Communications in Statistics, Theory and Methods 34 (6), 2005

Data Visualization by Pairwise Distortion Minimization

By Marc Sobel, and Longin Jan Latecki*

Department of Statistics and Department of Computer and Information Sciences
Temple University, Philadelphia, PA 19122.

We dedicate this paper to the memory of Milton Sobel who provided inspiration to us and the academic community as a whole.

ABSTRACT

Data visualization is achieved by minimizing distortion resulting from observing the relationships between data points. Typically, this is accomplished by estimating latent data points, designed to accurately reflect the pairwise relationships between observed data points. The distortion masks the true pairwise relationships between data points, represented by the latent data. Distortion can be modeled as masking dissimilarity measures between data points or, alternatively, as masking their pairwise distances. The latter class of models are encompassed by metric scaling methodology (MDS); the former are introduced here as competitors. The former class of models include Principal Components Analysis, which minimizes the global distortion between observed and latent data. We model distortion using mixtures of 'pairwise difference factor-analysis' statistical models. We employ an algorithm which we call 'stepwise forward selection' for purposes of identifying appropriate starting values and determining the appropriate dimensionality of the latent data space. We show that the pairwise factor-analysis models frequently better fit the data because they allows for direct modeling of pair-wise dissimilarities between data points.

Marc Sobel (marc.sobel@temple.edu) is an Associate Professor in the Department of Statistics, Fox School of Business and Management, 1810 N. 13th Street, Longin Jan Latecki (latecki@temple.edu) is an Associate Professor in the Department of Computer and Information Sciences (CIS) 314 Wachman Hall, 1805 N. Broad Street; Temple University, Philadelphia, PA 19122

1. INTRODUCTION

There has been considerable interest in both the machine learning and statistical modeling literature in comparing, registering, and classifying image data. From the practitioners perspective, there are a number of advantages if such algorithms are successful. First, algorithms of this sort can provide mechanisms for visualizing the data. Second, they provide a mechanism for learning the important features of the data. Because feature vectors typically live in very high dimensional spaces, reducing their dimensionality is crucial to most data-mining tasks. Many algorithms for reducing data dimensionality depend on estimating latent (projected) variables designed to minimize certain energy (or error) functions. Other algorithms achieve the same purpose by estimating latent variables using statistical models.

Generally, dimensionality reduction methods can be divided into metric and nonmetric methods. Metric methods start with data points (in a high-dimensional space) with observed pairwise distances between them. The goal of metric methods is to estimate latent data points, living in a lower dimensional space, whose pairwise distances accurately reflect the pairwise distances between the observed data points. Methods of this sort include those proposed by J. Sammon [2]. Nonmetric methods start with data points whose pairwise relationships are given by 'dissimilarities' which need not correspond to a distance. In contrast, principal components analysis minimizes the global distortion between observed and latent data values. Metric methods incorporate additional steps designed to provide constraints within which latent dissimilarities, having appropriate properties, can be optimally estimated. Methods of this sort include those of Kruskal [3]. In this paper we take as our starting point observed data points, living in a high-dimensional space. The pairwise relationships between these data points are represented by the corresponding relationships between latent data points, living in a low-dimensional space. The pairwise dissimilarities between observed data points are masked by noise. This noise could arise in many different settings; examples include:

(i) settings where partitioning data into groups is of paramount interest, and lack of

straightforward clusters can be modeled as the impact of noise on the pairwise relationships between data, and

(ii) settings where the 'energy' of data points is being modelled; in this case noise arises in evaluating the relationship between the energy of 'neighboring' data points.

Our approach is different from that of probabilistic principal components (see [4]) where noise masks the relationship between each individual data point and its latent counterpart. By contrast, in our approach noise masks pairwise dissimilarities between data points and analogous latent quantities; we will see below that this difference in approach allows us to biuld in some extra flexibility into the interpretation and modeling.of high-dimensional data. Our approach is similar in spirit to the approach employed in relational Markov models [5].

