Two Dimensional Projection Pursuit Applied to Gaussian Mixture Model Fitting

Natella Likhterov and Mayer Aladjem
Department of Electrical and Computer Engineering,
Ben-Gurion University of the Negev.
Beer-Sheva 84105 Israel

ABSTRACT

In this paper we seek a *Gaussian mixture model* (GMM) of an n-variate probability density function. Usually the parameters of GMMs are determined by a *maximum likelihood* (ML) criterion. A practical deficiency of ML fitting of GMMs is poor performance when dealing with high-dimensional data since a large sample size is needed to match the accuracy that is possible in low dimensions. We propose a method to fit the GMM to multivariate data which is based on the two-dimensional *projection pursuit* (PP) method. By means of simulations we compare the proposed method with a one-dimensional PP method for GMM. We conclude that a combination of one- and two-dimensional PP methods could be useful in some applications.

Keywords: Multivariate density estimation, Gaussian mixture models, Projection pursuit.

1. INTRODUCTION

Density estimation [3], [5], [6], [7] is an important issue for data analysis, because in most cases the density function is unknown and must be estimated. In this work we investigate the estimation of a *Gaussian mixture model* (GMM) of a multivariate probability density function. The GMM is a very important element of the statistical toolbox, in particular for pattern recognition. This model has proved quite useful in modelling complex distributions. Using a small number of normal components, one is able to model distributions that are far from normal.

The determination of the adjustable parameters of the GMM is usually carried out by an *expectation maximization* (EM) algorithm [7]. The EM procedure is very easy to implement, but there is difficulty with its poor performance when dealing with high-dimension data.

We propose a method to fit the GMM to multivariate data, which is based on the two-dimensional *projection pursuit* (PP) method [2]. It extends our method proposed in [1] previously.

We consider the problem of modelling a multivariate probability density function *p(x)* ($\mathbf{x} \in R^n$) on the basis of a data set

$$X=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}.$$

Here, $\mathbf{x}_i \in R^n$, i=1,2,..N are data points drawn from that density. We need a normalization of the data, called *sphering* [2]. The *sphered* **X** has a zero sample mean vector and identity sample covariance matrix. In the following explanation all operations are performed on the sphered data.

In this work we seek a GMM [5], [6], [7] of $p(\mathbf{x})$, which is a linear combination of M Gaussian densities.

$$\hat{p}(\mathbf{x}) = \sum_{j=1}^{M} \omega_j \phi_{\Sigma_j}(\mathbf{x} - \mathbf{m_j}). \quad (1)$$

Here, ω_j are the mixing coefficients, which are non-negative and sum to one and $\phi_{\Sigma_j}(\mathbf{x} - \mathbf{m_j})$ denotes the $N(\mathbf{m}_j, \mathbf{\Sigma}_j)$ density in the vector **x**.

2. DENSITY ESTIMATION BY TWO DIMENSIONAL PROJECTION PURSUIT

Following Friedman [2] we estimate the density p(**x**) by multiplication of K bivariate augmenting functions $f_k(.)$.

$$\hat{p}(\mathbf{x}) = \phi(\mathbf{x}) \prod_{k=1}^{K} f_k(\mathbf{a}_k^T \mathbf{x}, \mathbf{b}_k^T \mathbf{x}), \quad \mathbf{a}_k^T \mathbf{b}_k = 0, \quad (2)$$

where $\phi(\mathbf{x})$ is the n-variate standard normal density function $N(\mathbf{0}, \mathbf{I})$ in the vector **x**, $\mathbf{a}_k$ and $\mathbf{b}_k$ are orthonormal unit vectors specifying a projection plane in R^n and f_k is

$$f_k(y_1, y_2) = \frac{\hat{p}_k(y_1, y_2)}{\phi(y_1)\phi(y_2)}. \quad (3)$$

In (3) $\varphi(y_1)$ and $\varphi(y_2)$ denotes N(0,1) densities in the variables y_1 and y_2, and $\hat{p}_k(y_1, y_2)$ is a bivariate density approximation into the plane spanning $\mathbf{a}_k$, $\mathbf{b}_k$. We compute the direction vectors $\mathbf{a}_k$ and $\mathbf{b}_k$ using a two dimensional PP method [2, Secton 2]. The number K of the augmenting functions is set by a standard test of nonnormality [2, Section 7] and the bivariate density $\hat{p}_k(y_1, y_2)$ was approximated by a Legendre polynomial expansion [2, Section 4].

