

Latent Variable Models for Link Analysis of Similarity Data

Juan K. Lin
Department of Statistics
Rutgers University
Piscataway, NJ 08854

August 16, 2003

Abstract

There is a need for statistical models which can organize large collections of pair-wise similarity relationships between objects into meaningful clusters. Similarity data consists of non-negative quantitative measurements of similarity between object pairs. Examples include internet connectivity data and document word count data. We present various latent variable models for finding reduced rank structure in similarity data. Applications are presented in unsupervised clustering, targeted clustering based on pre-defined cluster relationships, and graph layout using reduced rank graph approximations.

1 INTRODUCTION

A major challenge confronting applied statistics and computer science is the analysis and organization of large datasets. In addition to statistics and computer science, this has become a major focus of researchers in the fields of applied math, data mining, electrical engineering and biology, with diverse applications such as web mining, collaborative filtering, customer relations management, information retrieval, gene function grouping and patient diagnosis. There is a pressing need for statistical models which can organize a large collection of pair-wise similarity relationships between objects into meaningful clusters. We present latent variable models with relatively few parameters for analyzing similarity data. In addition, we investigate dramatically faster variants of the EM algorithm for fitting the models.

We begin by describing a ubiquitous form of data called similarity data, or sometimes *affinity* or *dyadic* data (see e.g. Weiss 1999, Hofmann 2001). Given two discrete sets of objects, $\mathcal{X} = \{x_1, \dots, x_n\}$ and $\mathcal{Y} = \{y_1, \dots, y_m\}$, similarity data consists of a quantitative relationship measurement between each object pair (x_i, y_j) . This measurement can be in the form of a similarity score, or a co-occurrence count. The similarity data is often summarized as a two-way table with the n rows and m columns corresponding to the objects of \mathcal{X} and \mathcal{Y} respectively. The quantitative relation measurement between objects x_i and y_j is captured in the i -th column j -th row entry of the table. For the special case where the $\mathcal{X} = \mathcal{Y}$, the relation scores are between object pairs in a single domain, and the two-way table is a square matrix. A further consideration is whether the scores are symmetric, in which case the similarity data table will be a symmetric matrix.

Similarity data is common in today's computer age. Three specific examples of interest are:

- Internet connectivity data: A search engine like Google (*www.google.com*), analyzes the link structure of billions of internet web sites in order to rank the relevance of the web sites to specific queries. Here the similarity data consists of only a single domain, with \mathcal{X} as the set of 2 billion or so websites to be cataloged. The similarity data consists of a sparse square binary link matrix of order 2 billion by 2 billion. Since web links are directed, this link matrix is not symmetric in general. A second example of internet connectivity data consists quantitative traffic flow between routers. The similarity data consists of a single domain, with \mathcal{X} consisting of all routers being considered. The relationship between the object pair (x_i, x_j) is simply the amount of internet traffic flow from x_i to x_j . There is tremendous interest in discovering inherent clustering structure which arise from the self-organizing nature of the internet

- Information retrieval: The bag-of-words model is a commonly used framework for summarizing word documents in information retrieval. Here, the set \mathcal{X} corresponds to a collection of words of interest, and \mathcal{Y} corresponds to the set of documents. The similarity data consists of the number of occurrences of word x_i in document y_j . A challenge for the analysis of bag-of-words similarity data is to organize groups of words into themes, and collection of documents into hierarchies of topics.
- Microarray data: In microarray data analysis, gene expression, as quantified by fluorescence levels, are tabulated for a large collection of genes for different experimental scenarios. Here \mathcal{X} consists of a set of experiments while \mathcal{Y} is a set of genes. The domain \mathcal{X} can also be taken to be a set of patients. There is great interest in discovering gene function groupings, as well as the diagnostic classification of patients.

Though the datasets come from different fields, they all consist of pairwise relationships either between objects in a single domain (e.g. websites), or between objects in two separate domains (e.g. words-documents, and experiments-genes).

The organization of this paper is as follows. In Section 2, we present the probabilistic framework for unsupervised organization and dimension reduction based on relational data. We base the presentation on Google’s “random surfer” model, as well as related work on clustering based on similarity matrices. In Section 3, details of the latent variable models and improved fitting algorithms is presented. The basic building block of the latent variable model is presented first for symmetric similarity data, and then for the general case. Section 4 presents promising preliminary results of the model’s application to both information retrieval and microarray data. The paper concludes with a summary and discussion of future work.

