A General Probabilistic Framework for Clustering Individuals

Technical Report No. 00–09
Department of Information and Computer Science
University of California, Irvine

Igor Cadez, Scott Gaffney and Padhraic Smyth
Department of Information and Computer Science
University of California, Irvine
CA 92697-3425
[icadez,sgaffney,smyth]@ics.uci.edu

March 2000

Abstract

This paper presents a unifying probabilistic framework for clustering individuals or systems into groups when the available data measurements are not multivariate vectors of fixed dimensionality. For example, one might have data from a set of medical patients, where for each patient one has different numbers of time-series observations, each time-series of different lengths. We propose a general model-based framework for clustering heterogeneous data types of this form. We discuss a general Expectation-Maximization (EM) procedure for clustering within this framework and outline how it can be applied to clustering of sequences, time-series, histograms, trajectories, and other non-vector data. We show that a number of earlier algorithms can be viewed as special cases within this unifying framework. The paper concludes with several illustrations of the method, including clustering of two-dimensional histograms of red blood cell data in a medical diagnosis context, clustering of proteins from curves of gene expression data, and clustering of individuals based on their sequences of Web navigation.

Table 1: An example of individuals characterized by Web navigation sequences.

User 1	Session 1	2	3	2	2	3	3	3	1	1	1	3	1	3	3	3	3	1	2	3
	Session 2	3	3	3	1	1	1													
User 2	Session 1	7	7	7	7	7	7	7	7											
User 3	Session 1	1	5	1	1	5	1	5	1	1	1	1	1	1	1	1	1	1	5	
	Session 1	5	1	1	5															
	Session 3	1	3	3	1	5	1	1	1	1										

1 Introduction

Clustering is a fundamental and widely applied methodology in understanding and exploring large data sets. Clustering algorithms are typically applied to vector measurements of fixed dimension. For example, we may have a d measurements on a set of medical patients and we represent the measurements on individual i as a d-dimensional vector. For such data there are a wealth of different clustering methods available, both model-based methods (e.g., Fraley and Raftery, 1998) and distance-based methods (e.g., Jain and Dubes, 1986).

In this paper we are interested in the problem of clustering individuals given observed data about the individuals, where the observed data does not naturally occur in vector form (we elaborate on this below). We use the word "individuals" in a broad sense; it can encompass humans, animals, organisms, organizations, natural phenomena, mechanical systems, and so forth. Specific examples include clustering individuals based on their observed Web browsing behavior (Cadez et al., 2000), clustering animals given behavioral observations (Haccou and Meelis, 1992), clustering extra-tropical cyclones based on their temporal evolution (Blender et al., 1997), and clustering genes from expression data (Eisen et al., 1998). Clearly there is a notion of each individual in a sense representing a *dynamic system* and we wish to cluster these systems based on observations of their dynamic behavior.

Standard vector-based clustering techniques are not directly applicable to this type of clustering problem since the data arises in non-vector form (e.g., sequences, time-series, trajectories, etc) and we can have different amounts of observed data for different individuals. Table 1 shows an example of such data. The data represent categorized page requests from traces of Web navigation for different individuals. Each individual is characterized by different sessions (different sequences) and each of these sequences vary in length. We would like to be able to cluster individuals into groups based on their observed navigation behavior—but it is not at all obvious how one would go about this in a principled manner. Our goal in this paper is to formulate a general model-based clustering framework for clustering individuals when the data are non-homogeneous as in this example.

2 Background and Related Work

One approach to clustering in this context is the "feature-vector" approach, namely, to reduce the observed data to feature vectors of fixed dimensionality and use standard multivariate clustering techniques to cluster individuals given this representation. For example, Blender et al (1997) represent extra-tropical cyclones as a concatenation of 3 days worth of (x, y) latitude-longitude pairs, spaced at 6-hour intervals, to yield a 24-dimensional representation of each cyclone. The k-means clustering algorithm was then applied in this

24-dimensional space to find clusters of cyclones. Similarly, Eisen et al. (1999) represent gene expression data from different time-course experiments as fixed-dimensional vectors. Agglomerative hierarchical clustering is then applied to find clusters of genes using a vector distance measure. Although interesting scientific insights were produced in both of these cases, we argue that this "vector methodology" is not necessarily appropriate or adequate for clustering dynamic behavior. In particular, for sequential or temporal data, the conversion to vector form necessarily incurs a loss of information (e.g., in the examples above the temporal evolution of cyclones and genes (respectively) is not explicitly retained in the vector representation).

Another general approach to this problem is to define pairwise distances between all individuals in some manner (e.g., edit-distance for sequences) and then use distance-based clustering methods (e.g., hierarchical). Difficulties here arise with defining effective distance measures for complex problems. For example, if different individuals have different amounts of data (e.g., varying numbers of sequences per individual) there may be no principled way to directly account for this multiplicity of information via a distance function.

We propose instead a general probabilistic methodology for handling these issues, based on the framework of *generative mixture models*. This probabilistic framework is particularly useful for problems of this type since it allows us to directly address the two problems mentioned above of (1) modeling non-vector data in its "native" form, and (2) handling multiplicities of data sizes and data types across individuals.