3. GMM EXPANSION OF THE DENSITY ESTIMATION

Following the idea of our previous work [1] we approximate $\hat{p}_k(y_1, y_2)$ by a mixture of bivariate Gaussians

$$p_k(y_1, y_2) = \sum_{j=1}^{M_k} \omega_{kj} \varphi_{\Sigma_{kj}}(y_1 - \mu_{kj1}, y_2 - \mu_{kj2}) .(4)$$

Using the vector notation $\mathbf{y} = [y_1 \; y_2]^T$ and $\boldsymbol{\mu}_{kj} = [\mu_{kj_1} \; \mu_{kj_2}]^T$ we have (4) in the form

$$p_k(\mathbf{y}) = \sum_{j=1}^{M_k} \omega_{kj} \varphi_{\Sigma_{kj}}(\mathbf{y} - \boldsymbol{\mu}_{kj}) . \tag{5}$$

Substituting (5) into (3) we obtain $f_k(\boldsymbol{y})$ in the form of GMM

$$f_k(\mathbf{y}) = \sum_{j=1}^{Mk} \tilde{\omega}_{kj} \tilde{\varphi}_{\Sigma_{kj}}(\mathbf{y} - \tilde{\mathbf{m}}_{kj}) , \tag{6}$$

with

$$\tilde{\boldsymbol{\Sigma}}_{kj} = \left(\boldsymbol{\Sigma}_{kj}^{-1} - \mathbf{I}\right)^{-1} , \tag{7}$$

$$\tilde{\mathbf{m}}_{kj} = (\mathbf{I} - \boldsymbol{\Sigma}_{kj})^{-1} \mathbf{m}_{kj} , \tag{8}$$

$$\tilde{\omega}_{kj} = \omega_{kj} \frac{2\pi \cdot \left|\tilde{\boldsymbol{\Sigma}}_{kj}\right|^{1/2}}{\left|\boldsymbol{\Sigma}_{kj}\right|^{1/2}} \exp\{-\frac{1}{2}(-\mathbf{m}_{kj}^T \boldsymbol{\Sigma}_{kj}^{-1} \mathbf{m}_{kj}$$

$$+ \mathbf{m}_{kj}^T (\mathbf{I} - \boldsymbol{\Sigma}_{kj}) \cdot \boldsymbol{\Sigma}_{kj}^{-1} \mathbf{m}_{kj})\} . \tag{9}$$

Then substituting (6) into (2) we have

$$\hat{p}(\mathbf{x}) = \phi(\mathbf{x}) \prod_{k=1}^{K} \left(\sum_{j=1}^{Mk} \tilde{\omega}_{kj} \varphi_{\tilde{\Sigma}_{kj}}(\mathbf{y} - \tilde{\mathbf{m}}_{kj}) \right) . \tag{10}$$

Finally we employ the identity

$$\phi_{\Sigma}(\mathbf{x} - \mathbf{m}) \phi_{\mathbf{S}}(\mathbf{A}^{\mathbf{T}}\mathbf{x} \text{-} \mathbf{M}) = \alpha \phi_{\tilde{\Sigma}}(\mathbf{x} \text{-} \tilde{\mathbf{m}}) \tag{11}$$

with

$$\tilde{\boldsymbol{\Sigma}} = \left(\mathbf{A}\mathbf{S}^{-1}\mathbf{A}^T + \boldsymbol{\Sigma}^{-1}\right)^{-1} , \tag{12}$$

$$\tilde{\mathbf{m}} = \tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\mathbf{m} + \tilde{\boldsymbol{\Sigma}}\left(\mathbf{A}\mathbf{S}^{-1}\mathbf{A}^T\right)\mathbf{A}\mathbf{M} , \tag{13}$$

$$\alpha = \frac{\left|\tilde{\boldsymbol{\Sigma}}\right|^{1/2}}{2\pi \left|\boldsymbol{\Sigma}\right|^{1/2} \left|\mathbf{S}\right|^{1/2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} (\boldsymbol{\mu}^*)^T \left[(\mathbf{S}^*)^{-1} \boldsymbol{\Sigma} (\mathbf{S}^*)^{-1}\right] (\boldsymbol{\mu}^*) \\ + \mathbf{m}^T \left[\boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\right] \mathbf{m} \\ + (\boldsymbol{\mu}^*)^T [(\mathbf{S}^*)^{-1} \tilde{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1}] \mathbf{m} \\ + \mathbf{m}^T [\boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\Sigma}} (\mathbf{S}^*)^{-1}] (\boldsymbol{\mu}^*) \end{pmatrix} \right\} \tag{14}$$