The main goal of multidimensional scaling (MDS) is to minimize the distortion between pairwise data distances and the corresponding pairwise distances between their latent projections. This insures that the latent (or projected) data optimally reflect the internal structure of the data. MDS algorithms frequently involve constructing loss functions which prescribe (scaled) penalties for differences between observed and latent pairwise distances. See also [6] for a graphical analysis of MDS. MDS methods are used widely in behavioral, econometric and social sciences [7]. The most commonly used nonlinear projection methods within MDS involve minimizing measures of distortion (like those of Kruskal and Sammon) (e.g., Section 1.3.2 in [8] and [9]). These measures of distortion frequently take the form of loss or energy function. For purposes of comparison we focus on the loss (energy) function proposed by J. Sammon in [2].In this paper we compare MDS methods (as represented by that of Sammons) with those using pairwise-difference factor analysis methodology. We employ stepwise forward selection algorithms (see below) to provide good estimates of the dimension of the latent data space and

'starting' vectors appropriate for use with either of these two methodologies. Other starting value options which have been recommended include

(i) random starting values (see [2], [8], and [13]) and
(ii) starting at latent variable values arising from employing principal components analysis (PCA) [2].

The former option, still commonly used, fails to provide any useful training (information).

The latter option fails because PCA does not provide optimal (or near-optimal) solutions to minimizing the (noise) distortion of the data. In fact, as will be seen below in the examples, the distortion reduction for PCA generated latent variables is very small. For multi-dimensional scaling models, after employing stepwise forward selection algorithms for the aforementioned purposes, we typically use gradient descent methods (see e.g., [9] and [10]) to minimize Sammon's cost function. For factor analysis mixture models, after employing stepwise forward selection algorithms, we use the EM algorithm (see [11]) to provide estimates of the parameters. We partition the pairwise differences between data into two groups by determining data membership using EM-supplied probabilities and an appropriate threshold. The first group consists in those pairs of data with 'small' pairwise differences; the second in those pairs of data points with 'large' pairwise differences. The first group of pairs provides a mechanism for distinguishing data clusters; the second group provides a mechanism for distinguishing which pairs of points are 'different' from one another.

In the next section we compare two different ways of projecting the relationships between pairs of data points into latent k-dimensional space, denoted by R^k; typically k will be taken to be 2 or 3. Using the notation $\mathbf{F_1}, ..., \mathbf{F_n}$ for the observed feature vector data, multidimensional scaling is concerned with projecting a known real valued dissimilarity

function, $\{\mathbf{D(i,j) = D(F_i, F_j)}\}$ of the ordered pairs of features $\{\mathbf{F_i}, \mathbf{F_j}\}$ onto their latent k-dimensional functional counterparts $\{\|\mu_\mathbf{i} - \mu_\mathbf{j}\|\}$ $(1 \le i < j \le n)$. Sammons energy function provides a typical example of this. In this setting the latent r-vectors are chosen to minimize a loss (or energy) function of the form,

$$S_H(F \mid \mu) = \sum_{1 \le i < j \le n} \left(\frac{D(i,j) - \|\mu_i - \mu_j\|^2}{D(i,j)} \right)^2; \qquad (1.1)$$

in mu. We have in mind the example, $D(i,j) = \mathbf{l'(F_i - F_j)}$. Many variations on this basic theme have been explored in the literature (see [3]). As a counterpoint to this approach, we introduce the next section:

2. FORMULATING THE PROBLEM USING STATISTICAL MODELS

In this section we assume (as above) that feature vectors, associated with each data object are themselves observed. We employ a variant of probabilistic principal components models, introduced in [4]. Our variant is designed to take account of the fact that we seek to model the noise distortion between pairs of feature vectors rather than the noise distortion associated with each individual feature vector. We follow the main principle of MDS which is to map the data to a low-dimensional space in such a way that the distortion between data points is minimal. We introduce some necessary notation first. Let **'D(i, j)'** denote the dissimilarity between feature vectors $\mathbf{F_i}$ and $\mathbf{F_j}$ $\in \Box^k$ $(1 \le i<j \le n)$ (which is allowed to live in more than one dimension).