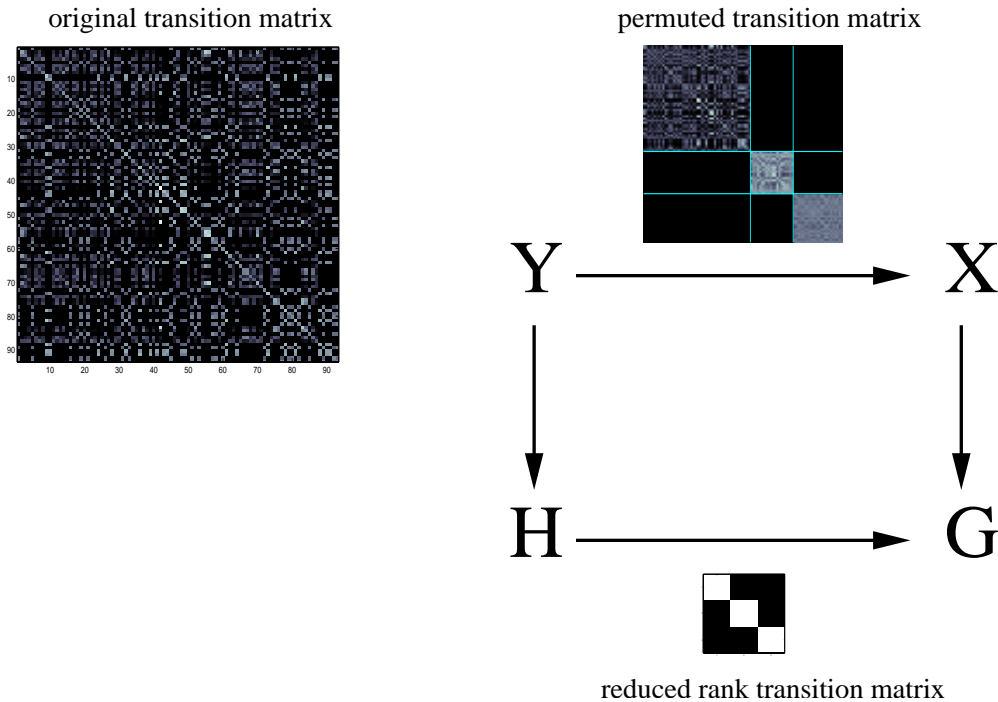


Figure 1: Approximate diagram motivating the latent variable models of similarity data. The original similarity matrix is normalized into a transition matrix, and is depicted on the left. On the right, the probabilistic mappings $p(h|y)$ and $p(g|x)$ accomplish a soft probabilistic grouping of the objects in \mathcal{Y} and \mathcal{X} into far fewer groups. The latent variables find cluster groupings with similar within cluster and between cluster transitions. In this simple case, the transition matrix permuted according to the soft clustering, and the diagonal 3x3 reduced rank transition matrix both reveal three distinct, well separated and clusters.

2 MODELING SIMILARITY DATA

The success of Google’s search engine is in large part due to the effectiveness of Google’s “random surfer” model. An individual surfing the web is modeled as a Markov process, transitioning from website to website according to the link structure of the web. With probability p , the web surfer is assumed to randomly follow a link on the current page being viewed, and with the remaining probability $(1 - p)$, the surfer resets and randomly views a website. The relational data matrix in this scenario is a square binary pairwise connectivity matrix between the websites. The random surfer model normalizes this connectivity matrix into a transition matrix by dividing each column by its sum. This transition matrix is then mixed with a uniform transition matrix with weights p and $(1 - p)$ respectively. The relevancy of a website assigned by Google’s PageRankTM algorithm is simply the steady-state probability of that website under the “random surfer” Markov process.

Similar to Google’s approach, our analysis of relational data begins with the specification of a discrete Markov process based on the data. For Google’s random surfer model, this involved simply normalizing the link connectivity matrix into a transition matrix. Whereas Google is only interested in assigning a ranking to websites, we are interested in a full probabilistic clustering of the websites.

For more general types of pairwise similarity matrices, recently there has been a great deal of interest across various research communities in spectral clustering methods which cluster based on eigenvectors of a normalized similarity matrix (see eg. Shi et.al. 2000, Weiss 1999, Meila et.al. 2001, Ng et.al. 2002). Following the recent approaches in the spectral clustering community, we normalize an similarity matrix into a transition matrix in order to quantify the similarity between objects in terms of transition probabilities.

Our approach is a two-staged analysis and organization of relational similarity data. First the relational information is modeled by an underlying transition matrix. Second we investigate various structured latent variable models for finding cluster groupings with similar within cluster and between cluster transitions. The intuition behind the clustering approach is depicted in Figure 1. The probabilistic clustering in the latent model framework is a natural outcome of finding a reduced rank approximation of a Markov process.

3 LATENT MARKOV ANALYSIS

We begin by describing various latent variable models for the reduced rank approximation of transition matrices that we wish to investigate. Two main categories of models, termed Latent Markov Analysis(LMA) models, are introduced (Lin 2003). We first address the case where the transition matrix is consistent with a reversible random walk. A more general case is subsequently addressed. Monotonically convergent EM-type algorithms are presented for all models. LMA is applied to clustering based on pairwise similarities, where similarities between observations are described probabilistically. A cluster-based permutation of the transition matrix results in its approximate block uniform reorganization. Furthermore, relationships between the inferred clusters are again described probabilistically by the reduced rank transition matrix. LMA simultaneously infers the clusters and abstracts the relationships between them, which can be represented in the form of a weighted graph. Finally, a “targeted” LMA model is introduced where the transition between latent cluster states are specified. This provides an algorithm which searches for clusters satisfying pre-specified relationships.