$$(\mathbf{S}^*)^{-1} = \mathbf{A}\mathbf{S}^{-1}\mathbf{A}^T , \tag{15}$$

$$\boldsymbol{\mu}^* = \mathbf{A}\mathbf{M} . \tag{16}$$

In (11)-(16) **A**=[**a**$_k$ **b**$_k$] is an n×2 projection matrix, $\phi_{\Sigma}(\mathbf{x} - \mathbf{m})$ is a n-variate normal density *N*(**m, Σ**) in vector **x** and $\phi_S(\mathbf{A}^{\mathbf{T}}\mathbf{x} - \mathbf{M})$ is the bivariate *N*(**M**, **S**) in vector **A**$^{\mathbf{T}}$**x**. The identity (11) shows that the multiplication of any n-variate normal density $\phi_{\Sigma}(\mathbf{x} - \mathbf{m})$ by any bivariate normal density $\phi_s(\mathbf{A}^{\mathbf{T}}\mathbf{x} - \mathbf{M})$ gives an n-variate normal density function $\phi_{\tilde{\Sigma}}(\mathbf{x} - \tilde{\mathbf{m}})$ scaled by a constant α. The proof of identity (11) is in Appendix.

After an interactive application of identity (11) to (10) the PP approximation (2) becomes the form of a GMM with $\tilde{M} = \prod_{k=1}^{K} M_k$ Gaussian components

$$\hat{p}(\mathbf{x}) = \sum_{j=1}^{\tilde{M}} \tilde{\omega}_j \phi_{\tilde{\boldsymbol{\Sigma}}_j}(\mathbf{x} - \tilde{\mathbf{m}}_j) . \tag{17}$$

Here $\tilde{\omega}_j$, $\tilde{\boldsymbol{\Sigma}}_j$ and $\tilde{\mathbf{m}}_j$ denote the parameter values calculated by expressions (11)-(16).

4. COMPARATIVE STUDIES

In this section we compare the performances of our new method (Sections 3) and our previous PP method [1]. We study a wide spectrum of situations in terms of the size N of the training samples drawn from 4-dimensional densities

$$p_{\mathbf{IJ}}(\mathbf{x}) = [\textstyle\sum_{j=1}^{3} \alpha_j g_{\mathbf{I}j}(x_1, x_2) g_{\mathbf{J}j}(x_3, x_4)], \tag{18}$$

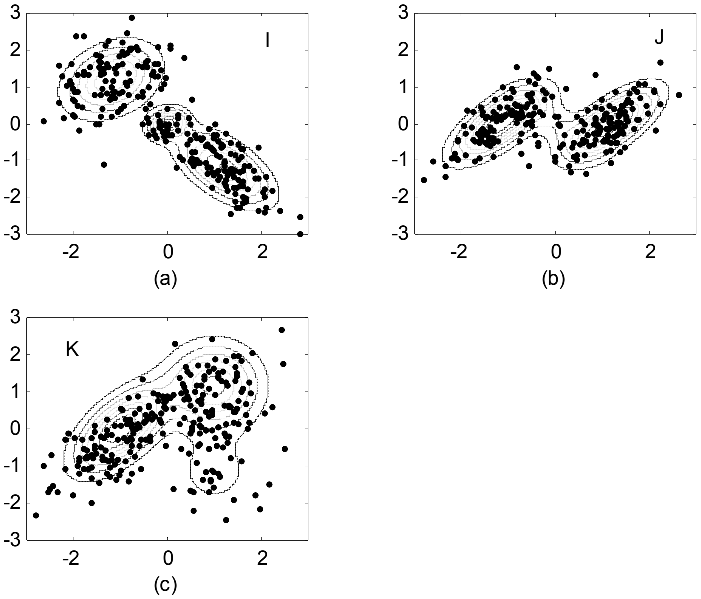
$$p_{\mathbf{IK}}(\mathbf{x}) = [\textstyle\sum_{j=1}^{3} \alpha_j g_{\mathbf{I}j}(x_1, x_2) g_{\mathbf{K}j}(x_3, x_4)]. \tag{19}$$

Here $g_{\mathbf{I}j}$, $g_{\mathbf{J}j}$, $g_{\mathbf{K}j}$ for j=1, 2, 3 are bivariate normal densities, $\alpha_1=\alpha_2=9/20$ and $\alpha_3=1/10$, $\mathbf{x}=[x_1 \; x_2 \; x_3 \; x_4]^T$. The parameters of $g_{\mathbf{I}j}$, $g_{\mathbf{J}j}$, $g_{\mathbf{K}j}$ are listed in Table1 taken from [8].