Special cases of the model-based probabilistic framework presented in this paper have been developed earlier for specific classes of data types and cluster models. In particular, the concept of using a generative model for clustering non-vector data has been independently pursued in several different contexts. For example, in clustering sequences, Poulsen (1990) introduced a particular form of Markov mixtures and an EM algorithm for modeling heterogeneous behavior in consumer purchasing data. More general versions of Markov mixtures were subequently independently developed by both Smyth (1997, 1999) and Ridgeway (1997), including a general EM framework for learning in this context. Clustering of regression curves has been investigated by Spath (1979), DeSarbo, Oliver, and Rangaswamy (1989), Wedel and Steenkamp (1991), and in a more general non-parametric form by Gaffney and Smyth (1999).

The work presented here represents a generalization of all of the above ideas within a single unified framework. Our work is more general in the sense that we explicitly discuss the case when different individuals have different amounts of data, which is an important factor in many practical applications. We note that once we define a proper likelihood over the data of interest, then specification of any specific EM algorithm (for any particular type of model) follows in a direct manner from the general principles of EM (e.g., Dempster, Laird, and Rubin, 1977; McLachlan and Krishnan, 1997). In fact this EM algorithm is quite similar to the "standard" EM algorithm, with the exception that different individuals have different effects on the estimation process depending on whether they have more or less observations (this is not surprising: see Section 4.3). Thus, our EM framework will be rather straightforward to any readers familiar with EM in general—for these readers we wish to emphasize the generality of the approach as evidenced by the diverse applications (Section 5). For readers less familiar with model-based clustering and EM in general, the goal of the paper is to demonstrate that a large number of non-trivial clustering problems can be elegantly handled within this model-based EM framework.

3 A Generative Mixture-Based Cluster Model

We propose the following generative framework for model-based clustering:

- An individual is randomly drawn from the overall population (universe) and is indexed by the letter i (the index represents the "individual").

- The individual is assigned to one of K clusters, $1 \leq k \leq K$, with probability $p(c_i = k)$, $\sum_k p(c_i = k) = 1$, $1 \leq k \leq K$, where we use c_i to indicate the cluster membership of individual i.

- Each cluster k, $1 \leq k \leq K$, has a data generating model $p_k(D|\Theta_k)$, where Θ_k are the parameters of the probability distribution p_k.

- Data D_i is now generated for an individual i by a data generating probability model $p(D_i|\Theta_k)$, once the cluster membership $c_i = k$ for individual i is known, and given Θ_k. D_i represents the heterogeneous data associated with individual i and can be quite general in form, e.g., sequences, images, histograms, text, vectors, or combinations of any of these, as long as we can define an appropriate density function p on such data.

As a simple (but concrete) example consider the case where for each individual we have a set of discrete-valued sequences, i.e., $D_i = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{n_i}\}$, where each sequence $\mathbf{s}$ can be of a different length. For example, each sequence could represent the observed record of page requests for individual i at a particular Web site, and the different sequences represent different sessions for that individual. We can use the framework above to model the overall data-generating process as follows:

- The population of Web users is divided into K groups or clusters, and a randomly chosen user i has probability $p(c_i = k)$ of belonging to cluster k, $\sum_k p(c_i = k) = 1$.

- The behavior of each cluster is governed by a finite-state Markov model (a stochastic finite-state machine) with parameters Φ_k (the initial state probabilities and transition matrix for that cluster). This Markov model provides a probabilistic generative model for sequences from that group. If the model has an absorbing state (an "end" state) then sequences drawn from this model may have different lengths, with a length distribution depending on Φ_k. The probability of a particular sequence $\mathbf{s}$ under the Markov model for cluster k is $p_k(\mathbf{s}|\Phi_k)$.

- There is a generative mechanism for session initiation, e.g., a distribution on the number n_i of sessions for individual i, for each group; e.g., a Geometric model with parameter λ_k, and distribution $q_k(n_i|\lambda_k)$.

- We could also couple the dynamic characteristics of the individual (as observed via the sequences $\mathbf{s}$) to other factors (covariates) such as demographic variables like age, income, etc. (We do not explore this explicitly in this paper, but see Smyth (1999) for how this type of coupling of static and dynamic behavior can be modeled and learned).

- Thus, in our model so far, the overall parameters for each cluster consist of $\Theta_k = (\Phi_k, \lambda_k)$.

- The probability of a set of sequences from individual i, $D_i = \{\mathbf{s}_1, \ldots, \mathbf{s}_{n_i}\}$, conditioned on assuming that i is a member of cluster k, can then be written as

$$\begin{aligned} p(D_i|c_i = k, \Theta_k) &= p(n_i, \mathbf{s}_1, \ldots, \mathbf{s}_{n_i}|k, \Theta_k) &(1)\\ &= q_k(n_i; \lambda_k) \prod_{j=1}^{n_i} p_k(\mathbf{s}_j|\Theta_k) \quad 1 \leq k \leq K &(2)\end{aligned}$$

  where we assume for simplicity that the sequences are conditionally independent given the model.

