Probabilistic Principal Component Subspaces: A Hierarchical Finite Mixture Model for Data Visualization

Yue Wang, Lan Luo, Matthew T. Freedman, and Sun-Yuan Kung, *Fellow, IEEE*

***Abstract*—Visual exploration has proven to be a powerful tool for multivariate data mining and knowledge discovery. Most visualization algorithms aim to find a projection from the data space down to a visually perceivable rendering space. To reveal all of the interesting aspects of multimodal data sets living in a high-dimensional space, a hierarchical visualization algorithm is introduced which allows the complete data set to be visualized at the top level, with clusters and subclusters of data points visualized at deeper levels. The methods involve hierarchical use of standard finite normal mixtures and probabilistic principal component projections, whose parameters are estimated using the expectation-maximization and principal component neural networks under the information theoretic criteria. We demonstrate the principle of the approach on several multimodal numerical data sets, and we then apply the method to the visual explanation in computer-aided diagnosis for breast cancer detection from digital mammograms.**

***Index Terms*—Computer-aided diagnosis, data visualization, hierarchical mixture distribution, information theoretic criteria, principal component neural network, soft clustering.**

I. Introduction

As a step toward understanding multivariate data sets, cluster information reveals insight that may prove useful in knowledge discovery since the growing volume of complex data are often high dimensional, multimodal, and lacking in prior knowledge [1]–[3], [6]. Several new visualization methods have been progressively developed to model and display the contents of the data sets [1], [3]–[6], [8], [11]. However, although such algorithms can usefully characterize the content of simple data sets, little comprehensive study has been reported that proves adequate in the face of multimodal and high dimensional data sets [1], [6], [11]. For example, a single projection of the data onto a visualization space may not be able to capture all of the interesting aspects of the data set. This motivates the consideration of a hierarchical visualization paradigm involving hierarchical statistical models and visualization spaces.

Once we explore the possibility of using many complementary visualization subspaces, cluster decomposition and dimensionality reduction are the two natural strategies. Cluster decomposition permits the use of relatively simple models for each of the local structures, offering greater ease of interpretation as well as the benefits of analytical and computational simplification. This philosophy for modeling complexity is similar in spirit to the *divide-and-conquer* principle [4], [7], [12]. On the other hand, dimensionality reduction allows better visual interpretation and less computational demand. Many researchers have recently proposed various methods to improve data visualization [3], [6]. The work most closely related to our methodology was reported by Bishop and Tipping in [1] and [9]. They introduce a hierarchical modeling and visualization algorithm based on a two-dimensional (2-D) hierarchical mixture of latent variable models, whose parameters are estimated using the expectation-maximization (EM) algorithm [1], [16]. The construction of the hierarchical tree proceeds top down in which the cluster decomposition is driven interactively by the user, and optimal projection is determined by maximum likelihood principle. There are three major potential limitations associated with the present approach [1], [11]. First, although a probability density is defined in the data space through a latent variable model, the *priori* and order of the mixture model are heuristically selected and an isotropic Gaussian conditional distribution is undesirably restricted, which may misrepresent the true data structures [4], [11], [19]. The second important limitation is that the parameters, including optimal projections, are determined by maximum likelihood, and this criterion need not always lead to the most interesting or interpretable visualization plots. For example, alternative models may be those that optimize other criteria such as the separation of clusters [1], [13]. An additional limitation of the solely user-driven scheme is its subjective nature, which may be highly influenced by the quality of visual interpretability. For example, alternative methods may be those which involve both information theoretic criteria and human input [4], [11], [22].

In this paper, we propose using standard finite normal mixtures (SFNM) and hierarchical visualization spaces for an effective data modeling and visualization. The strategy is that the top-level model and projection should explain the entire data set, best revealing the presence of clusters and relationships, while lower level models and projections should display internal structure within individual clusters, such as the presence of subclusters, which might not be apparent in the higher level models and projections. With many complementary mixture models and

Manuscript received September 18, 1999; revised February 28, 2000. This work was supported in part by the National Institutes of Health under Grant 1R21CA83231-01 and the Department of Defense under Grants DAMD17-98-1-8045 and DAMD17-98-1-8044.

Y. Wang and L. Luo are with the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064 USA (e-mail: wang@pluto.ee.cua.edu).

M. T. Freedman is with the Department of Radiology and the Lombardi Cancer Center, Georgetown University, Washington, DC 20007 USA.

S. Y. Kung is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kung@ee.princeton.edu).

Publisher Item Identifier S 1045-9227(00)04040-6.

visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet still conveys considerable cluster information. Based on the concept of combining finite mixture modeling [16] and principal component projection [1], [11] to guide cluster decomposition and dimensionality reduction, the particular advantages of our algorithm are as follows.

1) At each level, a probabilistic principle component extraction is performed to project the softly partitioned data set down to a 2-D visualization space, leading to an effective dimensionality reduction, allowing effective separation and visualization of local clusters [1], [5], [12].
2) Learning from the data directly, information theoretic criteria are used to select model structures and estimate its parameter values, where the soft partitioning of the data set results in a standard finite normal mixture distribution best fitted to the data [4], [18]–[22].
3) By alternatively performing principal component projection and finite mixture modeling, a complete hierarchy of complementary projections and refined models can be generated automatically, allowing a new paradigm of knowledge discovery [1]–[3], [6].

There are several major differences between our work and the previous most related research [1], [8]–[10]. First, we consider cluster decomposition and dimensionality reduction as two separated but complementary operations, in which the criterion used to effective dimensionality reduction is the separation of clusters rather than maximum likelihood. The resulting projections in turn enhance the performance of cluster decomposition at the next level [1], [13]. Second, we impose a model selection procedure to determine the number of subclusters inside each cluster at each level using information theoretic criteria. This allows the algorithm to automatically determine whether a further split of a subspace should continue or terminate in completing the whole hierarchy [1], [22]. Furthermore, we develop a probabilistic adaptive principal components extraction (PAPEX) algorithm to estimate the top two or three principal axes [5], [12]. When the dimensionality of raw data is high, this approach is computationally very efficient [11]. Finally, our model defines a probability distribution in data space which naturally induces a corresponding distribution in projection space through a Radon transform [15]. This permits an independent procedure in determining values of model parameters without concurrent estimation of projection matrix. In Section II, we introduce the theory and method, and in Section III, we discuss the algorithm to generation of such subspaces. This is further extended to implement an interactive visualization environment in Section IV, where we first illustrate the operation of the algorithm using representative multimodal numerical data sets, and then apply the algorithm to the visual explanation of decision making process in computer-aided diagnosis for breast cancer detection. Finally, extensions to the applications, and the relationships to other approaches, are discussed in Section V.

II. Theory and Method

One of the difficulties inherent in data visualization is the problem of visualizing multidimensionality [1], [3], [6]. When there are more than three variables, it stretches the imagination to visualize their relationships. Fortunately, in data set with many variables, groups of variables often form clusters [10], [12], [13]. Thus, our approach includes two major complementary components: 1) dimensionality reduction by probabilistic principal component projection and 2) cluster decomposition by adaptive soft data clustering.

Assume the data points $\{\boldsymbol{t}_i\}$ in the data space come from K_0 clusters $\{\boldsymbol{\theta}_{\boldsymbol{t}1}, \cdots, \boldsymbol{\theta}_{\boldsymbol{t}k}, \cdots, \boldsymbol{\theta}_{\boldsymbol{t}K_0}\}$, where $\boldsymbol{\theta}_{\boldsymbol{t}k}$ is the Gaussian kernel parameter vector of cluster k in the model. Recently there has been considerable success in using the SFNM to model the distribution of a multimodal data set [1], [4], [7], [16], [23], [24], such that the data distribution takes a sum of the following general form:

$$p(\boldsymbol{t}) = \sum_{k=1}^{K_0} \pi_k g(\boldsymbol{t}|\boldsymbol{\theta}_{\boldsymbol{t}k}) \tag{1}$$

where π_k is the corresponding mixing proportion, with $0 \leq \pi_k \leq 1$ and $\sum \pi_k = 1$, and g is the Gaussian kernel. The problem of SFNM modeling addresses the combined estimation of regional parameters $(\pi_k, \boldsymbol{\theta}_{\boldsymbol{t}k})$ and detection of structural parameter K_0 in (1) based on the observations $\boldsymbol{t}$. One natural criterion used for estimating the parameter values is to minimize the distance between the SFNM distribution $f(\boldsymbol{t})$ and the data histogram $f_{\boldsymbol{t}}$. Suggested by information theory [16], [17], relative entropy (Kullback–Leibler distance) is a suitable measure, given by

$$D(f_{\boldsymbol{t}}\|f) = \sum_{\mathcal{T}} f_{\boldsymbol{t}}(\boldsymbol{t}) \log \frac{f_{\boldsymbol{t}}(\boldsymbol{t})}{f(\boldsymbol{t})}. \tag{2}$$

We have previously shown that distance minimization based on (2) is equivalent to the maximum likelihood (ML) estimation under a data independency approximation [4], and when K_0 is given, the ML estimate of the regional parameters can be obtained using the EM algorithm [12], [16], [23].