Table 1: Parameters of the bivariate normal densities [8]

I-density parameters	$g_{\mathbf{I}1}(x,y)$ N(-1.2,1.2;0.36,0.36,0.3)*	$g_{\mathbf{I}2}(x,y)$ N(1.2-,1.2;0.36,0.36,-0.6)	$g_{\mathbf{I}3}(x,y)$ N(0.,0;0.0625 ,0.0625 ,.0.2)
K-density parameters	$g_{\mathbf{K}1}(x,y)$ N(-1.2,0;0.36,0.36,0.7)	$g_{\mathbf{K}2}(x,y)$ N(1.2-,0;0.36,0.36,.0.7)	$g_{\mathbf{K}3}$(x,y) N(0.,0; 0.36, 0.36,-0.7)
J-density parameters	$g_{\mathbf{J}1}(x,y)$ N(-1,0;0.36,0.49,0.6)	$g_{\mathbf{J}2}(x,y)$ N(-1,1.1547 ;0.36,0.49,0)	$g_{\mathbf{J3}}(x,y)$ N(1,-1.1547 ;0.36,0.49,0)

* Here, for easy of presentation, N(μ_1, μ_2; σ_1^2, σ_2^2, ρ) denotes the bivariate normal density, where two marginal means and variances are μ_i and σ_i^2 for i=1, 2 and the correlation coefficient is ρ.



(a)

(b)

(c)

Figure 1. Contour plots of the bivariate densities. (a) $p_{\mathbf{I}}(x,y) = \sum_{j=1}^{3} \alpha_j g_{\mathbf{I}j}(x,y)$, (b) $p_{\mathbf{J}}(x,y) = \sum_{j=1}^{3} \alpha_j g_{\mathbf{J}j}(x,y)$, (c) $p_{\mathbf{K}}(x,y) = \sum_{j=1}^{3} \alpha_j g_{\mathbf{K}j}(x,y)$. There are 250 data points superimposed on the contours to show the locations of the training points drawn from these densities.

These densities have been carefully chosen because they combine the benchmarks widely used for comparison density estimation methods [8].

We study the influence of the sample size on the result for N=200, 400, 600, 800, 1000, 1500, 3000. An experiment for a given combination of particular settings, density function and sample size consisted of the following run. We drew a training sample of size N from an appropriate distribution. Then we normalized (*sphered* [2]) the data. Using this data we fitted GMMs by the method proposed in Section 3 and our previous method [1].

For our new method (Section 3) the number *K* of augmenting functions of the PP approximation (2) were set to *K*=1, 2 and the number M_k of the components of the bivariate GMMs (4) was M_k= 2, 3, 4, 5. For our previous method the setting of the parameter variation is explained in [1, Section 4].

We compared the performance of the density estimation by a criterion called *percentage of variance explained* (PVE) [3]. Among the variations of the parameter values we selected those corresponding to the largest (best) PVE. In Figs. 2, 3 we show the training sample size versus the "best" PVE. The solid line (—) shows the results for our current method and the dashed line (--) the result of method [1].

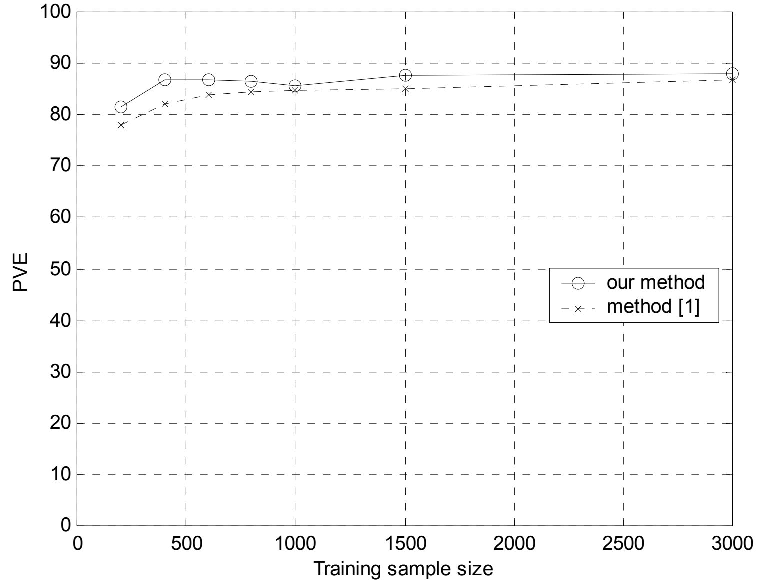


Figure 2. Estimation of $p_{IJ}(x_1, x_2, x_3, x_4)$ (18).