The Latent Markov Analysis approach to finding reduced rank approximations of transition matrices is depicted in Fig. 1. We begin with discrete random variables X and Y , and either a specified conditional probability or data sufficient for an empirical conditional probability $p(x|y)$. Let the number of states in X be n , and Y be m , where both n and m are large. Let \mathcal{S}^n denote the $n - 1$ dimensional simplex defined by $\sum_1^n p(x_i) = 1, x_i \geq 0$. This is the simplex over all possible n state multinomial distributions. The conditional probability, or transition matrix $p(x|y)$ is an operator which maps \mathcal{S}^m to \mathcal{S}^n by $p(x) = \sum_i p(x|y_i)p(y_i)$. Thus, $p(x|y)$ maps a distribution of Y into a distribution of X . The diagram in Fig. 1 depicts the reduced rank approximations of this mapping via mappings to and between low dimensional simplices \mathcal{S}^{k_1} and \mathcal{S}^{k_2} . This is accomplished in a latent variable model framework where discrete latent variables G with $k_1 (<< n)$ states and H with $k_2 (<< m)$ states are introduced. For the case where the $n \times m$ transition matrix $p(x|y)$ can be permuted into $k_1 \times k_2$ blocks with elements within each block being identical, the commutative diagram drawn in Fig. 1 can be made exact with suitably chosen mappings. However, in general the commutative diagram is only approximate. The diagram also shows that the latent variable model finds a reduced rank

approximation of the transition matrix by grouping similar states in X together via the mapping $p(x|g)$ and clustering similar states in Y together via $p(y|h)$. For the special case where the transition matrix is consistent with a reversible random walk, the mapping is from \mathcal{S}^n to \mathcal{S}^n . This case is presented in this section, while the general case shown in the diagram in Fig. 1 is presented in Section 3.2.

The LMA model is applicable to the analysis of two-way contingency table data (eg. Agresti 1990) consisting of either co-occurrence (Hofmann 2001) or conditional-occurrence data. For co-occurrence data, the data is sampled from the full joint distribution over two random variables X and X' . Conditional-occurrence data, on the other hand, consists of samples from the various conditional distributions of one random variable given all the states of the second random variable. Co-occurrence and conditional-occurrence data specify the empirical joint and conditional distributions respectively. We address the more general case of conditional-occurrence data since the joint distribution fully specifies the conditional distribution.

3.1 LMA Model and Algorithm: Symmetric Similarity Data

In this section we consider the case where the specified empirical transition matrix $\tilde{p}(x'|x)$ is consistent with a reversible random walk. This is the case when the transition matrix is consistent with a symmetric joint distribution matrix, and leads to a model applicable to the analysis of commonly seen symmetric similarity matrices. The more general model without the reversibility assumption is presented in the Section 3.2. The LMA model tailored for this symmetric case has significantly fewer parameters than the general LMA model, and thus will be much easier to fit.

The “symmetric” LMA model consists of latent variables H and H' , along with the following exchangeability and conditional independence assumptions:

1. X and X' exchangeable given H
2. H and H' exchangeable given X
3. $X \perp X'|H$,
4. $H \perp H'|X$

where $X \perp X'|H$ denotes that X and X' are conditionally independent given H . It should be emphasized here that this is not a graphical model, since the assumptions above cannot be expressed in either a directed or undirected graph.

Where there will be no confusion, we will use shorthand notations $p(x, x') = P(X = x, X' = x')$, where order of the arguments is explicitly maintained. Assumption (1) implies

$$P(X = x, X' = x'|H = h) = P(X = x', X' = x|H = h).$$

This assumption also implies that $p(x, x', h)$ is symmetric with respect to x and x' . Note that conditional exchangeability of X and X' given H is a stronger condition than exchangeability of random variables X and X' . Denoting $P(X = x|H = h) \equiv g(x|h)$, we have $P(X' = x'|H = h) = g(x'|h)$ from the symmetry assumption. Similarly, assumption (2) implies

$$P(H = h|X = x) = P(H' = h|X = x) \equiv w(h|x).$$

Assumption (3) justifies the relation $p(x, x', h) = p(h)g(x|h)g(x'|h)$, while assumption (4) implies $p(h, h', x) = p(x)w(h|x)w(h'|x)$.

Since $p(h, x) = w(h|x)p(x) = g(x|h)p(h)$, the parameters in the model are specified by the distributions $g(x|h)$ and $p(h)$. Without loss of generality we begin with a specified symmetric empirical joint distribution $\tilde{p}(x', x)$, since with the reversibility assumption the joint can be constructed after computing the stationary distribution under the transition matrix $\tilde{p}(x'|x)$.

The maximum likelihood estimation problem boils down to the minimum information divergence problem of finding $g(x|h)$ and $p(h)$ which minimizes

$$D(\tilde{p}(x', x) || \sum_h g(x|h)g(x'|h)p(h)).$$

The EM algorithm for this model results in the iterations:

E-step

$$p(h|x, x') = \frac{g(x|h)g(x'|h)p(h)}{\sum_h g(x|h)g(x'|h)p(h)}$$

M-step

$$p(h) = \sum_{x', x} p(h|x, x')\tilde{p}(x', x)$$

$$g(x|h) = \sum_{x'} \frac{p(h|x, x')\tilde{p}(x', x)}{p(h)}.$$

After convergence, the reduced rank transition matrix $p(h'|h)$ is computed by:

$$p(h'|h) = \frac{\sum_x p(x)w(h|x)w(h'|x)}{\sum_{x, h'} p(x)w(h|x)w(h'|x)}.$$

The symmetry assumptions significantly reduce the number of parameters in this model. In addition, the conditional independence assumptions impose a symmetry between X, X' and H, H' . It is important to note that this model is not a graphical model.