Explicitly, we assume that this dissimilarity measure **'D(i, j)'** 'lives' in a Euclidean space with dimensionality **p** (assumed to be less than or equal to the dimensionality **k** of the feature space). In the example, given below, we take $\mathbf{D(i,j) = F_i - F_j}$ $(1 \le i<j \le n)$, in which case p=k. Other examples include assuming that $\mathbf{D(i,j) = 1'\left(F_i - F_j\right)}$ (for a known p-vector **l)** The general linear statistical model assumed below takes the form:

$$\mathbf{D(i,j) = A^{(g)}\left(\mu_i^{(g)} - \mu_j^{(g)}\right) + \varepsilon_{i,j};\quad 1 \le i < j \le n} \qquad (2.1)$$

$'g'$ identifies the particular mixture model component;
(i.e., $\mathbf{'g(i,j) = s'}$ means that the pair (i,j) belong to mixture component s)
$\mathbf{'A^{(g)}'}$ are parametric $p \times q$ matrices indexed by the component π;
$\mathbf{'\mu_i^{(g)}'}$ are parametric $q \times 1$ latent vectors for feature $\mathbf{F_i}$ indexed by the component π and observation index 'i'. $(1 \le i<j<n)$.
$\mathbf{'\varepsilon_{i,j}'}$ is the pairwise noise distortion for features $\mathbf{F_i, F_j;}$ $(1 \le i<j \le n)$
It is assumed below that the errors $\mathbf{'\varepsilon_{i,j}'}$ are normally distributed with common variance $\left(\boldsymbol{\sigma}^{(g)}\right)^2 \mathbf{I}$.

While the dimensionality p of the D's (defined above) may be quite high, the dimensionality q of the latent mu vectors will typically be assumed to be quite small. (In the composite movie example analyzed in section 5, below, L is taken to be 5). The matrices $\mathbf{'A^{(g)}'}$ are (latent) projection matrices projecting paired differences between parametric latent μ vectors onto their feature vector paired difference counterparts. We use the EM algorithm [11] to estimate the model parameters under the assumption that the observed dissimilarities are given by $D(i,j) = F_i - F_j$ $(1 \le i<j \le n)$. The equations needed for purposes of doing this calculation are given in the appendix. In equation (2.2), below, we assume that the aforementioned mixture model, indexed by g, consists of exactly 2 components. The first component comprises pairs of

observations with small variance; the second comprises pairs of observations with large variance. The first component model is designed to characterize those pairs of feature vectors whose difference is well-approximated by the corresponding difference between their latent variable counterparts; the second, those pairs of feature vectors whose difference is not well-approximated by this difference. Specifically, we assume that the first component variance $[\sigma^{(1)}]^2$ is significantly smaller than the second, $[\sigma^{(2)}]^2$. First component model parameters were selected to minimize quantities of the form,

$$SS[g=1] = \sum_{i,j} \left\{ \frac{\mathbf{D}(i,j) - \hat{\mathbf{A}}^{(g=1)}(\hat{\boldsymbol{\mu}}_{\mathbf{i}}^{(g=1)} - \hat{\boldsymbol{\mu}}_{\mathbf{j}}^{(g=1)})}{\boldsymbol{\sigma}} \right\}^2 P(\mathbf{D}(i,j) \mid g=1) \qquad (2.2)$$

where the 'hatted' quantities are the EM algorithm estimates of the corresponding parameters and '$P(\mathbf{D}(i,j) \mid g=1)$' is the probability specified in the EM algorithm (see the appendix, below).

Model Fit and Assessment

We assess the fitness of data visualization models using Bayesian p-values [12]. This can be formulated as the probability that the information obtained from the model is less than expected under an aposteriori update of the data. Information quantities like those derived below are discussed in [13]. This kind of calculation is not possible for typical MDS models because they are not formulated as statistical models. In the model introduced at the beginning of this section, the information contained in the observed dissimilarity measures, assuming an uninformative prior and ignoring marginal terms, is,

$$INF\left(M \mid \mathbf{D}\right) = E\left\{\log\left(L\right) \mid \mathbf{D}\right\} = \sum_{1 \le i < j \le n} \mathrm{E}\left\{\log(\mathrm{L}_{i,j}) \mid \mathbf{D}\right\} \qquad (2.3)$$

where 'L' denotes the likelihood of the data and

$$`L_{i,j} = \left[1 \Big/ \sigma^{(g)} \right] \exp\left\{ - \frac{\left\| D(i,j) - A^{(g)}(\mu_i^{(g)} - \mu_j^{(g)}) \right\|^2}{\left(\sigma^{(g)}\right)^2} \right\},$$