There are three major problems associated with the current approach. First, when the dimensionality of the data space is high, the computational complexity of implementing the EM algorithm in $\boldsymbol{t}$-space is very high. Second, the initialization of the EM algorithm is often heuristically chosen, which may lead to both local optima and computational complexity. Finally, since the number of the local clusters in a particular data set is generally unknown, model selection is a prerequisite. A natural way, with greater practical applicability, to tackle these problems is to introduce user interaction with the system [1], [6]. Data mining and knowledge discovery are not processes that can be orchestrated *a priori*. Training algorithms and expected behavior can be specified, but the actual learning must follow for insight and spontaneous inspiration [6]. For example, by examining plots of principal component space, researchers often develop a deeper understanding of the driving forces that generated the original data, and effortlessly grasp the general characteristics of the data and propose an initial solution [1], [3], [6].

Principal component analysis (PCA) is an effective method for achieving dimensionality reduction [8], [9]. For a set of observed d-dimensional data vectors $\{\boldsymbol{t}_i\}$, $i \in \{1, \cdots, N\}$, the q principal axes $\boldsymbol{w}_m$, $m \in \{1, \cdots, q\}$, are those orthogonal

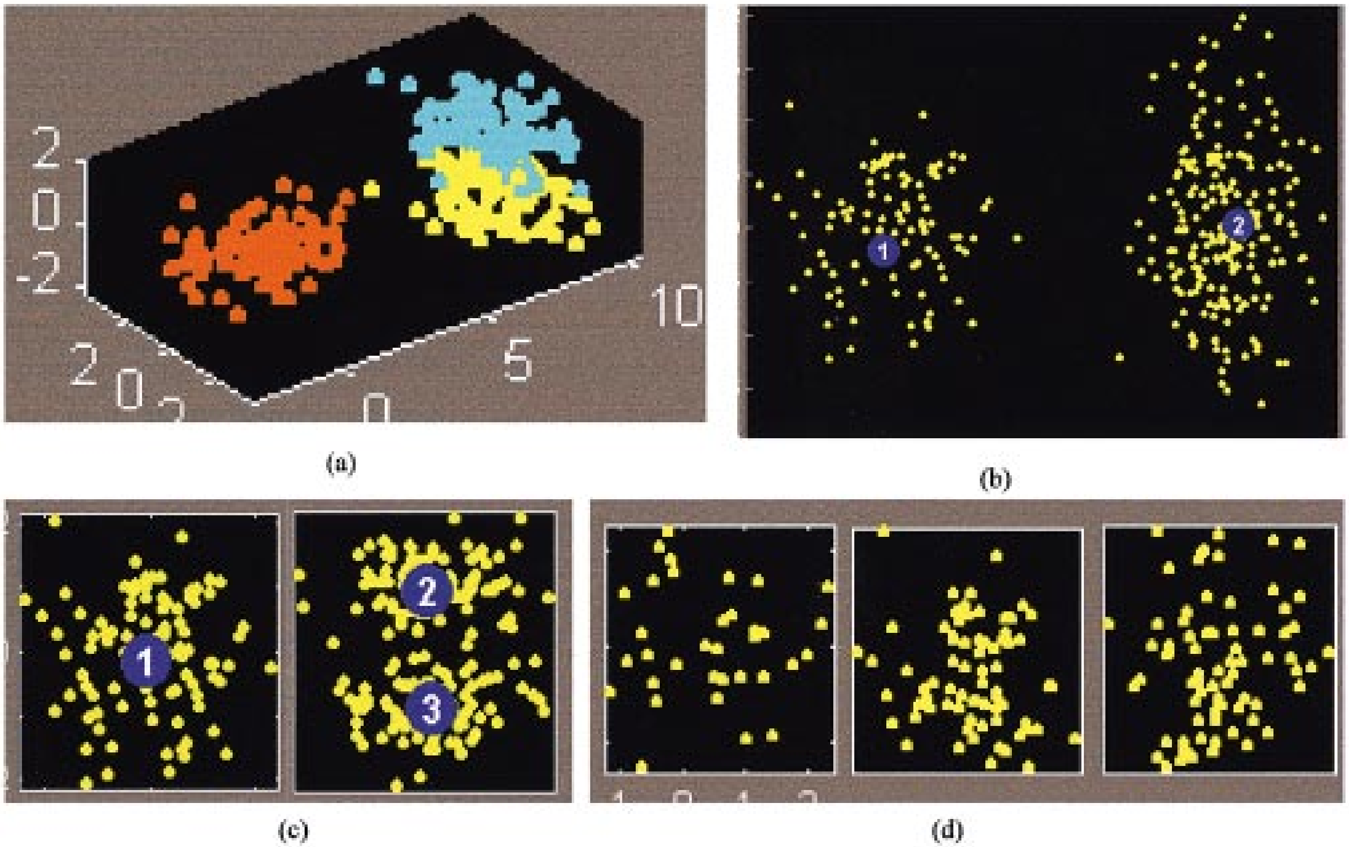


Fig. 1. Results of a demonstrative hierarchical cluster decomposition and dimensionality reduction with a simulated data set. The user initializes the centers of local clusters at the second level and the algorithm then completes the whole three-level hierarchy automatically.

axes onto which the retained variance under projection is maximal. It can be shown that the principal axes $\boldsymbol{w}_m$ are given by the q dominant eigenvectors (i.e., maximal eigenvalues) of the sample covariance matrix $\boldsymbol{C_t} = \sum_i (\boldsymbol{t}_i - \boldsymbol{\mu_t})(\boldsymbol{t}_i - \boldsymbol{\mu_t})^T/N$ such that $\boldsymbol{C_t}\boldsymbol{w}_m = \lambda_m \boldsymbol{w}_m$ and where λ_m is the eigenvalue and $\boldsymbol{\mu_t}$ is the sample mean. The vector $\boldsymbol{x}_i = \boldsymbol{W}^T(\boldsymbol{t}_i - \boldsymbol{\mu_t})$, where $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_q)$, is thus a q-dimensional reduced representation of the observed vector $\boldsymbol{t}_i$. The advantage of PCA is twofold: the projection onto the principal subspace: 1) minimizes the squared reconstruction error [9], [12] and 2) maximizes the separation of data clusters [13]. Although the effectiveness of applying PCA in an unsupervised manner is highly data dependent, our approach has a simple optimal appeal in that if the local clusters are linearly separable in a 2-D or three-dimensional (3-D) space, the principal component projections allow best separation of the clusters [13].

Suppose the data space is d-dimensional. Now consider a 2-D projection space $\boldsymbol{x} = (x_1, x_2)^T$ together with a linear transformation, that maps the data space to the projection space by $\boldsymbol{x} = \boldsymbol{W}^T(\boldsymbol{t} - \boldsymbol{\mu_t})$ where $\boldsymbol{W}$ is a $d \times 2$ matrix. For a normal distribution $p(\boldsymbol{t})$ over the data space, using the rules of probability, a similar reduced dimension probability distribution of the new variables $\{\boldsymbol{x}_i\}$ in the projection space is obtained from the convolution of the projection model with the true distribution over data space in the form of $f(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{t})p(\boldsymbol{t})\, d\boldsymbol{t}$ [1], [9], [14]. Since the conditional distribution $p(\boldsymbol{x}|\boldsymbol{t}) = \delta(\boldsymbol{x} - \boldsymbol{W}^T\boldsymbol{t} + \boldsymbol{W}^T\boldsymbol{\mu_t})$, where $\delta(.)$ is the delta function that $\delta(0) = 1$ and $\delta(\neq 0) = 0$, it can be shown that $f(\boldsymbol{x})$ is simply defined by the Radon transform of $p(\boldsymbol{t})$, i.e., $f(\boldsymbol{x}) = \int p(\boldsymbol{t})\delta(\boldsymbol{x} - \boldsymbol{W}^T\boldsymbol{t} + \boldsymbol{W}^T\boldsymbol{\mu_t})\, d\boldsymbol{t}$ [15]. According to the linear superposition property of Radon transform and the projection invariant property of normal distribution, if $p(\boldsymbol{t})$ is a SFNM distribution, the data distribution in the projection space has a similar reduced dimension form as (1)

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{k=1}^{K_0} \pi_k \int g(\boldsymbol{t}|\boldsymbol{\theta_{t}}_k)\delta(\boldsymbol{x} - \boldsymbol{W}^T\boldsymbol{t} + \boldsymbol{W}^T\boldsymbol{\mu_t})\, d\boldsymbol{t} \\ &= \sum_{k=1}^{K_0} \pi_k g(\boldsymbol{x}|\boldsymbol{\theta_{x}}_k). \end{aligned} \tag{3}$$

However, because of its global linearity, the application of PCA is necessarily somewhat limited [9], [10]. For example, the inherent multimodal nature of the data set may be completely obscured when it is projected onto the lower dimensional principal subspace. Thus, it is important to note that although the cluster structure of the data set may be evident from the higher dimensional plot of the raw data, it is quite conceivable to have the intrinsic cluster structure of the data concealed after a projection in the more general case of high-dimensional data sets [12]. An alternative paradigm is to model multimodal data set with a collection of local linear subspaces through probabilistic principal component analysis as shown in Fig. 1 [9]–[11]. The method is a two-stage procedure: a soft partitioning of the data space followed by estimation of the principal subspace within each partition. For the sake of computational simplicity, it is reasonable to consider the model parameter values being estimated first in the projection space and then further fine tuned in the data space [11].