- The probability of a set of sequences D_i for individual i, whose cluster membership is unknown, can be written as:

$$p(D_i|\Theta) = \sum_{k=1}^{K} p(D_i|c_i = k, \Theta_k)p(c_i = k), \qquad 1 \leq k \leq K \tag{3}$$

  i.e., as a *mixture* of data-generating processes, where the probability model for each component is specified by Equation 2.

- Finally, we can compute the probability that individual i belongs to cluster k by Bayes rule, i.e.,

$$p(c_i = k|D_i, \Theta) \propto p(D_i|c_i = k, \Theta_k)p(c_i = k). \tag{4}$$

The model above allows us to model heterogeneous data across different individuals in a fairly general framework. Using the model, our goal will be to estimate the Θ_k (cluster parameters) and cluster weights $p(c_k)$ given only observed data $D = \{D_1, \ldots, D_N\}$.

As we will see later in the paper, this probabilistic framework has some distinct advantages over alternative methods for clustering individuals. For example, clustering sequences of different lengths is not problematic. The key idea is that objects are defined to be similar in terms of common similarity to a model, expressed through the likelihood function $p(D_i|\Theta_k)$. Two objects are considered similar if they both have higher likelihood under one particular model than under any other model. For example, consider two clusters, each modeled by a Markov model, one favoring short runs of a's and b's, the other favoring longer runs. In this context, a short sequence *abaabab* and a longer sequence *baababbabbabaabaab* could both be considered similar relative to the short-run model, i.e., given these two models there is a much higher probability that they both originated from the short-run model than the long-run one.

We will show that we can learn the cluster model parameters Θ from a data set $D = \{D_1, \ldots, D_N\}$ describing N individuals, using an algorithm based on the EM procedure. We will further show that this algorithm generalizes the standard "vector-based" EM algorithm for mixture models in two directions that are entirely intuitive: (1) membership probabilities are associated with individuals rather than individual measurements, and (2) individuals with more (or less) data have more (or less) influence on the parameter estimation process.

4 A General EM-based Clustering Algorithm for Clustering Individuals

4.1 Statement of a General Algorithm

In this section we provide a general description of an EM algorithm that can be applied to mixture-model clustering of individuals. The model that we consider here is actually a special case of a Bayesian hierarchical framework for this class of clustering problems, but due to space limitations we do not pursue this Bayesian viewpoint further in this paper (see Cadez and Smyth, 1999, for more details).

We use notation similar to that introduced in the previous section. Specifically, let there be N individuals and let there be a data set D_i associated with each individual (note that we sometimes refer to the data set D_i itself as an *individual*). Each data set D_i consists of n_i observations $d_{ij}, 1 \leq j \leq n_i$, where each "observation" represents another smaller data subset. The data set for all individuals is denoted $D = \{D_1, D_2, \ldots, D_N\}$, with individual data $D_i = \{d_{i1}, d_{i2}, \ldots, d_{in_i}\}$. According to the generative cluster model each individual is assigned to a single cluster c_i $(1 \leq c_i \leq K)$ and has an associated probability density function of:

$$p(D_i|c_i, \Theta) = p(D_i|\Theta_{c_i}),$$

where Θ represents parameters for all the clusters: $\Theta = \{\Theta_1, \Theta_2, \ldots, \Theta_K\}$ and where c_i, $1 \leq c_i \leq K$ is the cluster identity of the ith individual.

We further assume that the observations are conditionally independent given the model parameters, so that we can write the probability of the individual (or an individual's data), given that the individual belongs to the cluster c_i, as:

$$p(D_i|c_i, \Theta) = \prod_{j=1}^{n_i} p(d_{ij}|\Theta_{c_i}). \tag{5}$$

This is equivalent to assuming in a Web-browsing scenario (for example) that an individual's Web navigation patterns in any one session are independent of their patterns from previous sessions, given the overall parameters governing that individual. While this is an approximation to what is really going on (there may in fact be some sequential session-to-session effects) we conjecture that it will be a reasonably accurate and useful assumption in practice.

We are interested in learning the maximum-likelihood (ML) or maximum a posteriori (MAP) parameter estimates given the data D, i.e., $\Theta_{\text{ML}} = \arg\max_\Theta\{p(D|\Theta)\}$, and $\Theta_{\text{MAP}} = \arg\max_\Theta\{p(D|\Theta)p(\Theta)\}$, where under the usual assumption that data from different individuals are conditionally independent given the underlying model we have

$$p(D|\Theta) = \prod_{i=1}^{N} p(D_i|\Theta) \tag{6}$$

known as the likelihood.