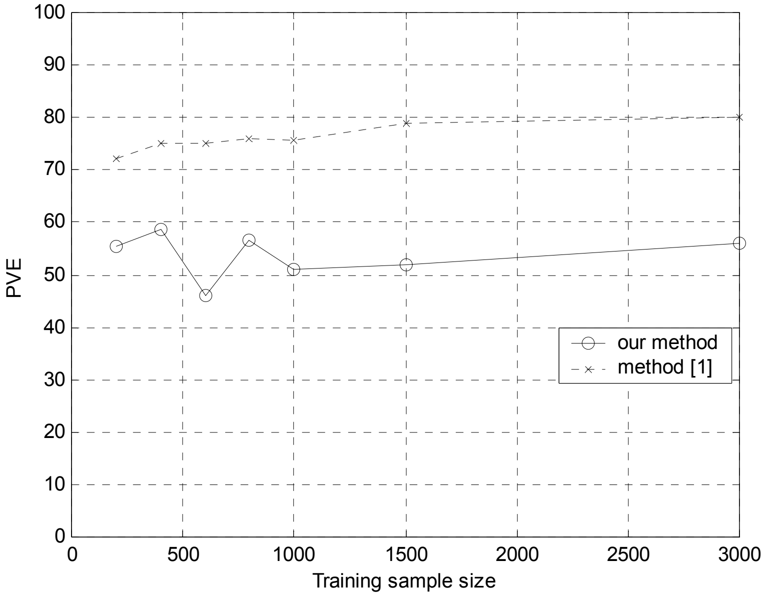


Figure 3. Estimation of $p_{IK}(x_1, x_2, x_3, x_4)$ (19).

Observing Fig. 2 we conclude that the proposed two-dimensional PP method yields better results (exhibits higher PVE values) than our previous one-dimensional PP method [1] for estimation p_{IJ}(x) (18) for all samples sizes. The latter is consistent with the observation in [4, Section 7] that in some cases the two dimensional PP could find data structure that the one dimensional PP may miss.

The result for p_{IK}(x) (19) in Fig. 3 shows the better performance of the method [1]. This is due to the fact that the orthogonal constraint ($\mathbf{a}_k^T \mathbf{b}_k = 0$) in (2) restricts the searching of $\mathbf{a}_k$ and $\mathbf{b}_k$. Consequently a compromise between one- and two-dimensional PP seems to be useful for applications. Combination of the PP methods is the object of our current research.

5. CONCLUSION

We have proposed a method for fitting GMMs based on the two dimensional *projection pursuit* (PP) strategy proposed by Friedman [2]. In Section 3 we showed that the PP density estimation implies a GMM model for a specific setting of augmenting functions. The derived formulae (11)-(16) allow us to set the parameters of the GMM implied by the PP estimation. In Section 4 we give the results of a comparative study of our method and the method [1]. We concluded that a combination of the one- and two-dimensional PP methods could be useful for the applications.

6. APPENDIX: PROOF OF THE IDENTITY (11)

According to the expressions of $N(\mathbf{m}, \mathbf{\Sigma})$ and $N(\mathbf{M}, \mathbf{S})$ we have for left-hand part of the equality (11)

$$\phi_{\Sigma}(\mathbf{x}-\mathbf{m})\phi_{\mathrm{S}}(\mathbf{A}^{\mathbf{T}}\mathbf{x}\text{-}\mathbf{M}) = \frac{\exp\left\{-\frac{1}{2}\gamma\right\}}{(2\pi)^{\frac{n}{2}+1}|\mathbf{\Sigma}|^{\frac{1}{2}}|\mathbf{S}|^{\frac{1}{2}}} \tag{20}$$

with

$$\gamma = (\mathbf{x}-\mathbf{m})^T\mathbf{\Sigma}^{\mathbf{-1}}(\mathbf{x}-\mathbf{m}) + (\mathbf{A}^{\mathbf{T}}\mathbf{x}\text{-}\mathbf{M})^T\mathbf{S}^{\mathbf{-1}}(\mathbf{A}^{\mathbf{T}}\mathbf{x}\text{-}\mathbf{M}). \tag{21}$$