The association of a SFNM distribution with PCA offers the possibility of being able to visualize complex data structures through a mixture of probabilistic principal component subspaces. By a simple extension of the maximum a posterior for data classification in the standard K_0-ary Bayes hypothesis testing [12], [17], we can obtain a principal component projection along the desired axes onto which a particular portion of the data set is highlighted, by weighting all of the data points in the whole data set with their posterior probabilities belonging to that portion. This involves a soft clustering of the data points in which instead of any given data point being assigned exclusively to one principal component subspace, the responsibility for its generation is shared among all of the subspaces.

Under the SFNM model defined by (1), the posterior Bayesian probability z_{ik} of a given data point $\boldsymbol{t}_i$ belonging to cluster k is

$$z_{ik} = \frac{\pi_k g(\boldsymbol{t}_i|\boldsymbol{\theta}_{\boldsymbol{t}k})}{p(\boldsymbol{t}_i)} \tag{4}$$

where $k = 1, 2, \cdots, K_0$ and $\sum_k z_{ik} = 1$. These posterior probabilities, together with the computational simplicity of performing PCA (involving no more than finding the top q eigenvectors of the covariance matrix of the data points) make it a good candidate for the linear subspace in the mixture. The q principal components define the local subspace assumed for the multimodal. The contributions of the input to the k subspace are the activities of the weighted data points $\{\boldsymbol{t}_{ik}\}$ for input cluster k. This can be obtained by $\boldsymbol{t}_{ik} = z_{ik}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})$, where $\boldsymbol{\mu}_{\boldsymbol{t}k}$ is the weighted sample mean of cluster k

$$\boldsymbol{\mu}_{\boldsymbol{t}k} = \frac{\sum\limits_i z_{ik}\boldsymbol{t}_i}{\sum\limits_i z_{ik}}$$
$$\boldsymbol{C}_{\boldsymbol{t}k} = \frac{\sum\limits_i z_{ik}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})^T}{\sum\limits_i z_{ik}}. \tag{5}$$

The subspaces for the focused clusters are generated by a localized linear PCA such that $\boldsymbol{C}_{\boldsymbol{t}k}\boldsymbol{w}_{mk} = \lambda_{mk}\boldsymbol{w}_{mk}$. It is important to understand that each component in (1) now corresponds to an independent subspace model with parameters $\boldsymbol{\theta}_{\boldsymbol{x}k}$ and $\boldsymbol{W}_k$, where $\boldsymbol{W}_k = (\boldsymbol{w}_{1k}, \boldsymbol{w}_{2k}, \cdots, \boldsymbol{w}_{qk})$. More precisely, consider the vector $\boldsymbol{x}_{ik} = z_{ik}\boldsymbol{W}_k^T(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})$ to be a q-dimensional reduced representation of k-cluster focused vector $\boldsymbol{t}_{ik}$, the corresponding probability distribution is defined by

$$g(\boldsymbol{x}|\boldsymbol{W}_k, \boldsymbol{\theta}_{\boldsymbol{x}k}) = \int g(\boldsymbol{t}|\boldsymbol{\theta}_{\boldsymbol{t}k})\delta(\boldsymbol{x} - \boldsymbol{W}_k^T\boldsymbol{t} + \boldsymbol{W}_k^T\boldsymbol{\mu}_{\boldsymbol{t}k})\, d\boldsymbol{t} \tag{6}$$

where the data mapping by $\boldsymbol{W}_k$ leads to an independent Radon transform. To interpret the corresponding set of visualization subspaces, it may be useful to plot all of the data points on every plot. For this, we may create a k-cluster focused projection in k-subspace by plotting the vector $\boldsymbol{x}_{ik}$, or display the density of "gray-level" in proportion to the contribution which each point has for k-subspace with $h[\boldsymbol{W}_k^T(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})] = z_{ik}$.

An important issue concerning unsupervised cluster decomposition is the detection of the structural parameter K_0, called model selection [4], [11], [12], [16], [22]. This is indeed particularly critical in real-world applications where the structure of the data patterns may be arbitrarily complex [2]. We propose to use two information theoretic criteria, i.e., the Akaike information criterion (AIC) [18] and minimum description length (MDL) [19], to guide model selection. The major thrust of this approach has been the formulation of a model fitting procedure in which an optimal model is selected from the several competing candidates such that the selected model best fits the observed data, under Jaynes' minimax entropy principle stated as "*the parameters in a model which determine the value of the maximum entropy should be assigned values which minimize the maximum entropy*" [20], [21]. For example, AIC tries to reformulate the problem explicitly as an *approximation* of the true structure by the model, implying that AIC will select the model that gives the minimum value defined by

$$\mathrm{AIC}(K_a) = -2\,\log(\mathcal{L}_{ML}) + 2K_a \tag{7}$$

where $\mathcal{L}_{ML}$ is the maximum likelihood of the model and K_a is the number of free adjustable parameters in the model. From a quite different point of view, MDL reformulates the problem explicitly as an information coding problem in which the best model fit is measured such that it assigns high probabilities to the observed data while at the same time the model itself is not too complex to describe [19]. A model is selected by minimizing the total description length defined by

$$\mathrm{MDL}(K_a) = -\log(\mathcal{L}_{ML}) + 0.5K_a\,\log N \tag{8}$$

where the penalty term in MDL takes into account the number of observations. It should be pointed out that when the cluster separability is poor, the performance of these two information theoretic criteria may not be reliable [18], [22].

As discussed above, the SFNM model identification is first performed over $\boldsymbol{x}$-space. However, a mapping from $\boldsymbol{t}$-space to $\boldsymbol{x}$-space may have the intrinsic cluster structure concealed, leading to an incorrect correspondence between (1) and (3). We now extend the mixture representation of (1) to form a hierarchical mixture model generally enough to be applicable to mixtures of any parametric density model. Based on the discussion of a two-level system consisting of a single Radon transform at the top level and a mixture of K_0 normal distributions at the second level, we can reformulate the hierarchy to a third level by associating a group $\mathcal{G}_k$ of SFNM models with each model k in the second level, given by

$$p(\boldsymbol{t}) = \sum_{k=1}^{K_0} \pi_k \sum_{j=1}^{L_{k,0}} \pi_{j|k}\, g(\boldsymbol{t}|\boldsymbol{\theta}_{\boldsymbol{t}(k,j)}) \tag{9}$$

where $\pi_{j|k}$ again correspond to a set of mixing proportions, one for each k, with $\sum_j \pi_{j|k} = 1$. The formation of the hierarchy is guided by the model selection over $\boldsymbol{x}$-subspaces, where each level of the hierarchy corresponds to a generic model, with lower levels giving more focused and interpretable representations. Once again each component in (9) now corresponds to an independent subspace model with Radon transform $g(\boldsymbol{x}|\boldsymbol{\theta}_{\boldsymbol{x}(k,\,j)}) = \int g(\boldsymbol{t}|\boldsymbol{\theta}_{\boldsymbol{t}(k,\,j)})\delta(\boldsymbol{x} - \boldsymbol{W}_{(k,\,j)}^T\boldsymbol{t} + \boldsymbol{W}_{(k,\,j)}^T\boldsymbol{\mu}_{\boldsymbol{t}(k,\,j)})\,d\boldsymbol{t}$.

III. Algorithms

Based on the theory behind hierarchical mixtures of probabilistic principal component subspaces we have discussed above, we now present the description of our algorithm involving major steps of the visual hierarchy construction. Although the tree structure of the hierarchy may be empirically defined [1], [9], a more interesting effort is to build the tree *automatically and interactively*. Guided by the two information theoretic criteria, our algorithm progressively proceeds by fitting a series of submodels to the clusters of the data set, in which model order is selected automatically and algorithm initialization is driven interactively. A schematic summary of the algorithm is as follows.

```
1) Project the data set onto a single
   $\boldsymbol{x}$-space, in which $\boldsymbol{W}$ is determined from
   the sample covariance matrix $\boldsymbol{C_t}$ by fitting
   a single Gaussian model to the data
   set over $\boldsymbol{t}$-space.
2) Learn $f(\boldsymbol{x})$ for $K = K_{\rm MIN}, \cdots, K_{\rm MAX}$, in
   which the values of $\pi_k$ and $\boldsymbol{\theta}_{\boldsymbol{x}k}$ are initialized
   by the user and estimated by
   the EM algorithm over $\boldsymbol{x}$-space.
3) Calculate the values of AIC and MDL for
   $K = K_{\rm MIN}, \cdots, K_{\rm MAX}$, and select a model
   with $K_0$ which corresponds to the minimum
   of AIC and MDL. The model parameters
   obtained in $\boldsymbol{x}$-space will be used to initialize
   the model parameters in $\boldsymbol{t}$-space
   for the learning in step 4.
4) Learn $f(\boldsymbol{t})$ with $K_0$, in which the values
   of $\pi_k$, $z_{ik}$, $\boldsymbol{\mu}_{\boldsymbol{t}k}$, and $\boldsymbol{C}_{\boldsymbol{t}k}$, are fine tuned
   by the EM algorithm over $\boldsymbol{t}$-space.
5) Determine $\boldsymbol{W}_k$ from $\boldsymbol{t}_{ik}$ or $\boldsymbol{C}_{\boldsymbol{t}k}$, and plot
   $\boldsymbol{x}_{ik}$ or $h[\boldsymbol{W}_k^T(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})]$ onto $\boldsymbol{x}$-subspaces at
   the second level for visual evaluation,
   for $k = 1, 2, \cdots, K_0$.
6) Learn $\mathcal{G}_k(\boldsymbol{t})$ by repeating steps 2–4 and
   construct $\boldsymbol{x}$-subspaces at the third level
   by repeating step 5, for $k = 1, 2, \cdots, K_0$.
7) Complete the whole hierarchy under the
   information theoretic criteria, and plot
   all $\boldsymbol{x}$-subspaces for visual exploration
   and explanation.
```