For the model introduced in section 2, the right hand side of equation (2.3) can be approximated, omitting terms which don't involve the observed dissimilarity measures, by

$$INF(M \mid D) \approx \hat{I}NF(M \mid \mathbf{D}) = - \sum_{1 \le i < j \le n} \left\{ \left(\frac{D(i,j) - \hat{A}^{(g=1)}(\hat{\mu}_i^{(g=1)} - \hat{\mu}_j^{(g=1)})}{\hat{\sigma}^{(g=1)}} \right)^2 \hat{P}(i,j \mid g=1) \right\}$$

$$- \sum_{1 \le i < j \le n} \left\{ \left(\frac{D(i,j) - \hat{A}^{(g=2)}(\hat{\mu}_i^{(g=2)} - \hat{\mu}_j^{(g=2)})}{\hat{\sigma}^{(g=2)}} \right)^2 \hat{P}(i,j \mid g=2) \right\} \qquad (2.4)$$

where the hatted quantities are all the EM algorithm estimates (see the appendix) (see [13] for a more complete discussion of information quantities like that given in equation (2.4)). Posterior updates of the dissimilarities were simulated via:

$$D^*(i,j) \sim \begin{cases} \mathrm{N}\left\{ \hat{\mathrm{A}}^{(\mathrm{g}=1)}(\hat{\mu}_i^{(g=1)} - \hat{\mu}_j^{(g=1)}), \left(\hat{\sigma}^{(g=1)}\right)^2 \right\} & \text{wprob} \quad \mathbf{\hat{P}(i,j \mid g=1)} \\ \mathrm{N}\left\{ \hat{\mathrm{A}}^{(\mathrm{g}=2)}(\hat{\mu}_i^{(g=2)} - \hat{\mu}_j^{(g=2)}), \left(\hat{\sigma}^{(g=2)}\right)^2 \right\} & \text{wprob} \quad \mathbf{\hat{P}(i,j \mid g=2)} \end{cases} \qquad (2.5)$$

$(1 \le i < j \le n)$. ('N(*1,*2)' refers to the normal distribution with mean *1 and variance *2).

The posterior Bayes p-value is equal to:

$$Bayes\ pvalue = P\left(\hat{I}NF(M \mid \mathbf{D}^*) < \hat{I}NF(M \mid \mathbf{D}) \mid \mathbf{D} \right) \qquad (2.6)$$

(the probability in equation (2.6) being calculated over the distribution specified by equation (2.5)). For the models examined below the (Bayesian p-values) were all between 80 and 90% indicating good fits.

3. ALGORITHMS EMPLOYED FOR MDS DATA VISUALIZATION

We use 'online' gradient descent algorithms to estimate parameters in the MDS approach to data visualization [6]. The gradient of Sammon's energy function with respect to the parametric vector r_i restricted to terms involving r_j; $j \neq i$ is:

$$\Delta_i(E \mid j) = \left\{ \frac{\left\| \mu_i - \mu_j \right\| - D_{i,j}}{D_{i,j}} \right\} \left\{ \frac{\mu_i - \mu_j}{\left\| \mu_i - \mu_j \right\|} \right\} \qquad (3.1)$$

The analogous quantity with 'i' and 'j' switched is: $\Delta_j(E \mid i) = -\Delta_i(E \mid j)$ $(1 \le i < j \le N)$.

An 'online' gradient descent algorithm can, in theory, be based on an iterative calculation of the r-vectors by updating r-vectors using the following iterative steps:

$$\mu_{\mathbf{i}}^{(\mathbf{new})} = \mu_{\mathbf{i}}^{(\mathbf{old})} - \varepsilon \Delta_{\mathbf{i}}(\mathbf{E} \mid \mathbf{j}) \qquad (3.2)$$

$$\mu_{\mathbf{j}}^{(\mathbf{new})} = \mu_{\mathbf{j}}^{(\mathbf{old})} - \varepsilon \Delta_{\mathbf{j}}(\mathbf{E} \mid \mathbf{i})$$