The EM algorithm is a general technique for finding ML or MAP parameters Θ when some aspect of the data is considered "missing." In a mixture context, the missing data consists of the cluster labels c_i for each individual: if we knew these labels then parameter estimation would be quite straightforward. The EM algorithm can be viewed as operating in two steps. In the E step one calculates class-conditional probabilities $p(c_i|D_i, \Theta)$ for each individual under each of the K cluster models using the current value of the parameters Θ. In the M step one updates parameters Θ by weighting each individual according to their class-conditional probability. This yields a very intuitive algorithm that is guaranteed to lead to a sequence of Θ's which have non-decreasing likelihood or posterior probability, i.e., under fairly broad conditions it will find at least a local maximum of the ML or MAP objective function.

In the next section we illustrate how this algorithm can be applied to the case when we have different amounts of data for different individuals, a generalization of the standard EM framework.

4.2 A Specific Illustration of this EM-Algorithm Framework

As a specific example of this general framework, we revisit in more detail the problem discussed in Section 3 of clustering discrete-valued sequences, taking values from 1 to M. We will assume that we are using a mixture of Markov chains model. Each individual i can have n_i sequences (e.g., different observed sessions of Web navigation behavior). The generative model has the general mixture form described in Section 4, where the component model for each observation in each cluster, $p(d_{ij}|\Theta_k)$ in Equation 5, takes the form of a Markov model $p(\mathbf{s}|\Theta_k)$. We could of course use any sequential model that defines a density on possible sequences, but choose the Markov model for simplicity of illustration.

The model parameters for each Markov cluster Θ_k consist of an initial state probability vector $\pi_k(s)$ and an $M \times M$ transition matrix $T_k(s_2|s_1)$, where s, s_1, s_2 denote discrete states, $1 \leq s, s_1, s_2 \leq M$. In addition, there is a set of weights α_k that defines the mixing proportions of the component models (previously denoted as $p(k)$). The intuition is that different clusters will have different Markov behavior and we wish to learn these different behaviors from observed sequences.

We follow the general approach outlined in Section 3 to define a likelihood and in Section 4 to derive an associated EM algorithm. An important point is that this procedure is quite general—one simply encodes one's assumptions about the generative nature of the model via the likelihood, and the associated EM algorithm follows directly.

Let $D_i = \{\mathbf{s}_{i,1}, \ldots, \mathbf{s}_{i,j}, \ldots, \mathbf{s}_{i,n_i}\}$ be the data for the ith individual, where $\mathbf{s}_{i,j}$ is the jth sequence observed for this individual (from before, $\mathbf{s}_{i,j}$ corresponds to d_{ij}, the jth subset of data (or observation) for an individual i). From the definition of a Markov chain we can define the likelihood of any particular sequence $\mathbf{s}_{i,j}$, conditioned on a particular cluster c_i with parameters Θ_{c_i} as

$$p(\mathbf{s}_{i,j}|c_i = k, \Theta_{c_i}) = \pi(s_{i,j,1}) \prod_{l=1}^{L_{i,j}-1} T_k(s_{i,j,l+1}|s_{i,j,l}) \quad 1 \leq k \leq K \tag{7}$$

where $s_{i,j,l}$ denotes the lth element of the jth sequence for individual i, and $L_{i,j}$ is the length of the jth sequence for individual i.

Thus, the probability of all of the data D_i from individual i, conditioned on cluster c_i, can be written as:

$$p(D_i|c_i = k, \Theta_{c_i}) = \prod_{j=1}^{n_i} p(\mathbf{s}_{i,j}|c_i, \Theta_{c_i}) \tag{8}$$

in accordance with Equation 5. Here we do not explicitly model the distribution on n_i, i.e., a distribution modeling how many sequences are produced by members of cluster c_i. In effect this is equivalent to assuming a uniform (flat) prior on n_i.

Since we do not know a priori which cluster individual i came from, their marginal probability given the model parameters can then be written in mixture model form as:

$$P(D_i|\Theta) = \sum_{k=1}^{K} p(D_i|c_i = k, \Theta_{c_i})\alpha_k \tag{9}$$

and where the full likelihood $p(D|\Theta)$ can be written as a product of these terms over i as in Equation 6.

Equations 7, 8, 9, and 6 completely specify our generative model for the observed data D. In this manner, a full likelihood for all of the observed data can be constructed in a straightforward systematic manner by building up first from a model of an individual sequence (Equation 7), to a model of how a set of sequences are generated for an individual (Equations 8 and 9), to a model of how data for a set of individuals is generated (Equation 6).

Once the likelihood is defined in this manner, the EM procedure is relatively straightforward to define. The E-step is a straight-forward evaluation of

$$p(c_i = k|D_i, \Theta) = \frac{p(D_i|c_i = k, \Theta_k)p(c_i = k)}{\sum_{u=1^K} p(D_i|c_i = u, \Theta_u)p(c_i = u)} \qquad 1 \le k \le K \tag{10}$$

via the likelihood terms defined above. The M-step becomes:

$$\begin{aligned} \alpha_k^{\text{new}} &= \frac{1}{N}\sum_{i=1}^{N} p(c_i = k|D_i, \Theta) \\ \pi_k^{\text{new}}(s) &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} p(c_i = k|D_i, \Theta)\delta(s, s_{ij1})}{\sum_{i=1}^{N} p(c_i = k|D_i, \Theta)} \\ T_k^{\text{new}}(s_2|s_1) &= \frac{\sum_{i=1}^{N} p(c_i = k|D_i, \Theta) r_i^{s_1 -> s_2}}{\sum_{i=1}^{N} p(c_i = k|D_i, \Theta)} \end{aligned} \tag{11}$$

This equation states that the new mixing proportions are proportional to the membership probabilities, while the new initial state probabilities and transition probabilities are obtained by counting initial states and transitions and weighing them by the membership probabilities. The term $r_i^{s_1->s_2}$ represents the count of transitions from state s_1 to state s_2 in all the sequences associated with individual i.