Our algorithm begins by determining $\boldsymbol{W}$ for the top level projection. For low dimensional data sets, we directly evaluate the covariance matrix $\boldsymbol{C_t}$ to find $\boldsymbol{W}$ [10], [12]. For high dimensional cases, since only the top two eigenvectors of the covariance matrix of the data points are of the interest, it may be computationally more efficient to apply our previously developed APEX neural networks [5] to find $\boldsymbol{W}$ directly from the data points $\boldsymbol{t}_i$ (Step 1). On the basis of this single $\boldsymbol{x}$-space, given a fixed K, the user then selects $(K_{\rm MIN},\, K_{\rm MAX})$ and points $\boldsymbol{\mu}_{\boldsymbol{x}k}$ on the plot corresponding to the centers of apparent clusters. The EM algorithm can be applied to allow a SFNM [see (3)] to be fitted to the projected data through the following two-stage [16], [23], [24] form:

E-Step

$$z_{ik}^{(n)} = \frac{\pi_k^{(n)} g\left(\boldsymbol{x}_i|\boldsymbol{\theta}_{\boldsymbol{x}k}^{(n)}\right)}{f\left(\boldsymbol{x}_i|\pi_k^{(n)},\, \boldsymbol{\theta}_{\boldsymbol{x}k}^{(n)}\right)} \tag{10}$$

M-Step

$$\pi_k^{(n+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{ik}^{(n)} \tag{11}$$

$$\boldsymbol{\mu}_{\boldsymbol{x}k}^{(n+1)} = \frac{\sum_{i=1}^{N} z_{ik}^{(n)}\boldsymbol{x}_i}{\sum_{i=1}^{N} z_{ik}^{(n)}} \tag{12}$$

$$\boldsymbol{C}_{\boldsymbol{x}k}^{(n+1)} = \frac{\sum_{i=1}^{N} z_{ik}^{(n)}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}k}^{(n)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}k}^{(n)}\right)^T}{\sum_{i=1}^{N} z_{ik}^{(n)}} \tag{13}$$

where at each complete cycle of the algorithm, we first use an "old" set of parameter values to determine the posterior probabilities $z_{ik}^{(n)}$ using (10). These posterior probabilities are then used to obtain "new" values $\pi_k^{(n+1)}$, $\boldsymbol{\mu}_{\boldsymbol{x}k}^{(n+1)}$, and $\boldsymbol{C}_{\boldsymbol{x}k}^{(n+1)}$ using (11)–(13). The algorithm cycles back and forth until the value of relative entropy [see (2)] reaches its minimum (Step 2). It can be shown that, at each stage of the EM algorithm, the relative entropy decreases unless it is already at a local minimum [16]. The model selection procedure will then determine the optimal number K_0 of models to fit at the next level down using the two information theoretic criteria, where $K_a = 6K_0 - 1$ including $2K_0$ means, $2K_0$ variances, K_0 correlation coefficients, and $K_0 - 1$ mixing factors (Step 3). The resulting points $\boldsymbol{\mu}_{\boldsymbol{t}k}^{(0)}$ in data space, obtained by $\boldsymbol{\mu}_{\boldsymbol{t}k}^{(0)} = \boldsymbol{W}\boldsymbol{\mu}_{\boldsymbol{x}k}^{(\infty)} + \boldsymbol{\mu}_{\boldsymbol{t}}$, are then used as the initial means of the respective submodels. Since the mixing proportions π_k are projection-invariant, we simply assign a 2 × 2 unit matrix to the remaining parameters of the covariance matrix $\boldsymbol{C}_{\boldsymbol{t}k}$. Once again the EM algorithm can be applied to allow a SFNM [see (1)] with K_0 submodels to be fitted to the data over $\boldsymbol{t}$-space. In order to obviate the need to store all the incoming observations, and change the parameters immediately after each data point, it may be computationally more efficient to apply our previously developed probabilistic self-organizing map (PSOM), an incremental EM algorithm [4], to estimate $p(\boldsymbol{t})$.

With a soft partitioning of the data set using the PSOM, data points will now effectively belong to more than one cluster at any given level. Thus, the effective input values are $\boldsymbol{t}_{ik} = z_{ik}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})$ for an independent visualization subspace k in the hierarchy. We then extend our APEX algorithm to a probabilistic version, i.e., PAPEX [5], [25], to determine $\boldsymbol{W}_k$, summarized as follows (Step 4).

1) Initialize the feedforward weight vector $\boldsymbol{w}_{mk}$ for $m = 1, 2$, and the feedback weight vector a_k to small random values at time $i = 1$. Assign a small positive value to the learning rate parameter η.

2) Set $m = 1$, and for $i = 1, 2, \cdots$, compute

$$y_{1k}(i) = \boldsymbol{w}_{1k}^T(i) z_{ik}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k}) \tag{14}$$

$$\boldsymbol{w}_{1k}(i+1) = \boldsymbol{w}_{1k}(i) + \eta[y_{1k}(i) z_{ik}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k}) - y_{1k}^2(i)\boldsymbol{w}_{1k}(i)]. \tag{15}$$

For large i we have $\boldsymbol{w}_{1k}(i) \longrightarrow \boldsymbol{w}_{1k}$, where $\boldsymbol{w}_{1k}$ is the eigenvector associated with the largest eigenvalue of the covariance matrix $\boldsymbol{C}_k$.

3) Set $m = 2$, and for $i = 1, 2, \cdots$, compute

$$y_{2k}(i) = \boldsymbol{w}_{2k}^T(i) z_{ik}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k}) + a_k(i) y_{1k}(i) \tag{16}$$

$$\boldsymbol{w}_{2k}(i+1) = \boldsymbol{w}_{2k}(i) + \eta[y_{2k}(i) z_{ik}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k}) - y_{2k}^2(i)\boldsymbol{w}_{2k}(i)] \tag{17}$$

$$a_k(i+1) = a_k(i) - \eta[y_{2k}(i) y_{1k}(i) + y_{2k}^2(i) a_k(i)]. \tag{18}$$

For large i we have $\boldsymbol{w}_{2k}(i) \longrightarrow \boldsymbol{w}_{2k}$, where $\boldsymbol{w}_{2k}$ is the eigenvector associated with the second largest eigenvalue of the covariance matrix $\boldsymbol{C}_k$.

Having determined principal axes $\boldsymbol{W}_k$ of the mixture model at the second level, we will construct the visualization subspaces by plotting each data point $\boldsymbol{t}_i$ at the corresponding $\boldsymbol{x}_{ik}$. Thus, if one particular point takes most of the contribution for a particular component, then that point will effectively be visible only on the corresponding subspace (Step 5).

Determination of the parameters of the models at the third level can again be viewed as a two-step estimation problem, in which further split of the models at the second level is determined within each of the subspaces over $\boldsymbol{x}$-space, and then the parameters of the selected models are fine tuned over $\boldsymbol{t}$-space. Similarly, the resulting model estimated over $\boldsymbol{x}$-space are then used to initialize the means of the respective submodels over $\boldsymbol{t}$-space. The corresponding $\mathcal{G}_k(\boldsymbol{t})$ can again be estimated using the EM or PSOM algorithm [4], [16], [23] to allow a SFNM distribution with $L_{k,0}$ submodels to be fitted to the data. In the E-step, the posterior probability that data point $\boldsymbol{t}_i$ belongs to submodel j is given by

$$z_{i(k,j)} = z_{ik} z_{ij|k} = z_{ik} \frac{\pi_{j|k} g(\boldsymbol{t}_i | \boldsymbol{\theta}_{\boldsymbol{t}j|k})}{\mathcal{G}(\boldsymbol{t}_i | \boldsymbol{\theta}_{\boldsymbol{t}k})} \tag{19}$$

where z_{ik} are constants estimated from the second level of the hierarchy. The corresponding M-step includes

$$\pi_{j|k} = \frac{\sum_{i=1}^{N} z_{i(k,j)}}{\sum_{i=1}^{N} z_{ik}} \tag{20}$$

$$\boldsymbol{\mu}_{\boldsymbol{t}(k,j)} = \frac{\sum_{i=1}^{N} z_{i(k,j)} \boldsymbol{t}_i}{\sum_{i=1}^{N} z_{i(k,j)}} \tag{21}$$

$$\boldsymbol{C}_{\boldsymbol{t}(k,j)} = \frac{\sum_{i=1}^{N} z_{i(k,j)} (\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}(k,j)})(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}(k,j)})^T}{\sum_{i=1}^{N} z_{i(k,j)}}. \tag{22}$$

With the resulting $z_{i(k,j)}$ in $\boldsymbol{t}$-space, we can apply the PAPEX algorithm to estimate $\boldsymbol{W}_{(k,j)}$, in which the effective input values are expressed by $\boldsymbol{t}_{i(k,j)} = z_{i(k,j)}(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}(k,j)})$. The next level visualization subspace is generated by plotting each data point $\boldsymbol{t}_i$ at the corresponding $\boldsymbol{x}_{i(k,j)} = z_{i(k,j)} \boldsymbol{W}_{(k,j)}^T (\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}(k,j)})$ in (k, j)-subspace (Step 6).