We have already remarked on the problem of training a large number of μ vectors for the purpose of starting the gradient descent and EM algorithms. We show below how to select a small number 'l' of vantage objects $\mathbf{v_1}, ..., \mathbf{v_l}$ from among the observed input objects such that the Sammon's energy function,

$$E(D \mid v_1, ..., v_l) = \sum_{1 \le i < j \le l} \left\{ D(i,j) - \sqrt{\sum_{p=1}^{l} \left[\frac{\left(D(\mathbf{v}_p, \mathbf{F_i}) - D(\mathbf{v}_p, \mathbf{F_j}) \right)^2}{D(i,j)^2} \right]} \right\}^2 \qquad (3.3)$$

is fairly small. This provides us with well-trained (i.e., well-fitted) starting r-vectors given by

$$\mu_i^{(0)} = \left(D(\mathbf{v_1}, i), ..., D(\mathbf{v_p}, i) \right); \quad i=1,...,n$$

Since for purposes of visualization l=2 or l=3 is typically sufficient to insure small values of $E(\mathbf{D} \mid \mathbf{v_1}, ..., \mathbf{v_l})$ for a moderate sized data set, two or three vantage vectors usually suffice in this case [14]. For a large number n of observed data points vantage objects can be obtained by the stepwise forward selection process described below. We note that this process improves on the adhoc procedures used heretofore [13].

Stepwise Forward Selection

At each stage s=1,...,l, the stepwise forward selection algorithm selects one new vantage object $\mathbf{v}_s$ that is added to the set of previously chosen objects $\mathbf{v_1}, ..., \mathbf{v_{s-1}}$. The vantage object v_s is chosen to satisfy:

$$v_s = \arg\min_{v \in A} E\left(D \mid v_1, ..., v_{s-1}, v\right) \qquad (3.4)$$

where $'\arg\min_{v \in A}'$ denotes the vector in A for which the minimum value of $E(D \mid v_1, ..., v_{s-1}, v)$ is reached. At stage s, having chosen the vantage object v_s, we prune the objects by comparing the energies $E(D \mid v_1, ..., v_{i-1}, v_{i+1}, ..., v_s)$ (i=1,….,s) with the energy $E(D \mid v_1, ..., v_{s-1}, v_s)$ (i=1,...,s). If any of them are smaller than the latter energy, we remove the vantage object v_i and return to the next step of the process.

4. Experimental Results

In this section we examine the performance of the proposed algorithms on various data sets. We begin with the classical Iris data set [7]. The Iris data is composed of 150 vectors each having 4 components. It is known that there are 3 clusters, each having 50 points; these consist of one clear cluster, denoted by ‘A’ below, and two clusters, ‘B’ and ‘C’ that are hard to distinguish from one another. We first compare 3 dimensional projections obtained using Sammon’s algorithm (cf., equation (1.1)) with input vantage vectors produced via stepwise

forward selection [15]. Figure 1a, below, shows a 3 dimensional projection of the Iris data obtained using the classical Principal Components algorithm for data visualization; the distortion measure for this estimate (computed using Sammon's energy function) is 3,255. Figure 1b, below, shows a 3 dimensional projection of the Iris data obtained using Sammons data visualization algorithm; the distortion measure (computed using Sammons energy function) for this estimate is 544. As can be seen, we cannot clearly distinguish between clusters B and C using PCA. By contrast, clusters B and C can be clearly distinguished using Sammons data visualization algorithm.