The computational complexity of this algorithm at a high-level is the same as the standard multivariate EM algorithm for mixtures, i.e., linear in the total number N of individuals, in the total number of observations $\sum_{ij} d_{ij}$, and in the number of iterations of the EM algorithm. Within each iteration, for each observation, the computation of the M-step and the E-step will be model-dependent. For Markov mixtures, for example, the complexity is linear in the sum of the lengths of all sequences (i.e., linear in the total number of discrete symbols observed), and thus the overall algorithm retains its linearity. For more complex component models, the complexity can be higher.

4.3 Differences between this Framework and "Standard EM"

Recall that the standard approach to mixture modeling involves a single observation vector per individual i. Thus, we cluster vectors rather than individuals. The framework outlined in this paper differs in two important aspects from this standard EM approach:

1. In the E-step of Equation 10 the membership weights are associated with individuals i rather than with data observations d_{ij}. This is quite intuitive: we can view this as having the memberships of each data observation d_{ij} being tied to a single probability membership distribution for individual i.

2. In the M-step, if one individual has more observations than another (e.g., either more sequences and/or longer sequences), then that individual's data D_i will get more weight in parameter estimation through the r_i terms in Equation 11. Although it is intuitive that this should be the case, it is worth pointing out that the probabilistic framework takes care of this issue of different amounts of data for different individuals (via the likelihood model) in an automatic and consistent fashion. In a non-probabilistic framework there is no obviously consistent framework for handling this issue.

5 Experimental Results

5.1 Clustering Individuals based on Web Browsing Behavior

Table 1 shows a small fraction of actual Web session data obtained from a large commercial software company. Details of the data are proprietary but can be characterized in the following manner. Page-requests from a Web-browser are recorded on individual's client machines and downloaded nightly to a central archive. Sessions are defined (somewhat arbitrarily) as any set of page requests where the gap between successive requests does not exceed 30 minutes. Page requests are automatically categorized into 10 different categories, based on the nature of the Website from which the page is requested.

We investigated the use of Markov mixtures for this problem. Thus, for this particular data set, we have the following instantiation of our general approach:

- **Type of Data:** Discrete-valued sequences, variable length, multiple sequences for some individuals

- **Model:** mixture of Markov components

- **Parameters:** weight, initial state distribution, and transition matrix per cluster.

The data set contains 7,845 individuals with a total of 13,082 sessions (sequences). Each sequence consists of page requests categorized into 10 categories (states). In the results presented here we clustered the data into $k = 20$ clusters.

Figure 5.1 shows three transition matrices summarizing three interesting clusters. Each square in the figure represents a joint probability of the transition $s_i- > s_j$. Note that this information not only takes into account how likely it is that a user will go from state s_i to state s_j, but also how often the user is in state s_i. The first cluster contains users that mostly stay within a single category (high self transition probability). The second cluster shows users that tend to navigate among three different categories but stay slightly longer in the first two. The third cluster shows users that mostly stay in a single state, but occasionally make a short visit to another state.

Viewing transition matrices are one way to visualize resultant clusters. For a larger-scale study and additional results with a similar type of data see Cadez et al. (2000, submitted).

5.2 Clustering Genes using Gene Expression Data

Gene expression data provides another direct application for our clustering framework. In an effort to classify and understand the behavior of the human genome, scientists employ tiny DNA mircoarrays which afford the placement of thousands of distinct genes in a compact

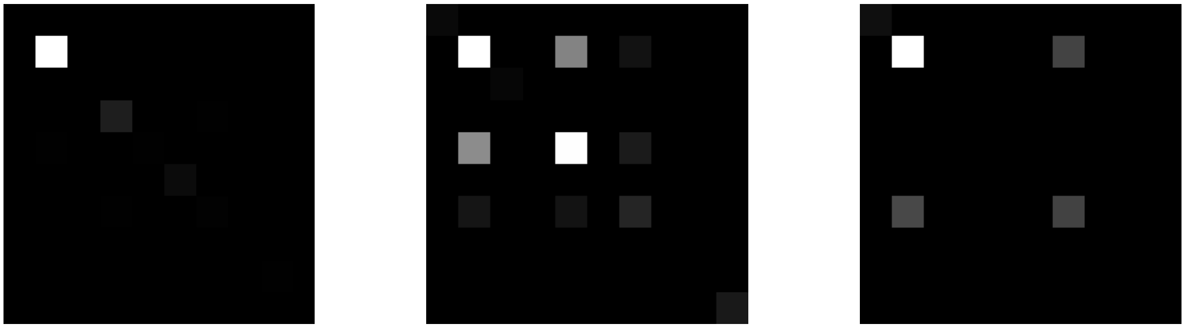

Figure 1: Three joint–transition probability matrices for the Web page–request data. Each square represents joint probability $p(s_i, s_j)$, where s_i and s_j are all combinations of the 10 available states. Lighter squares represent higher probability.