First we note that for $\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}$ we have

$$\gamma = (\mathbf{x}-\mathbf{m})^T\mathbf{\Sigma}^{\mathbf{-1}}(\mathbf{x}-\mathbf{m}) + (\mathbf{x}\text{ - }\mathbf{AM})^T\mathbf{AS}^{\mathbf{-1}}\mathbf{A}^{\mathbf{T}}(\mathbf{x}\text{ - }\mathbf{AM}). \tag{22}$$

Then we seek for n×n matrices **B** and **C**, and scalar γ^* that imply (21) in the form

$$\gamma = (\mathbf{x}-\tilde{\mathbf{m}})^T\tilde{\mathbf{\Sigma}}^{\mathbf{-1}}(\mathbf{x}-\tilde{\mathbf{m}}) - \gamma^* \tag{23}$$

for

$$\tilde{\mathbf{m}} = \mathbf{Bm} + \mathbf{CAM}, \tag{24}$$

$$\tilde{\mathbf{\Sigma}} = \left(\mathbf{AS}^{-1}\mathbf{A}^T + \mathbf{\Sigma}^{-1}\right)^{-1}. \tag{25}$$

Combining (22) and (23), (24), (25) we have

$$\gamma^* = (\boldsymbol{\mu}^*)^T\left[(\mathbf{S}^*)^{-1}\mathbf{\Sigma}(\mathbf{S}^*)^{-1}\right](\boldsymbol{\mu}^*) + \mathbf{m}^T\left[\mathbf{\Sigma}^{-1}\tilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\right]\mathbf{m} + (\boldsymbol{\mu}^*)^T[(\mathbf{S}^*)^{-1}\tilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}]\mathbf{m} + \mathbf{m}^T[\mathbf{\Sigma}^{-1}\tilde{\mathbf{\Sigma}}(\mathbf{S}^*)^{-1}](\boldsymbol{\mu}^*), \tag{26}$$

where

$$(\mathbf{S}^*)^{-1} = \mathbf{AS}^{-1}\mathbf{A}^T, \tag{27}$$

$$\boldsymbol{\mu}^* = \mathbf{AM} \tag{28}$$

and

$$\mathbf{B} = \tilde{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}, \quad \mathbf{C} = \tilde{\mathbf{\Sigma}}\mathbf{AS}^{-1}\mathbf{A}^T. \tag{29}$$

Then substitution γ (23) into the right-hand part of (20) and using γ^*(26) we obtain the identity (11) for α (14). Finally, substitution **B** and **C** (29) into (24) we have $\tilde{\mathbf{m}}$ (13).

7. ACKNOWLEDGEMENT

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University of the Negev, Israel

8. REFERENCES

[1] M.E. Aladjem, "Projection pursuit fitting Gaussian mixture models", Eds. T.Caelli, Amin A., Duin R.P.W., Kamel M. and Ridder D., **Advances in Statistical, Structural and Syntactical Pattern Recognition, joint IAPR international Workshops SSPR 2002 and SPR 2002, Windsor, Ontario, Canada, August 2002, Lecture notes in Computer Science**, Springer, pp.380-388, 2002.

[2] J.H. Friedman, "Exploratory projection pursuit", **Journal of the American Statistical Association**, vol.82, pp.249-266, 1987.

[3] J.H. Friedman , W. Stuetzle, and A. Schroeder, "Projection pursuit density estimation", **Journal of the American Statistical Association**, vol.79, pp.599-608, 1984.

[4] P.J.Huber, "Projection pursuit" (with discussion), **The Annals of Statistics**, Vol.13, pp.435-525, 1985.

[5] G.J. McLanchlan and K.E. Basford, **Mixture Models: Inference and Applications to Clustering**, Marcel Dekker, 1988.

[6] D.M. Titterington, A.F.M. Smith, and U.E. Makov, **The Statistical Analysis of Finite Mixture Distributions**, New York: Willey, 1985.

[7] M.E. Tipping and C.M.Bishop, "Mixtures of Probabilistic Principal Component Analyzers", **Neural Computation**., vol. 11, pp.443-482, 1999.

[8] M.P. Wand and M.C. Jones, "Comparison of smoothing parameterizations in bivariate kernel density estimation", **Journal of the American Statistical Association**, Vol.88, No.422, pp.520-528, 1993.