3.1.1 Numerical Examples

We apply LMA to clustering based on a pairwise similarity matrix. From a set of observations $\vec{x}_1, \dots, \vec{x}_n$, the similarity matrix is defined as $A_{ij} = \exp(-\|\vec{x}_i - \vec{x}_j\|^2/2\sigma^2)$, where σ is a specified length scale. Following Meila et.al.(2001) and treating the observations as states in a random walk, we construct two discrete n -state random variables X and X' , with the joint distribution proportional to the similarity matrix, $p(x_i, x'_j) = kA_{ij}$ for all $i, j \in \{1, \dots, n\}$. Since the similarity matrix is symmetric, $\tilde{p}(x'|x)$ is consistent with a reversible random walk.

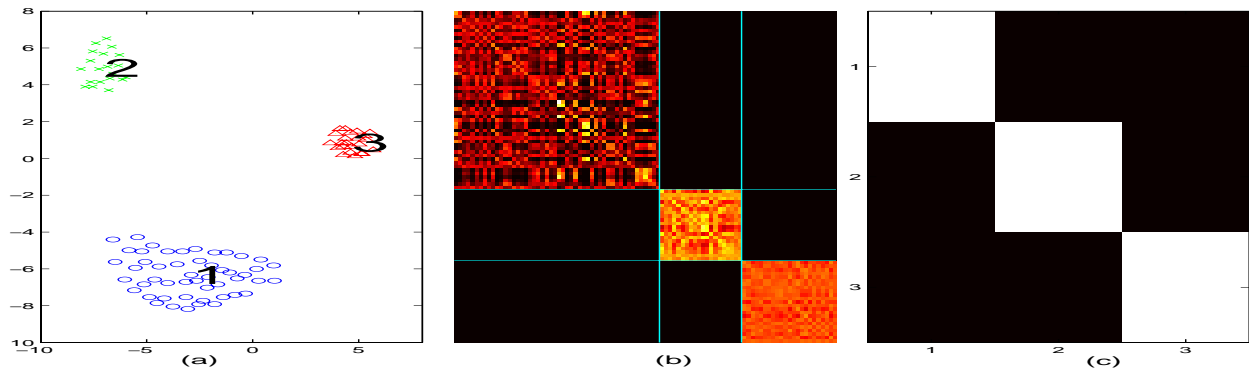


Figure 2: Latent Markov Analysis of self-transition matrix generated from the pairwise similarity matrix as described in the text. The well separated data is plotted on the left. The transition matrix, permuted in accordance with MAP assignment clustering is shown in the center. The reduced rank transition matrix is shown on the right.

The LMA clustering framework is summarized as follows. First similarities between objects are quantified probabilistically in the form of a transition probability between “object-states”. Second, a reduced rank approximation of the transition matrix is constructed via the LMA model. In the LMA clustering framework, similarities between clusters are again quantified probabilistically by the reduced rank transition matrix. Intuitively, similar “object-states” will have similar transitions to and from other states. This intuitive notion is demonstrated numerically in Fig. 2 for the simple example of three well separated clusters. The clustering of the observations is in accordance with the MAP assignment of each observation to the states of

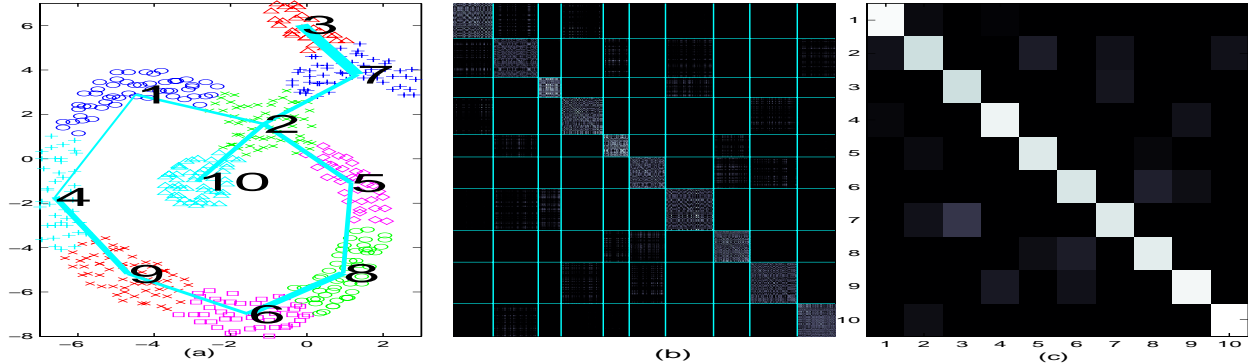


Figure 3: Latent Markov Analysis clustering based on a normalized similarity matrix. The observations are plotted on the left. The transition matrix, permuted according to LMA clustering is shown in the center. The reduced rank transition matrix is shown on the right.

the latent variable H . It should be noted that specification of the number of states in the latent variable H does not directly specify the number of clusters, since $p(h)$ can be very close to zero for a state, giving rise to a null cluster with no MAP assigned observations.

The 3 well separated clusters of observations are shown in Fig. 2(a). The transition matrix permuted in accordance with the MAP clustering assignment is shown in Fig. 2(b), and the 3×3 reduced rank transition matrix in Fig. 2(c). The algorithm is run for 100 iterations starting at random initial conditions with $\sigma = 2$ in A_{ij} . Since the reduced rank transition matrix is very close to the identity matrix, LMA approximately block-diagonalizes the transition matrix.