The construction of the entire tree structure hierarchy is automatically completed when no further data split is recommended by the information theoretic criteria in all of the parent subspaces (Step 7).

IV. Illustration and Application

We first illustrate the application of our algorithm to a simple synthetic data set. Fig. 1(a) shows a data set consisting of 450 data points generated from a mixture of three Gaussians in 3-D space. Each Gaussian is relatively flat (has small variance) in one dimension. Two of these pancake-like clusters are closely spaced, while the third is well separated from the first two. The dimensionality of this data set has been chosen to illustrate the basic principle of the approach. The global view of the raw data over $\boldsymbol{t}$-space clearly suggests the presence of three distinct clusters within the data.

To explore the data characteristics, we first perform a single global PCA to project each data point onto a single $\boldsymbol{x}$-space (top level), shown in Fig. 1(b). Both the user inspection and the two information theoretic criteria have clearly suggested the presence of two distinct clusters within the projected data set. Based on a soft clustering of the data points, we then apply

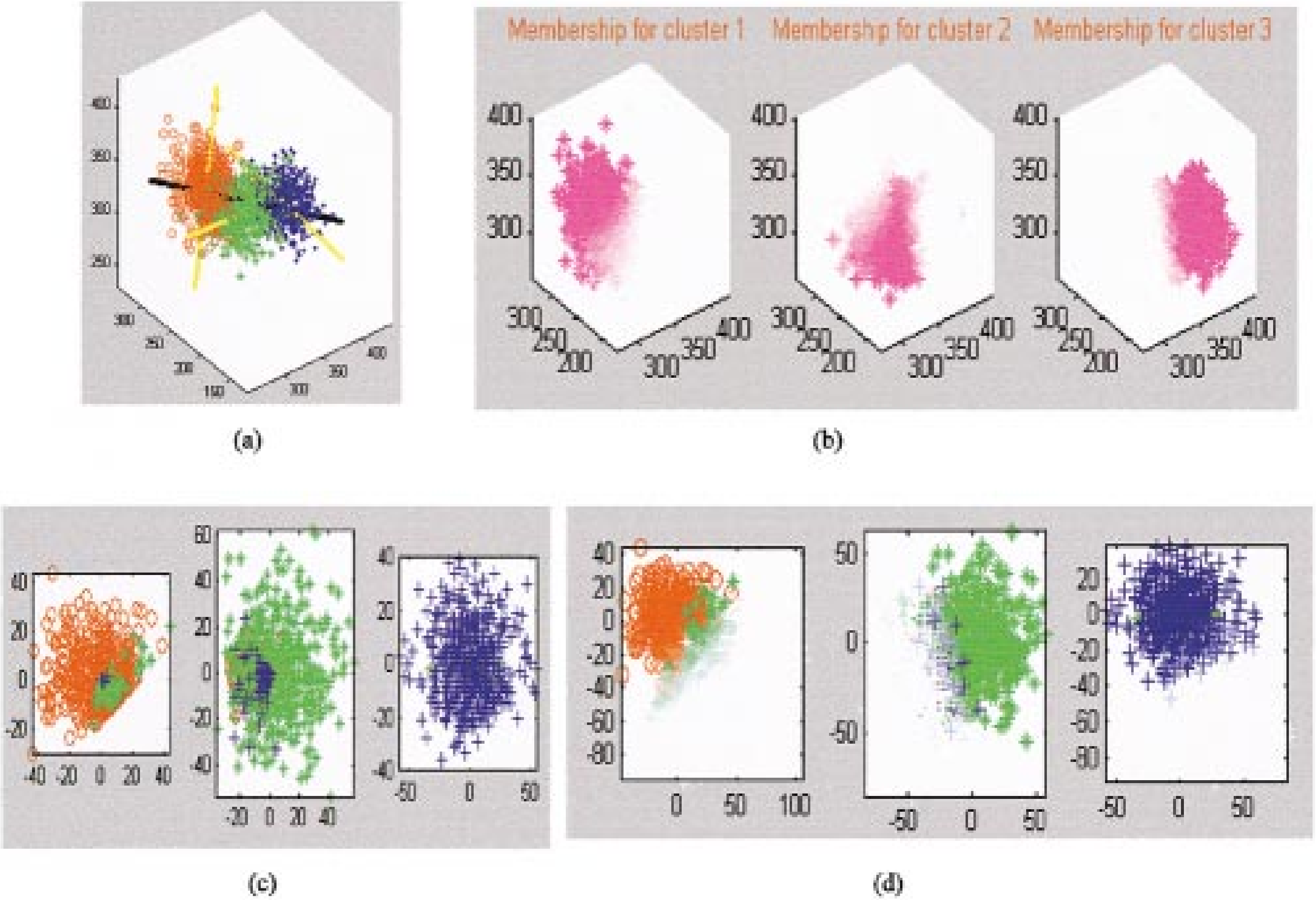


Fig. 2. Summary of final data visualization from a data set with three closely spaced clusters in the 3-D space. Both global and probabilistic principal axes are estimated. Each of the three subspaces is plotted with weighted data inputs. The plots of "data graphics" and "data images" are also generated.

PAPEX to both clusters and generate the two corresponding independent cluster-focused subspaces (second level), as shown in Fig. 1(c). Not to our surprise, the two information theoretic criteria have suggested a further split of cluster 2, but not of cluster 1. Once again by performing three independent PAPEX, the final cluster decomposition through the cluster-focused subspaces (third level) is completed shown in Fig. 1(d).

With this three-level hierarchical data exploration, the capable nature of the approach is evident as the interim two subspaces (second level) only attempt to highlight the data points which have already been modeled by their immediate ancestor (top level). Indeed, the model fitting procedure has successfully discovered all three data clusters. The original data clusters have been individually colored, and it can be seen that the red, yellow, and blue data points have been well separated and highlighted in the third-level subspaces.

As an example of a more complex problem, we consider a data set arising from a mixture of three closely spaced Gaussians consisting of 300 data points, shown in Fig. 2(a). Once again the original data clusters have been individually colored. We first apply APEX to extract the global principal axis, indicated by the black line in Fig. 2(a). The two information theoretic criteria have suggested the presence of three distinct clusters, where the user then selects three initial cluster centers and the EM/PSOM algorithm is applied to perform a soft clustering of the data points. This leads to a mixture of three independent probabilistic principal component subspaces whose principal axes are separately extracted, indicated by the yellow lines in Fig. 2(a). The contributions of each data point to these subspaces, in terms of its "gray-level" $h[\boldsymbol{t}_i] = z_{ik}$, are displayed over $\boldsymbol{t}$-space in Fig. 2(b).

Since the model selection and algorithm initialization are performed over $\boldsymbol{x}$-space with user's interaction, it may be helpful to investigate the visual effectiveness of dimensionality reduction using the probabilistic principal component projections [1], [6]. Based on the estimated $\boldsymbol{W}_k$, we have constructed each of the cluster-focused subspaces using both "data graphics" [e.g., in terms of $\boldsymbol{x}_{ik} = z_{ik}\boldsymbol{W}_k^T(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})$] and "data image" [e.g., in terms of $h[\boldsymbol{W}_k^T(\boldsymbol{t}_i - \boldsymbol{\mu}_{\boldsymbol{t}k})] = z_{ik}$] techniques. As a more overlapped case, Fig. 2(c) and (d) presents the plots of "data graphics" and "data image" from the data set, where "data graphics" emphasizes the contribution of a particular data point to that particular subspace concerning its geometric distance to the center of the cluster, while "data image" emphasizes the effectiveness of a data point reflecting its global appearance. It can be seen that the plot of each cluster is clean and well-shaped.