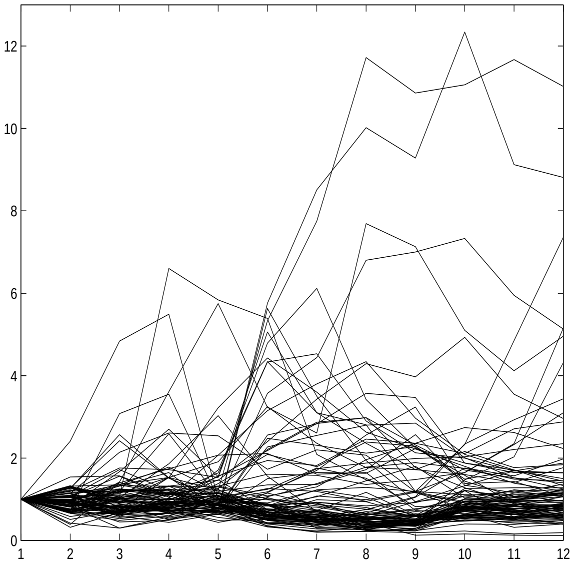

Figure 2: This picture shows the actual gene expression data. Note that only one-fifth of the data set is displayed.

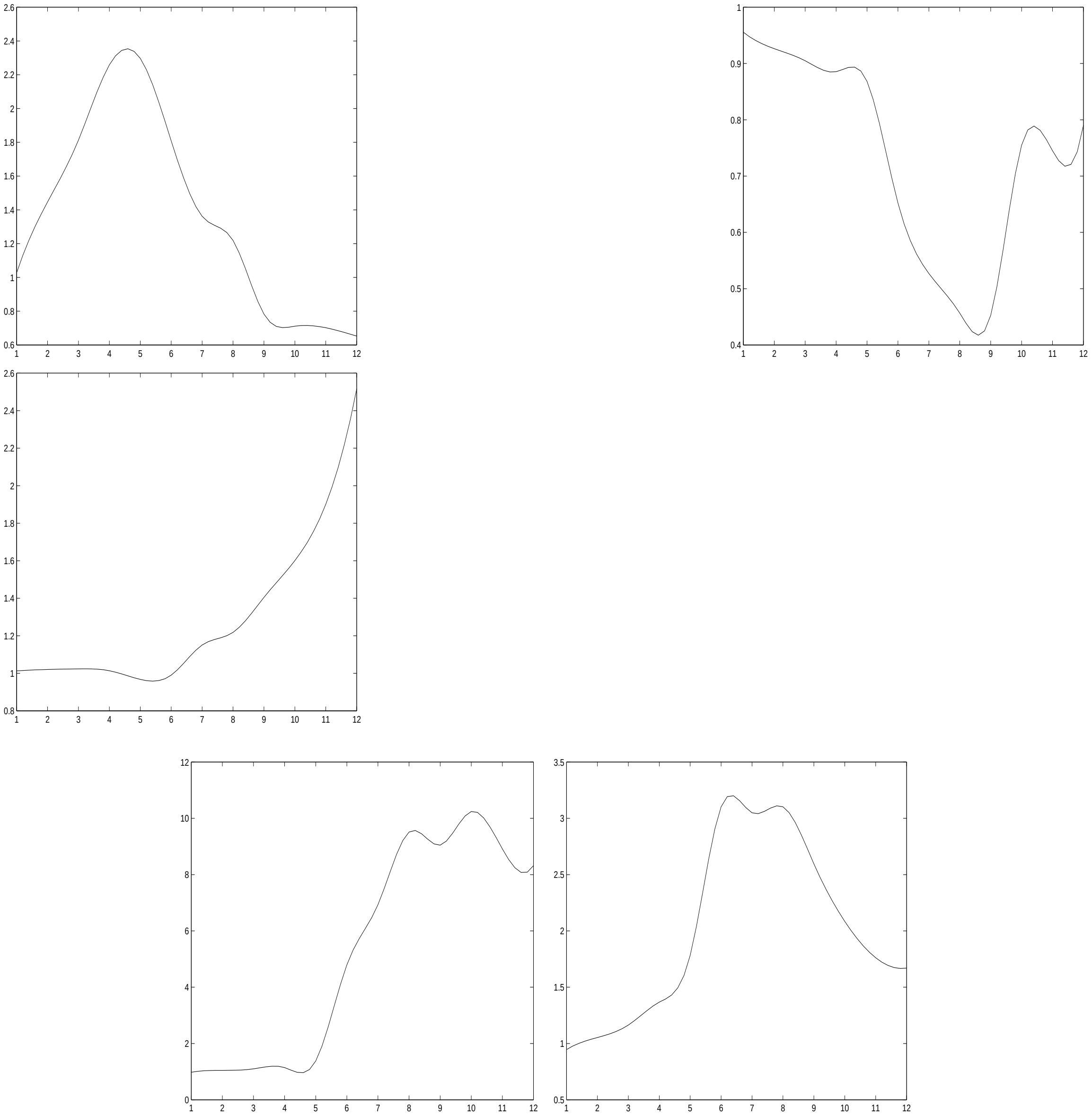

Figure 3: Kernel regression clusters returned on gene expression data.