In Fig. 3, we show the latent Markov analysis of a more complex configuration of observations. Since the observations are no longer well-separated, a block-diagonalization of the transition matrix cannot be accomplished by a re-permutation of the transition matrix. Here, LMA constructs a block uniform approximation of the transition matrix. The LMA algorithm is run for 100 iterations, with $\sigma = 1$. The clustering of the observations into ten states is depicted in Fig. 3(a) through the use of different plot symbols and colors. In addition, numerical labels of the corresponding latent variable states are superimposed over the observations. The transition matrix, re-organized in ascending order with respect to the cluster number is shown in Fig. 3(b). Transition matrices are shown in the figures with all columns summing to one.

In Fig. 3(c), the reduced rank transition between the ten states of the latent variable are shown. The block uniform transformation is apparent. The LMA block uniform transformation captures the relationships between the clusters, as quantified probabilistically in the reduced rank transition matrix. For example, from the first column in Fig. 3(c), one can see that cluster 1 is near clusters 2 and 4. Similarly, from the second column, one infers that cluster 2 is near clusters 1, 5, 7 and 10. This structural relationship is supported by the underlying transition matrix in Fig. 3(b). The probabilistically quantified cluster similarity is represented in Fig. 3(a) in the superimposed weighted graph, where the widths of the edges connecting the cluster numerical labels are proportional to $[p(h_i|h_j) + p(h_j|h_i)]$.

3.2 LMA Model: General Similarity Data

Here we introduce the Latent Markov Analysis model for the general case where no reversibility assumptions are made. We begin with a data in the form of a conditional-occurrence table $n(x|y)$, which specifies the empirical conditional distribution $\tilde{p}(x|y)$. In the bag-of-words model for text information retrieval, this conditional-occurrence table consists of the word counts for each given document. The LMA model is a graphical model with latent variables G and H , as depicted in the approximate commutative diagram in Fig. 1, with conditional independence assumptions $G \perp Y|H$ and $X \perp \{H, Y\}|G$. The corresponding graphical model is depicted in Fig. 4. The parameters of the model are $p(x|g)$, $p(y|h)$ and $p(g, h)$.

A similar model has been investigated by Hofmann and Puzicha (1998) for the fitting of an empirical joint

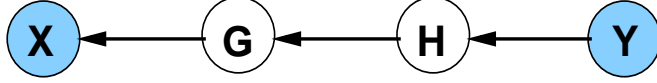


Figure 4: Directed graphical model for the general case LMA model.

distribution. In Section 3.2.1 we present a modified EM algorithm for the fitting of an empirical conditional distribution. Section 3.2.2 describes a combinatorially-inspired iterative I-projection algorithm for fitting the model with significantly faster convergence speed. Additional conditional independence assumptions which lead to a non-graphical model is discussed which seem to improve clustering in our numerical experiments.

3.2.1 Modified EM Algorithm

A modified EM algorithm for maximum likelihood parameter estimation results in the following iterative scaling algorithm:

$$\alpha(x, y) = \frac{\tilde{p}(x|y) \sum_{x,g,h} p(x|g)p(g, h)p(y|h)}{\sum_{g,h} p(x|g)p(g, h)p(y|h)}$$

$$p(g, h)^{new} = p(g, h) \sum_{x,y} \alpha(x, y)p(x|g)p(y|h)$$

$$p(y, h)^{new} = p(y|h) \sum_{x,g} \alpha(x, y)p(g, h)p(x|g)$$

$$p(x, g)^{new} = p(x|g) \sum_{y,h} \alpha(x, y)p(g, h)p(y|h),$$

where all parameters on the right hand side are estimates from the previous iteration. The conditional distribution parameters are computed by normalizing the joint distributions given above. The I-projection corresponding to the E-step in the EM algorithm has been modified since the data is in the form of a conditional-occurrence instead of a co-occurrence table. Equivalently, the constraint is in the form of a specified conditional distribution $\tilde{p}(x|y)$ instead of a joint. In the E-step, the I-projection was replaced with a reverse I-projection, minimizing the Kullback Liebler information divergence with respect to the second argument instead of the first.

To test out this model, we synthesized a 297×227 block uniform transition matrix out of a 20×16 matrix with elements uniformly chosen between 1 and 5. The full block uniform matrix is first normalized into a transition matrix, then *i.i.d* normally distributed noise of amplitude .003 is added, and the matrix renormalized. The rows and columns of this matrix were then separately permuted. A correct re-permutation of this matrix using a cyclic algorithm described below is shown in Fig. 5(c).

The modified EM algorithm presented above was experimentally found to be very slow to converge. In Fig. 5(b), the cluster permutation of the transition matrix is shown after 2000 iterations of the EM steps, with the latent variables G and H having 40 and 36 states respectively. The run-time in Matlab was 185 seconds on a P-III 550MHz PC.