In order to quantitatively evaluate the effectiveness of our approach with user interactions [6], we apply our algorithm to a synthesized testing data set given in Fig. 3 (upper left). Using

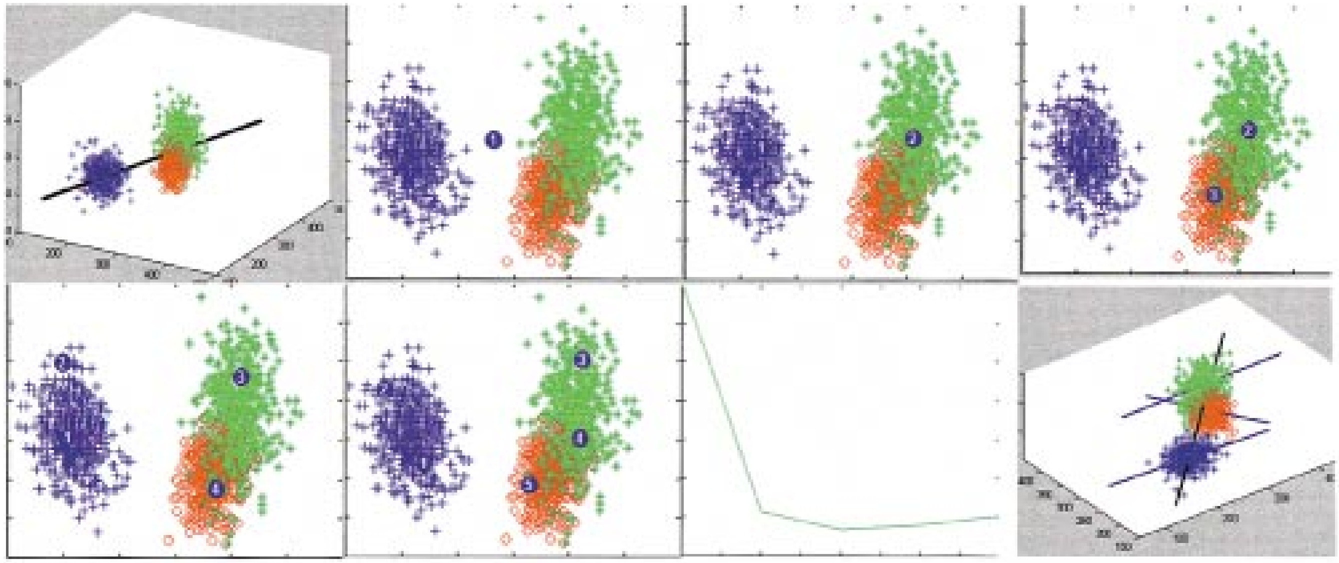


Fig. 3. Summary of a demonstrative interactive user interface for model selection and principal component projections for completing the mixtures of the subspaces.

the APEX algorithm we accurately estimate the top global principal axis, indicated by the back line. By projecting the data points onto a 2-D $\boldsymbol{x}$-space, all three data clusters are visible. This plot indicates that although the second advantage of PCA aforementioned is highly data dependent, when the data clusters are linearly separable in a projection space, the principal component projections allow effective separation of the clusters [13]. We then apply the two information theoretic criteria to examine this plots. In this case, we set $K_{\rm MIN} = 1$ and $K_{\rm MAX} = 5$. The minima of both AIC and MDL have clearly suggested a three-cluster data structure, as given by the curve in Fig. 3 (third block in the second row). Thus, a two-level SFNM model may be sufficient. We then conduct two experiments to assess the performance of our algorithm. Since all the model parameters are known in this case, the true top principal axes of the data clusters have been individually calculated. First, we compare the estimated top principal axes of the data clusters using our algorithm with the corresponding true top principal axes. From the lower right block in Fig. 3, it can be seen that the two sets of the top principal axes are perfectly matched (blue lines). Second, we use the global relative entropy (GRE) between the data histogram and the estimated SFNM model to measure the goodness of model fitting. The numerical result through our experiments indicates a very good performance with a GRE value of 0.008 nats.

User interaction with the algorithm is an important issue. We have developed a user-friendly graphical interface to facilitate the data visualization purpose, as shown in Fig. 3. By allowing the user to select the initial centers of the data clusters demonstrated in Fig. 3, our experience has convincingly indicated a great reduction of both computational complexity and local optimum likelihood. For example, compared to the results of model selection reported by Akaike [18] and Wax [22], the curves of the AIC and MDL generated by our algorithm are much more consistent and smooth, and user-initialized computation is five times (in average) faster than the random trials. It should be pointed out that although the final SFNM model can be estimated, the pathways of achieving cluster decomposition may be multiple. For example, in this case the user has the flexibility to select only two clusters in the second level and to further split the "right" cluster, thus to adopt a three-level hierarchy. We believe that this user-driven nature of the current algorithm is also highly appropriate for the visualization context [1], [11].

Since a more convincing example should involve more clusters with multiple levels, we have also applied our algorithm to the same data set used by Bishop and Tipping [1], shown in Fig. 4(a). This data set arises from a noninvasive monitoring system used to determine the quantity of oil in a multiphase pipeline containing a mixture of oil, water, and gas [1]. The experiment gives 12 diagnostic measurements in total. Our interim goal is to visualize the structure of the data in the original 12-dimensional (12-D) space. A data set consisting of 1000 points is obtained synthetically and the data is expected to have an intrinsic dimensionality of two corresponding to the two dominant components (e.g., oil and water). However, the presence of different flow configurations leads to numerous distinct clusters. We then apply our algorithm to perform a cluster discovery. Results from partially fitting the oil flow data using a three-level hierarchical model are given in Fig. 4. It should be pointed out that since the "right" answer to this real-world data set is not available, we are not able to validate this new result. However, we believe that this example has clearly been highly successful, note how the selected single cluster (number 2) in the upper level plot, is discovered to be two quite separated clusters at the second level.

As a final example, we consider the visual explanation in computer-aided diagnosis (CAD) for breast cancer detection. As a step toward improving the performance of CAD system, we have put considerable efforts to conduct various studies and develop reliable image enhancement and lesion segmentation

