maximum-entropy expectation-maximization (MEEM) algorithm relies on a maximum entropy penalty which provides an iterative method to obtain a smooth estimate of the density function. Whereas, the proposed attractive-repulsive expectation-maximization (AREM) algorithm aims to achieve a balance between over-smoothing and over-fitting by reaching equilibrium between attraction and repulsion penalties characterized by Gibbs and inverse-Gibbs densities, respectively. The simulation results using the proposed algorithms show superior performance compared to traditional density estimation methods for image reconstruction and sensor field estimation from randomly scattered samples.

6. REFERENCES

[1] E. Parzen, "On estimation of a probability denstiy function and mode," *Anals of Math. Statistics*, vol. 33, pp. 1065–1076, 1962.

[2] N. Balakrishnan and D. Schonfeld, "A maximum entropy kernel density estimator with applications to function interpolation and texture segmentation," in *SPIE Proceedings of Electronic Imaging: Science and Technology. Conference on Computational Imaging IV*, San Jose, California, 2006.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[4] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438.

[5] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. 1998, Kluwer.

[6] C. Bishop, *Neural Networks for Pattern Recogntion*, Oxford Univ. Press, 1995.

[7] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition*, Academic Press, San Diego, CA, USA, 1999.

[8] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.