3.2.2 Fast Cyclic I-projection Algorithm

Since the performance of the modified EM-algorithm was not encouraging, we investigated a combinatorially motivated algorithm to try to find the 20 and 16 clusters of the states in X and Y respectively. Instead of iterative projections of the full joint $p(x, y, g, h)$, we construct the following cyclic I-projection algorithm of various marginals. Given initial values for $p(x|g)$, $p(y|h)$ and $p(g, h)$:

- **cycle A:**

1. $p(x, y, h) = p(y|h) \sum_g p(x|g)p(g, h)$
2. Iterative scaling of $p(x, y, h)$ subject to the constraints $p(x|y)$ and $X \perp Y|H$.

- **cycle B:**

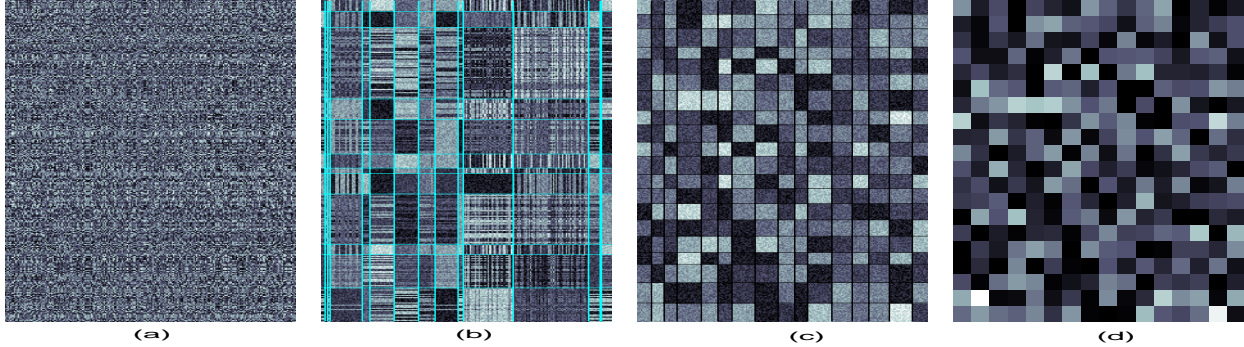


Figure 5: Latent Markov Analysis of a synthetic approximately block uniform transition matrix. (a) Specified transition matrix. (b) Re-permuted transition matrix using modified EM algorithm. (c) Re-permutation using cyclic I-projection algorithm. (d) Reduced rank transition matrix.

1. $p(x, g, h) = p(x|g)p(g, h)$
2. Iterative scaling of $p(x, g, h)$ subject to constraints $p(x, h)$ and $X \perp H|G$.

- **cycle A’:**

1. $p(x, y, g) = p(x|g) \sum_h p(y|h)p(g, h)$
2. Iterative scaling of $p(x, y, g)$ subject to the constraints $p(x|y)$ and $X \perp Y|G$.

- **cycle B’:**

1. $p(y, g, h) = p(y|h)p(g, h)$
2. Iterative scaling of $p(y, g, h)$ subject to constraints $p(y, g)$ and $Y \perp G|H$.

The iterative scaling algorithms in the cycles with specified joint distributions are the same algorithms for probabilistic latent semantic analysis (Hoffman 2001) and non-negative matrix factorization (Lee and Seung 1999). The cycles with specified conditionals are implemented with the modified E-step I-projection. Convergence of this cyclic algorithm was *dramatically* faster than for the algorithm given in Section 3.2.1.

In numerical experiments, many latent variable states under MAP assignment represented null clusters with no assigned observations. These states have very small corresponding probabilities in $p(h)$ or $p(g)$. We augmented the cyclic I-projections algorithm with a trimming step where these null-cluster states were removed. For $p(x|g)$, $p(y|h)$, and $p(g, h)$, a simple trimming and renormalization accomplished this task. This added trimming significantly improved the clustering. Finally, we found that the additional assumptions $X \perp H|Y$ and $G \perp \{H, Y\}|X$, which provide a nice symmetry amongst the random variables greatly improved the clustering. It should be noted that these additional conditional independence assumptions lead to a model which is not graphical. These additional model assumptions seem to introduce additional regularization which greatly helps the convergence. In the trimming step, these conditional independence assumptions are used to compute $p(g|h) = \sum_{x,y} p(g|x)p(x|y)p(y|h)$. The clustering and reduced rank approximation result using this algorithm is shown in Fig. 5(c,d). The resulting clustering in Fig. 5(c) is in exact accordance with the constructed block uniform mosaic structure. The approximate block uniform structure of the transition matrix after permutation (Fig. 5(c)) visually justifies its reduced rank approximation (Fig. 5(d)). The algorithm was run for 40 iterations of the 4 cycles, with trimming and regularization of $p(g|h)$ after every 10 iterations. Within each cycle, 20 iterative scalings were performed. The run-time in Matlab was 93 seconds on a P-III 550MHz PC.

Figure 6 shows an experimental comparison of the convergence speeds of the EM algorithm and two specifications of the cyclic I-projection algorithm. For comparable iteration cycles, the convergence for the I-projection algorithms are dramatically faster.