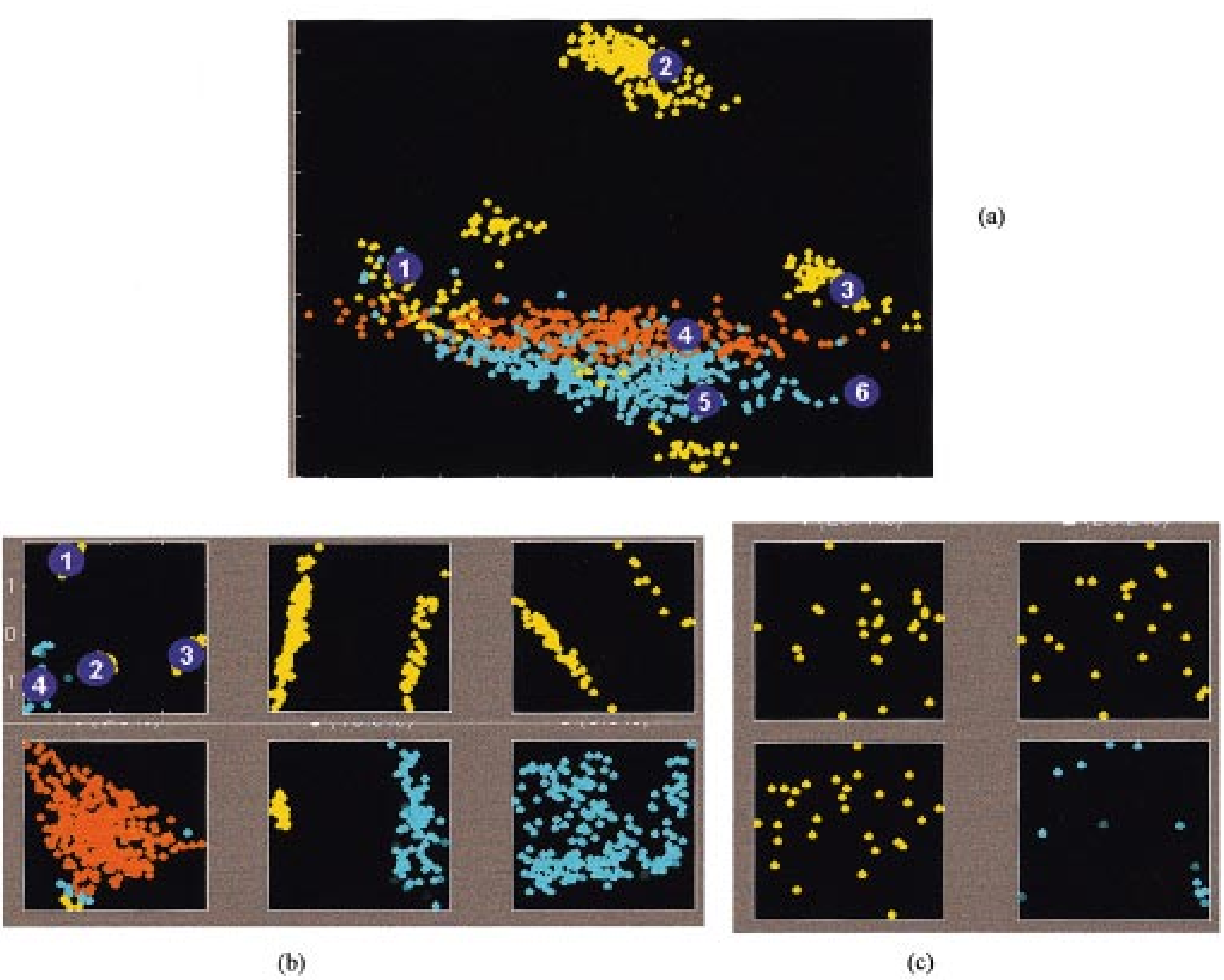
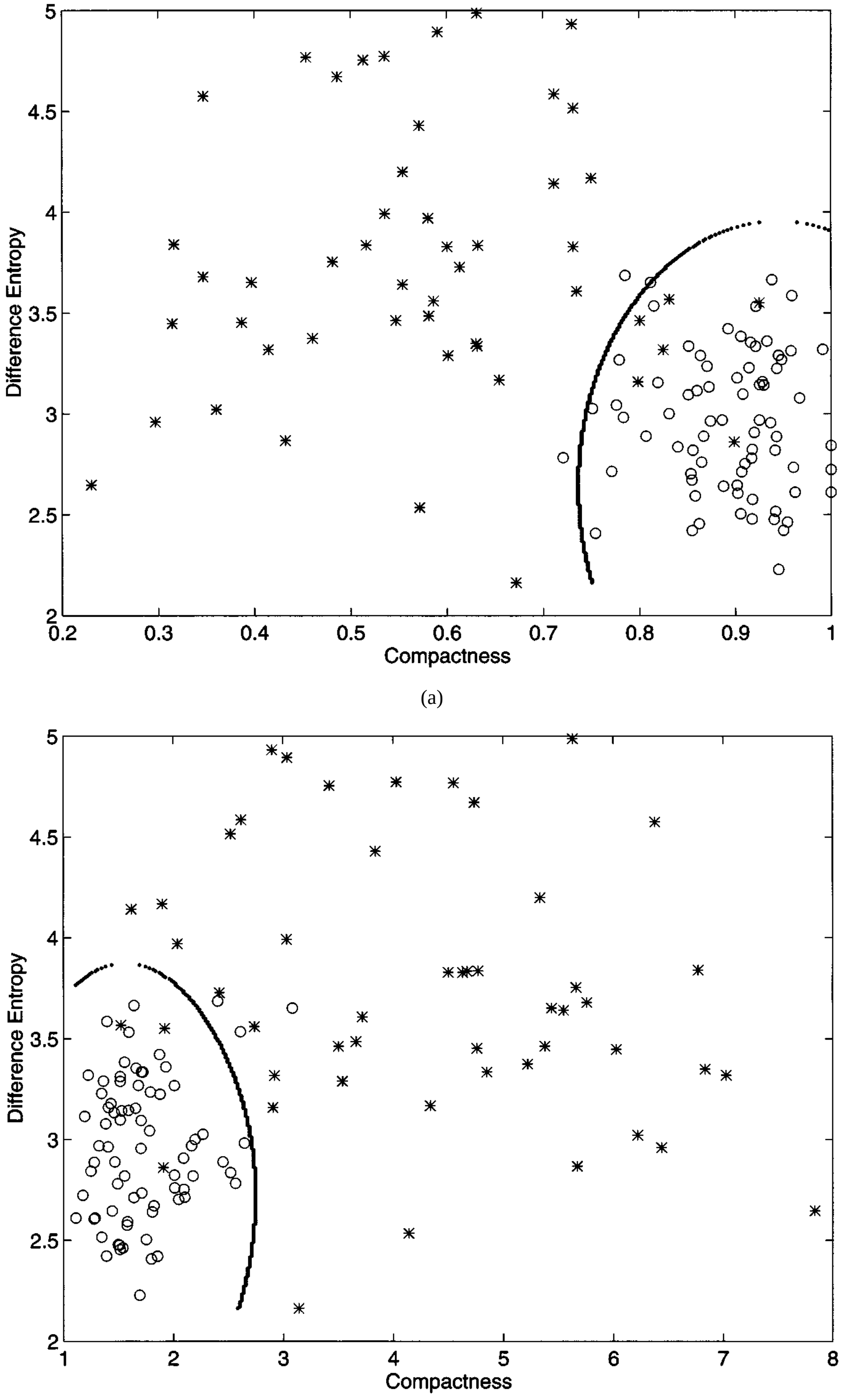


Fig. 4. Summary of the interim results from the oil flow data set. The data set is living in a 12-D space with numerous local clusters. A multiple level model is adopted with a hierarchical visual exploration for cluster discovery.

techniques [4]. More precisely, we try to make both the hidden data patterns and the neural network “black box” to be as transparent as possible to the user (e.g., radiologists and patients) through interactive visual explanation. The clinical goal is to eliminate the false positive sites that correspond to normal dense tissues with mass-like appearances through featured discrimination. We adopt a mathematical feature extraction procedure to construct our database from all the suspicious mass sites localized by the enhanced segmentation [4]. The optimal mapping of the data points is then obtained by learning the generalized normal mixtures and decision boundaries, where a probabilistic modular neural network is developed to carry out both soft and hard clustering [4]. The joint histogram of the featured database extracted from true and false mass regions are investigated and the features that can better separate the true and false mass sites are selected [4]. Our experience has suggested that three imagery features, i.e., site area, compactness, and difference entropy, were having good discrimination and reliability properties.

We then use our previously developed algorithm [4] to distinguish the true masses from false masses based on the features extracted from the suspected regions. 150 mammograms were selected from the mammogram database. Each mammogram contained at least one mass case of varying size and location. The areas of suspicious masses were identified following the proposed procedure with biopsy proven results. In a typical experiment, we have selected a 3-D feature space consisting of compactness I, compactness II, and difference entropy. It should be noticed that the feature vector can easily extend to higher dimensionality. A training feature vector set was constructed from 50 true mass ROI’s and 50 false mass ROI’s, where ROI stands for *region of the interest*. In addition to the decision boundaries recommended by the computer algorithms, a visual explanation interface has also been integrated with hierarchical projections. Fig. 5(a) shows the database map selection with compactness definition I and difference entropy. Fig. 5(b) shows the database map selection with compactness definition II and difference entropy. Our experience has suggested that the recognition rate



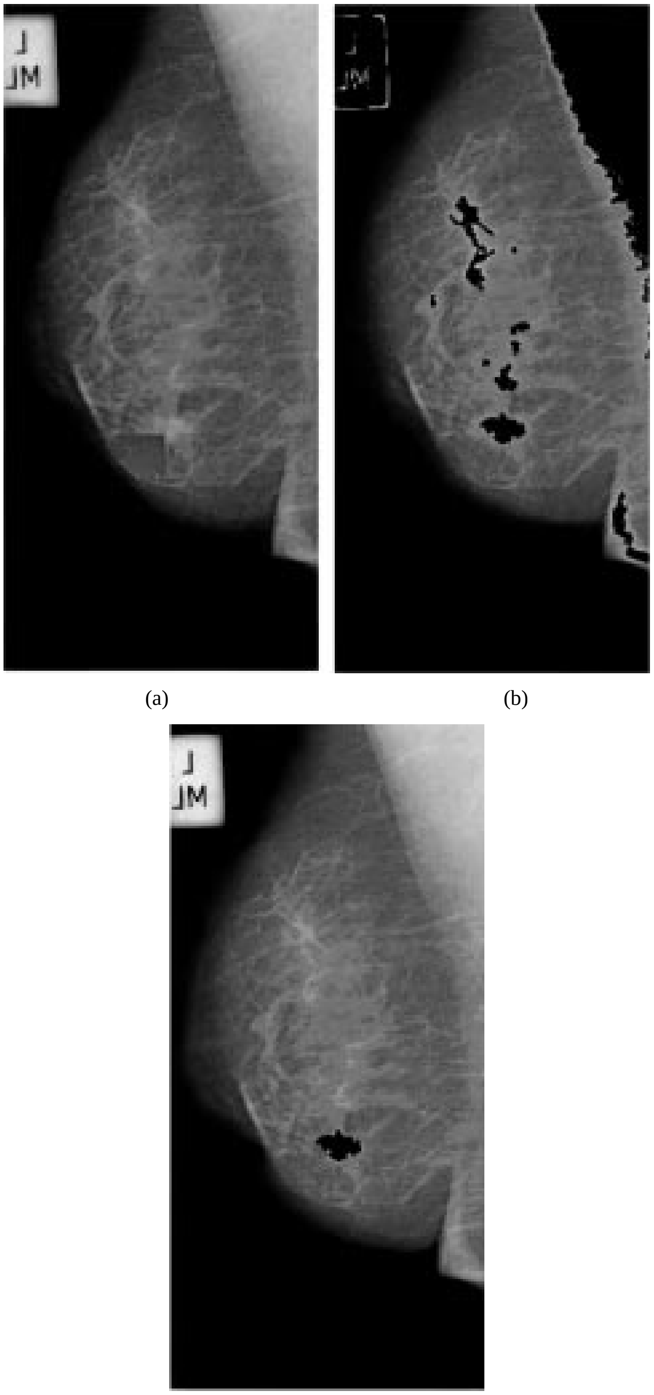
(a)

(b)

Fig. 5. Selected feature maps of the database in computer-aided diagnosis for breast cancer detection.

with compactness I are more reliable than that with compactness II.