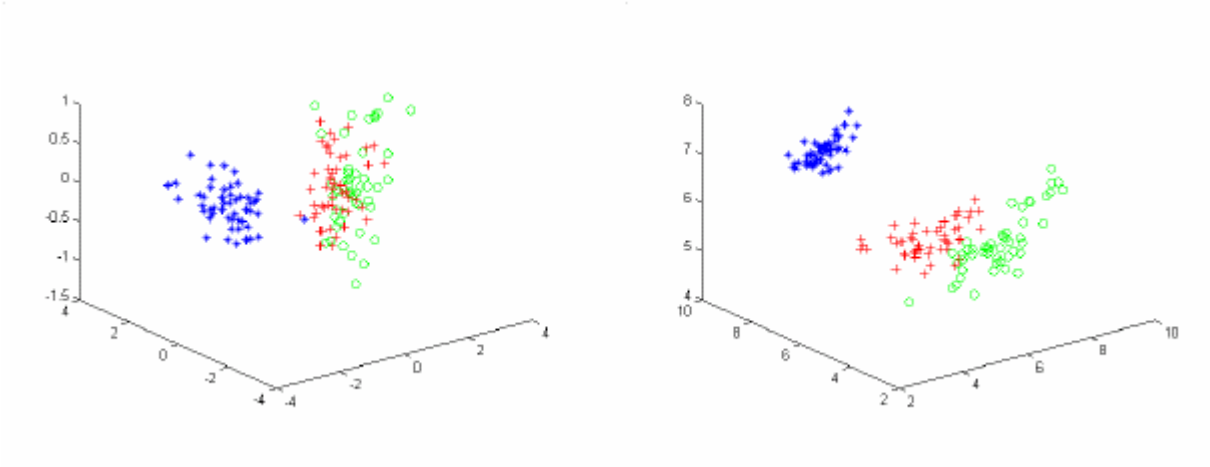

Figure 1(a on the left and b on the right) 3 Dimensional projections of the Iris data (a): obtained by classical PCA, and (b) using Sammons algorithm for data visualization (employing vantage vectors produced by stepwise forward selection.

We now turn to evaluating two dimensional projections for the data set referred to below as 'composite movie', below. Composite movie is composed of 10 shots (each having 10 frames) taken from 4 different movies; these consist in:

a) 4 shots taken from the movie, 'Mr. Beans Christmas': frames 1 to 40.

b) 3 shots taken from the movie 'House Tour': frames 41 to 70.

c) 2 shots taken from a movie we created (referred to below as 'Mov1'): frames 71:90.

d) 1 shot from a movie in which Kylie Minogue is interviewed: frames 91 to 100.

The frames can viewed at the site: http://www.cis.temple.edu/\~\latecki/ImSim . Using image processing techniques described in [16], we assign a vector with 72 features to each of the 100 frames. We obtain a data set consisting of one hundred 72 component feature vectors. 'Composite Movie' has two hierachical grouping levels; it can be grouped using shots and separately using movies. We expect to distinguish both between the shots and, on a higher level, between the movies. The best data visualization algorithm (cf., figure 2a) for this data set was obtained using the pairwise difference factor analysis mixture model outlined in section 2; we used starting vantage vectors, computed using stepwise forward selection. As can be seen in Figure 2a, below, there are 4 clear clusters that belong together in the upper left corner of the figure. They represent the 4 shots grouped to form excerpts from Mr Beans Christmas. In the lower right corner, we see two clear clusters. These are two shots from the movie referred to as 'Mov1'. The 3 shots from 'House Tour' are represented by the 3 rightmost clusters in the middle of the figure. Figure 2b below, employs Sammon's data visualization algorithm, using gradient descent (see section 3) with the same vantage vectors. Sammon's data visualization gave a significantly worse picture of the data. This is demonstrated by the fact that the movies are no longer grouped correctly. For example, the four clusters from Mr. Bean's christmas are mixed with clusters from the other movies in the lower right hand quadrant.