rectangular array. The microarrays allow scientists to test the responses (or expressions) of a set of various genes under specific stimuli.

A typical gene expression data set contains a set of sequences whose values measure the level of response for a certain set of genes over time. For example, a data set might contain 1000 sequences measuring the responses of 1000 different genes over a 24 hour period. This data set might contain 20 response measurements in each sequence, yielding 20,000 measurements in total.

Scientists find it useful to be able to cluster a set of genes into several groups, in which each group contains genes evincing similar behavior. Armed with such a clustering, scientists may be able to more easily probe the functional purpose or role for which these "super-groups" are responsible.

The results in this section were obtained from a gene expression data set containing 517 sequences, each consisting of 12 measurements. The measurements were taken over a 24 hour period. The details are as follows.

- **Type of Data:** real-valued sequences as a function of time, fixed-length
- **Model:** mixture of kernel-regression components
- **Parameters:** weight, bandwidth

Eisen et al. (1998) describe this data set in greater detail and provide a clustering using standard hierarchical (average-linkage) clustering.

Here we show that it is a simple matter to cluster this data using our unified framework. We chose to model this data with a mixture model containing kernel regression mixture components. This type of model seems more natural than employing a standard hierarchical approach (vector-based) since you are explicitly modelling a gene's expression dependent on time. See Gaffney and Smyth (1999) for a more detailed discussion of kernel regression mixture models.

Figure 2 shows a picture of a portion of the data set. The y-axis shows the level of response and the x-axis simply indexes time. Figure 3 shows 5 of the clusters that were returned when the data was analyzed. The returned clusters match those found by Eisen et al. (1998) rather closely.

5.3 Clustering Patients based on Red-Blood Cell Cytograms

Red blood cell cytograms are two-dimensional histograms of red blood cells' volume and hemoglobin concentration. They are routinely obtained from flow-cytometric machines available in some medical labs. Understanding properties of joint distribution of volume and hemoglobin concentration is an important diagnostic tool in discovering patients with some types of blood disorders like iron deficient anemia (IDA) (McLaren, C. E. 1996). Even more important scientific insights can be obtained by examining a population of individuals, where each individual can be represented by one or more cytograms. For example, this analysis can reveal variability among individuals, variability of a single individual over time, or it can be used to classify new patients.

In an earlier study (Cadez et al, 1999) we showed how each cytogram can be characterized as an 11–dimensional vector. This way, each individual has one or more 11–dimensional vector data points associated with him or her. In the experiments presented here we specifically look at 2 cytograms per individual, the so called "duplicates". Duplicates exist since

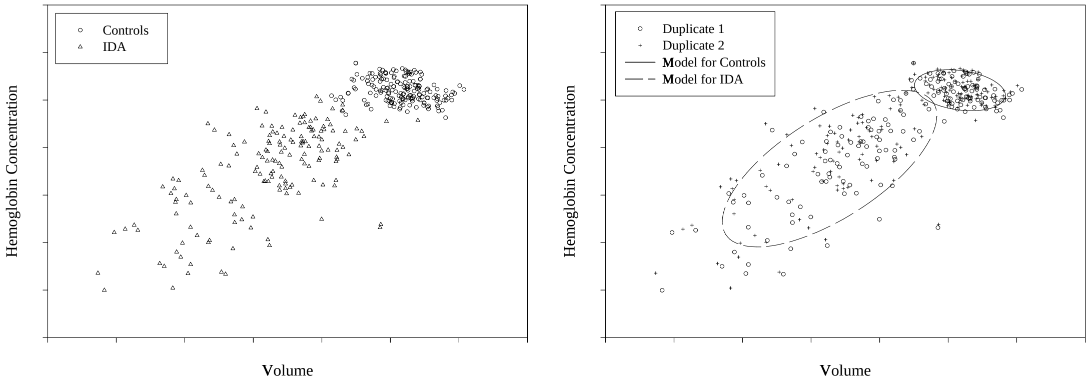

Figure 4: Red-blood cell patient data (a) labeled into control and iron deficient classes with two measurements per patient, (b) results of running our clustering framework on the data in (a) with the class labels removed.

each blood sample is typically analyzed twice with a 15 minute interval in between. The red blood cell data set consists of 90 controls (healthy individuals) and 82 patients with IDA. Figure 4 shows the 2 most important dimensions (out of 11 available) that can be viewed as an approximation to cytogram mean. Each dot represents a single duplicate, so there are twice as many dots as there are individuals. The left figure shows true classification, while the right figure shows results of our clustering (using 11–dimensional vector data and with class labels removed). The right figure also shows different duplicates with different symbols revealing variability of individual data.