We have conducted a preliminary study to evaluate the performance of the algorithms in real case detection, in which 6–15 suspected masses per mammogram were detected and required further clinical decision making. We found that the proposed visual explanation approach, together with CAD system, can reduce the number of suspicious masses with a sensitivity of 84% at a specificity of 82% (1.6 false positive findings per mammogram) based on the database containing 46 mammograms (23 of them have biopsy proven masses). Fig. 6 shows a representative mass detection result on one mammogram with a stellate mass, indicated by the arrow in Fig. 6(a). After appropriate feature extraction, ten sites with brightest intensity were selected, shown in Fig. 6(b). The featured vectors of these candidates were submitted against the estimated "probability cloud" for visual explanation as a decision support, together with the opinion recommended by our CAD system. The final results indicated that



(a) (b)

(c)

Fig. 6. Summary of the results of mass detection from a typical clinical case.

the stellate mass lesion was correctly detected and confirmed by our experienced radiologists, shown in Fig. 6(c). It should be pointed out that in this real-world application, a higher recognition rate may be controlled by the domain experts in balancing the tradeoff between the *false positive* and *false negative* rates [4].

V. Discussion

We have presented a novel approach to visual explanation for data mining and knowledge discovery, which is both statistically principled and visually effective. This method, as illustrated by the well-planned simulations and pilot applications in computer-aided diagnosis, can be very capable of revealing hidden structure within data. It is important to emphasize that in relation to previous work [1], [8]–[10], one interesting consideration with the present algorithm is that the models are determined by the information theoretic criteria, and this criterion will not only select the most appropriate model structure, but also allow a user-driven portfolio as a double check. This approach promotes a self-consistent fitting of the whole tree, so that an automated procedure for generating the hierarchy becomes reality [1]. In addition, since we perform model selection and parameter initialization first over the projection space, the computational complexity is greatly reduced in compared to the maximum likelihood estimation in full dimension. Our case study of a seven-dimensional (7-D) data set has indicated at least a 50% reduction of the computational time. Other possible advantages include the determination of data projection by maximum the separation of clusters which in turn optimizes the other crucial operations such as model selection and parameter initialization [13] and data rendering algorithms which permit user or hypothesis driven nature of the data projection [11].

Another important consideration with the present approach is the measure of quality in visual explanation [3]. This is not a glamorous area, but progress in this area is eminently critical to the future success of visual exploration [6]. What is the correct matrix for a direct projection of a particular multimodal data set? How effective was a particular visualization tool? Did the user come to the correct conclusion? It may be agreeable that the benchmark criteria in visual exploration are very different and difficult [6]. As shared by Bishop and Tipping [1], we believe that in data visualization there is no objective measure of quality, and so it is difficult to quantify the merit of a particular data visualization technique, and the effectiveness of such a techniques is often highly data dependent. The possible alternative is to perform a rigorous psychological evaluation using a simple and controlled environment, or to invite domain experts to directly evaluate the efficacy of the algorithm for a specified task. For example, we can compare the domain expert's performances with and without the system aid. In that case, the receiver operating characteristic (ROC) method may be used to evaluate the performance of our algorithm when used by the radiologists. While the optimality of these new techniques is often highly data dependent, we would expect the hierarchical visualization model to be a very effective tool for the data visualization and exploration in many applications. We are currently investigating further applications to the molecular classification of cancer based on gene array data sets.

Acknowledgment

The authors would like to thank H. Li of the University of Maryland at College Park and T. Adali of the University of Maryland at Baltimore County for their valuable scientific input to this work.

References

[1] C. M. Bishop and M. E. Tipping, “A hierarchical latent variable model for data visualization,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 282–293, Mar. 1998.

[2] T. R. Golub *et al.*, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, pp. 531–537, Oct. 1999.

[3] E. R. Tufte, *Visual Explanation: Images and Quantities, Evidence and Narrative*. Cheshire, U.K.: Graphics, 1996.

[4] Y. Wang, S. H. Lin, H. Li, and S. Y. Kung, “Data mapping by probabilistic modular networks and information theoretic criteria,” *IEEE Trans. Signal Processing*, vol. 46, pp. 3378–3397, Dec. 1998.

[5] S. Y. Kung, *Principal Component Neural Networks*. New York: Wiley, 1996.

[6] G. M. Nielson, “Challenges in visualization research,” *IEEE Trans. Vis. Comput. Graph.*, vol. 2, pp. 97–99, June 1996.

[7] M. I. Jordan and R. A. Jacobs, “Hierarchical mixture of experts and the EM algorithm,” *Neural Computat.*, vol. 6, pp. 181–214, 1994.

[8] N. Kambhatla and T. K. Leen, “Dimension reduction by local principal component analysis,” *Neural Computat.*, vol. 9, no. 7, pp. 1493–1516, 1997.

[9] M. E. Tipping and C. M. Bishop, “Mixtures of probabilistic principal component analyzers,” *Neural Computat.*, vol. 11, pp. 443–482, 1999.

[10] G. E. Hinton, P. Dayan, and M. Revow, “Modeling the manifolds of images of handwritten digits,” *IEEE Trans. Neural Networks*, vol. 8, pp. 65–74, Jan. 1997.

[11] L. Luo, Y. Wang, and S. Y. Kung, “Hierarchy of probabilistic principal component subspaces for data mining,” in *Proc. IEEE Workshop Neural Nets for Signal Processing*, Aug. 1999.

[12] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[13] K. Etemad and R. Chellappa, “Separability-based multiscale basis selection and feature extraction for signal and image classification,” *IEEE Trans. Image Processing*, vol. 7, Oct. 1998.

[14] R. Gray and L. Davisson, *Random Processes—A Mathematical Approach for Engineers*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[15] R. N. Bracewell, *Two-Dimensional Imaging*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[16] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[18] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Automat. Contr.*, vol. AC-19, no. 6, pp. 716–723, 1974.

[19] J. Rissanen, “Modeling by shortest data description,” *Automat.*, vol. 14, pp. 465–471, 1978.

[20] E. T. Jaynes, “Information theory and statistical mechanics,” *Phys. Rev.*, vol. 106, no. 4, p. 620–630/171–190, May 1957.

[21] J. Rissanen, “Minimax entropy estimation of models for vector processes,” in *System Identification: Advances and Case Studies*. New York: Academic, 1987, pp. 97–117.

[22] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” *IEEE Trans. Acoust. Speech, Signal Processing*, Apr. 1985.

[23] L. I. Perlovsky and M. M. McManus, “Maximum likelihood neural networks for sensor fusion and adaptive classification,” *Neural Networks*, vol. 4, pp. 89–102, 1991.

[24] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny, “Bayesian approaches to Gaussian mixture modeling,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1133–1141, Nov. 1998.

[25] S. Y. Kung, K. I. Diamantaras, and J. S. Taur, “Adaptive principal component extraction (APEX) and applications,” *IEEE Trans. Signal Processing*, vol. 42, pp. 1202–1217, May 1994.

Yue Wang received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 1995.

Since 1996, he has been an Assistant Professor of electrical engineering at The Catholic University of America, Washington, DC. He is also affiliated with Georgetown University and Johns Hopkins University as an Adjunct Assistant Professor of radiology and bioinformatics. His research interests include intelligent bioinformatics, information visualization, adaptive neural computing, biomedical imaging, and image/video analysis and understanding.

Dr. Wang is the recipient of a 1998 Department of Defense (CDMRP) Career Development Award.

Lan Luo received the B.S. and M.S. degrees in biomedical engineering from Chongqing University, China, in 1993 and 1996, respectively. She is currently working towards the Ph.D. degree at North Carolina State University, Raleigh.

She performed research at the Imaging and Intelligent Informatics Laboratory for one year. Her interests include computer networks, database management, and medical imaging.

Matthew T. Freedman received the A.B. degree in general science in 1963 from the University of Rochester, Rochester, NY, and the Ph.D. degree in medicine in 1967 from the State University of New York, Brooklyn.

Currently, he is an Associate Professor of Radiology at the Georgetown University Medical Center, Washington, DC. He is a General Radiologist with clinical and research activities in chest, musculoskletal, and breast imaging. He is Director of Mammography Research and Clinical Director of the Center for Imaging Science and Information Systems.

Sun-Yuan Kung (F'88) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

From 1977 to 1987, he was a Professor of electrical engineering—systems at the University of Southern California, Los Angeles. Since 1987, he has been a Professor of electrical engineering at Princeton University, Princeton, NJ. He has authored more than 300 technical publications, including three books *VLSI Array Processors* (Englewood Cliffs, NJ: Prentice-Hall, 1988) (with Russian and Chinese translations), *Digital Neural Networks* (Englewood Cliffs, NJ: Prentice-Hall, 1993), and *Principal Component Neural Networks* (New York: Wiley, 1996).

Dr. Kung received the 1992 IEEE Signal Processing Society's Technical Achievement Award for his contributions on “parallel processing and neural network algorithms for signal processing.” He was appointed as an IEEE-SP Distinguished Lecturer in 1994. He received the 1996 IEEE Signal Processing Society's Best Paper Award. Since 1990, he has served as an Editor-In-Chief of the *Journal of VLSI Signal Processing*.