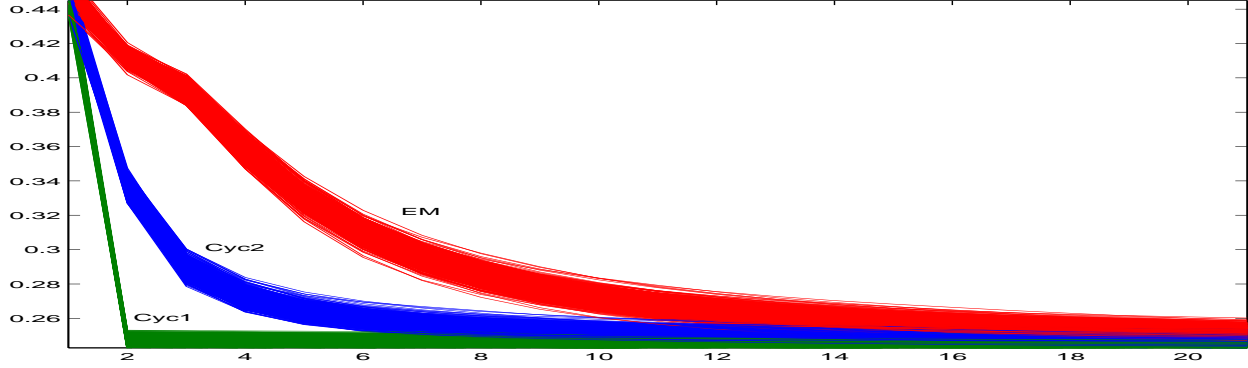


Figure 6: Experimental comparison of the EM algorithm with the cyclic I-projection algorithm. The KL-Divergence is plotted as a function of equivalent iteration number. For the EM algorithm (labeled *EM*), each equivalent iteration is 20 E and M steps. For *Cyc2*, each equivalent iteration is 10 iterations of the cycles A, B, A', B' , while for *Cyc1*, it is 5 iterations of cycles $A, B, A', B', A, B, A', B'$. Convergence and running time is significantly faster for the cyclic I-projection algorithm.

3.3 Targeted LMA: specification of prior $\tilde{p}(h', h)$

The LMA model can be extended to the case where the latent variables are jointly observed. Here we focus on the symmetric LMA model, though this can be extended to the more general case. The scenario is as follows: transition pairs between H and H' are observed in addition to transition pairs between X and X' . The goal of targeted LMA is to relate the variables through probabilistic mappings between both X, H , and X', H' . Targeted relational LMA clustering looks for a clustering solution which respects the probabilistic relationships between clusters as specified by the observed empirical distribution $\tilde{p}(h', h)$. Here we seek a maximum entropy solution for the full joint distribution, subject to the constraints given by the empirical distributions of $\tilde{p}(x', x)$ and $\tilde{p}(h', h)$. The algorithm we implemented is again an I-projection based algorithm consisting of repeated iterations of first the EM step presented in Section 3.1, followed by its symmetric dual obtained by mapping $x, x' \leftrightarrow h, h'$

E'-step

$$p(x|h, h') = \frac{w(h|x)w(h'|x)p(x)}{\sum_h w(h|x)w(h'|x)p(x)}$$

M'-step

$$p(x) = \sum_{h', h} p(x|h, h')\tilde{p}(h', h)$$

$$w(h|x) = \sum_{h'} \frac{p(x|h, h')\tilde{p}(h', h)}{p(x)},$$

where $p(h, x) = w(h|x)p(x) = g(x|h)p(h)$ is used to relate the parameters between the two EM steps. This algorithm consists of a sequence of I-projections in each cycle.

In Fig. 7, numerics for targeted LMA relational clustering is shown. The specified relation $\tilde{p}(h'|h)$ in Fig. 7(c) corresponds to three cyclically related clusters for states 1,2 and 3, a root cluster attached to three leaf clusters for states 4,5,6,7, and linearly related clusters 8,9,10,11. The targeted LMA model correctly groups the data into clusters that respect the specified relationships between the clusters, as seen in Fig.7(ab).

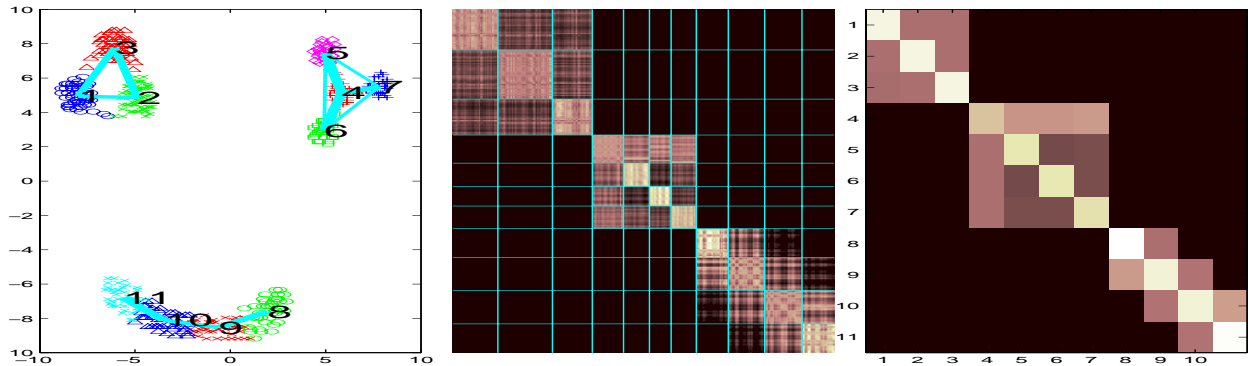


Figure 7: Hierarchical targeted relational clustering. The two-dimensional data is plotted on the left. The target reduced rank transition matrix is shown on the right. The re-permuted transition matrix is shown in the center.