To summarize, this data set and model can be described as:

- **Type of Data:** Vector data, variable number of data points per individual
- **Model:** mixture of Gaussian components
- **Parameters:** weight, mean, and covariance matrix per cluster.

6 Conclusion

We presented a unifying probabilistic framework for clustering individuals or systems into groups when the available data measurements are not multivariate vectors of fixed dimensionality. a. We derived a general EM algorithm for clustering this type of data and demonstrated its usefulness within several applied examples. A key idea in this paper is the fact that one is able to plug in an appropriate model for a data set and apply it within our general framework to generate parameter estimates and cluster models in a straightforward and consistent manner. This allows the data analyst to take full advantage of all of the available data on an individual without having to develop specialized algorithms for each different type of data present.

Acknowledgements

The work described in this paper was supported by the National Science Foundation under Grant IRI-9703120. The authors gratefully acknowledge the contributions of Christine McLaren in providing the red blood cell data.

References

Blender, R., Fraedrich, K., and Lunkeit, F. (1997) Identification of cyclone-track regimes in the North Atlantic. *Quart J. Royal Meteor. Soc.*, 123, 727–741.

Cadez, I. V., and Smyth P. (1999) Probabilistic clustering using hierarchical models, ICS Technical Report 99–16, UC Irvine.

Cadez, I. V., C. E. McLaren, P. Smyth, and G. J. McLachlan 'Hierarchical models for screening of iron-deficient anemia,' in *Proceedings of the 1999 International Conference on Machine Learning*, I. Bratko and S. Dzeroski (eds.), Los Gatos: CA, Morgan Kaufmann, 77–86, June 1999.

Cadez, I. V., Heckerman, D., Meek, C., Smyth, P., White, S. (2000) 'Visualization of navigation patterns on a Web site using model-based clustering,' submitted for publication.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM-Algorithm,' *Journal of the Royal Statistical Society, Series B*, 39, pp.1–38.

DeSarbo, W., Oliver, R. L., and Rangaswamy, A. (1989), 'A simulated annealing approach to clusterwise linear regression,' *Psychometrika*, 54(4), pp.707—36.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) 'Cluster analysis and display of genome-wide expression patterns,' *Proc. Natl. Acad. Sci. USA*, 95(25), pp. 14863–68.

Fraley, C. and A. E. Raftery, 'How many clusters? Which clustering method? Answers via model-based cluster analysis,' *Computer Journal*, 41, 578–588, 1998.

Gaffney, S., and P. Smyth, 'Trajectory clustering using mixtures of regression models,' in *Proceedings of the ACM 1999 Conference on Knowledge Disovery and Data Mining*, S. Chaudhuri and D. Madigan (eds.), New York, NY: ACM, 63–72, August 1999.

Haccou, P. and Meelis, E. (1992) *Statistical Analysis of Behavioral Data: An Approach based on Time-Structured Models*, New York, NY: Oxford University Press.

Jain, A. and Dubes, R. (1988) *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D., 'Hidden Markov models in computational biology: applications to protein modeling,' it J. Mol. Bio., 235:1501–1531, 1994.

McLachlan, G. J., and Krishnan, T., *The EM Algorithm and Extensions*, New York: John Wiley and Sons, 1997.

McLaren, C. E. (1996) 'Mixture models in haematology: a series of case studies,' *Statistical Methods in Medical Research*, 5(2):129-53.

Poulsen, C. S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behavior. *International Journal of Research in Marketing*, 7, 5–19.

Rabiner, L. R., C. H. Lee, B. H. Juang, and J. G. Wilpon, 'HMM clustering for connected word recognition,' *Proc. Int. Conf. Ac. Speech. Sig. Proc*, IEEE Press, 405–408, 1989.

Ridgeway, G., 'Finite discrete Markov process clustering,' Technical Report TR 97-24, Microsoft Research, Redmond, WA, 1997.

Smyth, P., 'Clustering sequences using hidden Markov models,' in *Advances in Neural Information Processing 9*, M. C. Mozer, M. I. Jordan and T. Petsche (eds.), Cambridge, MA: MIT Press, 648–654, 1997.

Smyth, P., 'Probabilistic model-based clustering of multivariate and sequential data,' in *Proceedings of the Seventh International Workshop on AI and Statistics*, D. Heckerman and J. Whittaker (eds), San Francisco, CA: Morgan Kaufman.

Spath, H. (1979) 'Clusterwise linear regression,' *Computing*, 22(4), pp. 367–73.

Wedel, M. and Steenkamp, J. B. (1991) 'A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation,' *Journal of Marketing Research*, 28, pp.385–96.

Wedel, M. and Kamakura, W. A. (1998) *Market Segmentation: Conceptual and Methodological Foundations*, Boston, MA: Kluwer Academic Publishers.