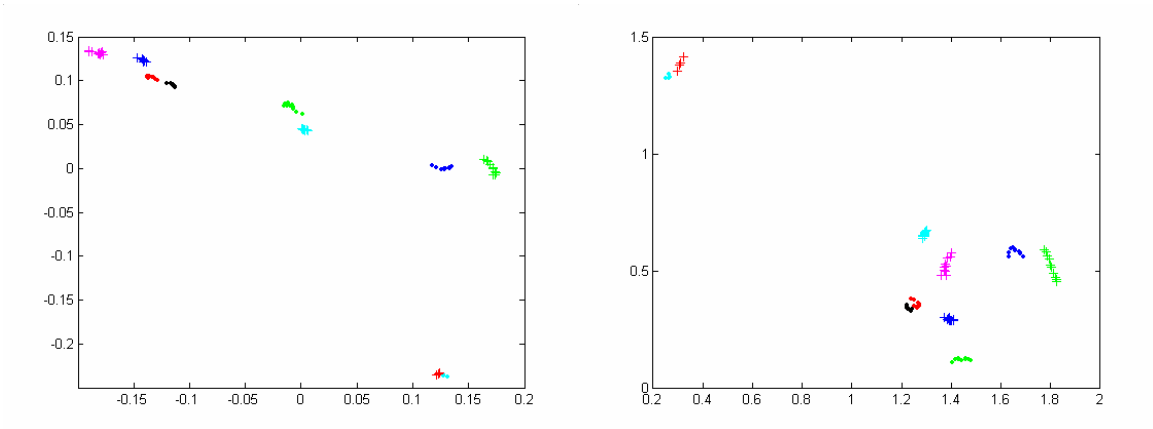

Figure 2 The2 dimensional projections of the composite movie data obtained by (a) the pairwise difference factor analysis mixture algorithm (on the left) and (b) Sammons algorithm computed using gradient descent (on the right).

5. Conclusions and Future Research

We have introduced stepwise forward selection algorithms and demonstrated their value in providing starting values for factor mixture models and Multidimensional scaling algorithms.

It has been shown that pairwise difference factor mixture models provide good data visualization for a wide variety of data when vantage vectors, constructed using stepwise forward selection, are used to generate appropriate starting values. Our examples illustrate that factor mixture models frequently provide better data visualization than Multidimensional Scaling algorithms, designed for the same purpose. Their superiority arises as a result of their flexibility in modeling data distortion. We have shown how to assess the fitness of factor mixture models and used these results to assess fit in the examples presented above. We would like to extend our current work to include mixture factor models which incorporate intra-component correlations.

Appendix

Calculations via the EM algorithm needed for the mixture factor difference model:

In this section we describe the Expectation-Maximization (EM) algorithm [10] used to estimate the latent variables and parameters introduced in section 2 above. For purposes of clarity we repeat the formulation of our model:

$$\mathbf{D(i,j) = A^{(\pi)}\left(\mu_i^{(\pi)} - \mu_j^{(\pi)}\right) + \varepsilon_{i,j};\quad 1 \le i < j \le n} \qquad \text{(A.1)}$$

'$\boldsymbol{\pi}$' identifies the particular mixture model component;

(i.e., '$\boldsymbol{\pi(i,j) = s}$' means that the pair (i,j) belong to mixture component s)

'$\mathbf{A^{(\pi)}}$' are parametric $p \times q$ matrices indexed by the component π;

'$\boldsymbol{\mu_i^{(\pi)}}$' are parametric $q \times 1$ latent vectors for feature $\mathbf{F_i}$ indexed by the component π and observation index 'i'. $(1 \le i<j<n)$.

'$\boldsymbol{\varepsilon_{i,j}}$' is the pairwise noise distortion for features $\mathbf{F_i, F_j;}$ $(1 \le i<j \le n)$

It is assumed below that '$\boldsymbol{\varepsilon_{i,j}}$' that the errors are normally distributed with common variance $\left(\boldsymbol{\sigma^{(\pi)}}\right)^{\mathbf{2}}$.

In the notation below, $\mu_i^{(old;g)}$ (respectively, $\mu_i^{(g;new)}$) denotes the 'old' or previous value (respectively, new or updated value) of the latent parameter μ_i for the g'th component (g=1,2). (i=1,…,n). Analogous notation is used to characterize the projection matrix A. We also employ the notation, $\mu_{(-i)}^{(g;old)}$ for the average of the old (or previous) mu-parameters of the g'th component excluding the i'th; similar notation applies to the the new (or updated) parameters (g=1,2; i=1,…,n). Then, employing the notation,

$$\text{P(i,j;g)=P(D(i,j)|g)=}\frac{\exp\left\{-(1/2[\sigma^{(g;old)}]^2)\left\|\text{D-A}^{(\text{g; old})}(\mu_i^{(g;old)} - \mu_j^{(g;old)})\right\|^2\right\}}{\sum_{\kappa=1}^{2}\exp\left\{-(1/2[\sigma^{(\kappa;old)}]^2)\left\|\text{D-A}^{(\kappa;old)}(\mu_i^{(\kappa;old)} - \mu_j^{(\kappa;old)})\right\|^2\right\}} \qquad \text{(A.2)}$$

for the probability weight attached to the observed pair of dissimilarity measure $D(i,j)$, we update the latent mean vectors $\mu_i^{(g;new)}$ (g=1,2; i=1,…,n) via,

$$\hat{\mu}_i^{(g;new)} \leftarrow \frac{\sum_{j\neq i}\left(A^{(g;new)\prime}A^{(g;new)}\right)^{-1}A^{(g;new)\prime}\left(D_{i,j}+A^{(g;new)}\mu_{(-i)}^{(g;new)}\right)P(i,j;\pi)}{\sum_{j\neq i}P(i,j;g)} \qquad (A.3)$$

The back projection matrix ' $A^{(g;new)}$ ' is updated using the formula,

$$A^{(g;new)} \leftarrow \frac{\sum_{i<j}D_{i,j}\left(\mu_i^{(g;new)}-\mu_j^{(g;new)}\right)\left\{\sum_{i<j}\left(\mu_i^{(g;new)}-\mu_j^{(g;new)}\right)\left(\mu_i^{(g;new)}-\mu_j^{(g;new)}\right)'\right\}^{-1}P(i,j;\pi)}{\sum_{i<j}P(i,j;g)} \qquad (A.4)$$

for g=1,2. We upgrade the variances $\sigma^{(g;new)}$ via,

$$\sigma^{(g;new)} \leftarrow \sqrt{\frac{\sum_{i<j}\left\|D_{i,j}-A^{(g;new)}(\mu_i^{(g;new)}-\mu_j^{(g;new)}\right\|^2P(i,j;g)}{\sum_{i<j}P(i,j;g)}} \qquad (A.5)$$

.

Bibliography

[1] Jolliffe,I.T. *Principal Component Analysis*, Springer-Verlag, **1986**

[2] Sammon, J.W., Jr., A nonlinear mapping for data structure analysis, IEEE Trans. Comput. **1969,** 18, 401—409.

[3] T.F. Cox and M.A. Cox. *Multidimensional Scaling,* Chapman and Hall, **2001.**

[4] Bishop, M., and Tipping, M.E. A Hierarchical Latent Variable Model for Data Visualization, IEEE Transactions on Pattern Analysis and Machine Intelligence, **1983,** 20,3,281-293.

[5] Koller, D. Probabilistic Relational Models, invited contribution to, *Inductive Logic Programming, 9th International Workshop (ILP-99),* Saso Dzeroski and Peter Flach, Eds, Springer Verlag, **1999,** pp. 3-13.

[6] McFarlane, M., and Young F.W., Graphical Sensitivity Analysis for Multidimensional Scaling, Journal of Computational and Graphical Statistics, **1994,** 3, 1, 23-33.

[7] Kohonen, T. *Self-organizing maps*, Springer-Verlaag, New York, **2001.**

[8] Faloutsos C., and Lin, K.-I. FastMap: A fast algorithm for Indexing,Data-Mining and Visualization of Traditional and Multimedia Datasets, Proc. ACM SIGMOD International Conference on Management of Data**, 1995**, 163--174.

[9] Mao, J. and Jain, A.K.: Artifical Neural Networks for Feature Extraction and Multivariate Data Projection. IEEE Transactions on Neural Networks **1995**, 6,2.

[10] Lerner, Boaz, Guterman, Hugo, Aladjem, Mayer, Dinstein, Itshak, and Romem, Yitzhak,On pattern classification with Sammon's Nonlinear Mapping - An Experimental Study, Pattern Recognition, **1998,** 31, 371-381.

[11] Laird, N.M., and Rubin, D.B., Maximum likelihood for incomplete data via the em algorithm, Journal of Royal Statistical Society, **1977**, 39, pp. 1—38.

[12] Gelman, Carlin, Stern and Rubin, *Bayesian Data Analysis*, Chapman and Hall, **1995.**

[13] Maclachlan, G. and Peer, D., *Finite Mixture Models,* Wiley Series in Probability and Statistics, **2000.**

[14] Fraley, C. and Raftery A.E. , How Many Clusters? Which clustering method? Answers via Model Based Cluster Analysis, Computer Journal, **1999**, 41, pp297:306.

[15] Jolliffe, I.T.*, Principal Components Analysis,* Springer series in statistics, 2nd edition, **2002.**

[16] Latecki, L.J., and Wildt, D., Automatic Recognition of Unpredictable Events in Videos, Proceedings of the International Conference on Pattern Recognition (ICPR), **2002**, 16.