4 Promising Preliminary Results on Real Data

4.1 Unsupervised text document classification

We present some promising preliminary results in the analysis of real datasets. We begin with an information retrieval application of automated text document classification. The analysis is done on the MED dataset (see e.g. Deerwester 1990), publically available at <ftp://ftp.cs.cornell.edu/pub/smart/med/>. The dataset consists of a collection of medical abstracts, organized into 30 topical labels. Following Kolenda et al. (2002), we restricted our attention to only the 124 abstracts in the first five topic groups, and 1159 words after common words were removed. The bag-of-words similarity data consists of a 1159 by 124 matrix of the number of occurrences of each of the 1159 words in each abstract. From the MED dataset the topics are labeled:

- crystalline lens in vertebrates, including humans.
- relationship of blood and cerebrospinal fluid
- oxygen concentrations or partial pressures. a method of interest is polarography.
- electron microscopy of lung or bronchi.
- tissue culture of lung or bronchial neoplasms.

We normalized the similarity data into a transition matrix and used the cyclic I-projection algorithm to fit a general LMA model with 5×5 latent states. The clustering of the 124 abstracts by the model were then compared with the original topic labels. We found a strong correspondence between the original labeled clusters and the clusters found using LMA. Specifically, the five topic groups were accurately labeled for 70.3%, 100%, 77.3%, 82.6% and 96.1% of the abstracts respectively. This is competitive with the leading algorithms.

4.2 Patient classification from gene expression data

We applied LMA to the publically available gene expression study of leukemia dataset (Golub et al.). The study was made of 72 subjects, of which 25 had acute lymphoblastic leukemia (*ALL*), and 47 had acute myeloid leukemia (*AML*). The *ALL* cases are further classified into 38 cases of the *B*-cell type (*ALL - B*) and 9 of the *T*-cell type (*ALL - T*). The gene expression levels for each subject were standardized to have zero mean and variance 1. After standardization, we selected the 200 genes with the largest variance across subjects. The unsupervised LMA clustering of the resulting 72 subjects by 200 genes similarity matrix is shown in Fig. 8. Comparing the subject labels *ALL - B*, *ALL - T* and *AML* with the LMA clusters, we found only 3 classification errors.

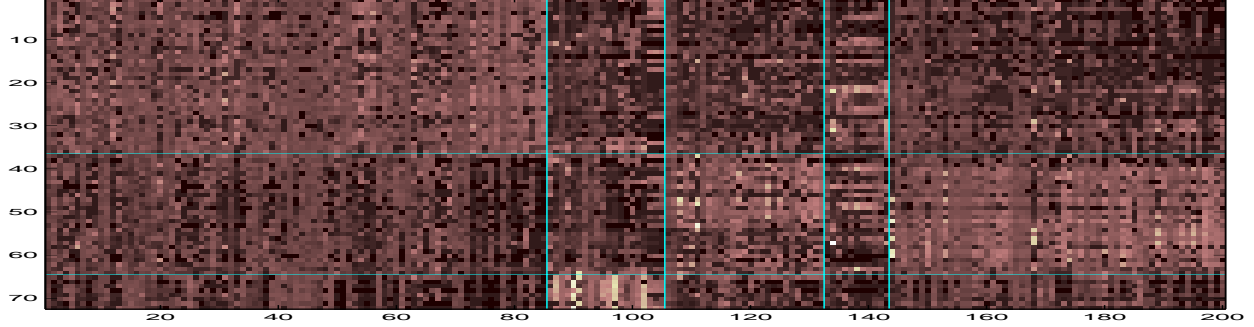


Figure 8: Simultaneous unsupervised clustering of leukemia microarray data into three patient clusters and five gene clusters. The 72 subjects and 200 genes are represented in the rows and columns respectively of the similarity matrix. Comparing the resulting clustering with the subjects’ original *ALL – B*, *ALL – T* and *AML* leukemia labels, only three patients are misclassified.

5 Discussion

In the reversible random walk case, the symmetric LMA clustering results in a single permutation of both the rows and columns of the transition matrix, applicable to the analysis of symmetric similarity matrix data. For the general LMA model, the latent variables provide separate permutations of the rows and columns. Related work include Friedman et al.(2001), where the clustering was presented in an “information bottleneck” framework specified via two networks G_{in} and G_{out} , and Lee et al.(1999) and Hofmann(1998, 2001) who applied their latent variable models to the clustering of words and documents. We foresee a few contributions of the presented research in the context of latent variable clustering models. The symmetric LMA has the benefit of a reduced number of parameters, as well as an additional conditional independence assumption which allows for the extraction of the transition matrix between the latent variables. A *targeted clustering* model was also introduced which allows for the targeted extraction of clusters with pre-specified inter-cluster relationships.

The dramatic improvements in convergence speed motivates further investigation of various iterative I-projection based algorithms. Another extension of this work will be to construct *hierarchical* models which lead to scalable algorithms that hierarchically clustering previously found clusters of objects. In the presented formalism, since relations between clusters are again described by a transition matrix, clusters can again be clustered with another LMA model. The approximate commutative diagram for the latent variable reduced rank model is shown in Fig. 9. Preliminary numerical results show improved clustering performance - in particular, the ability to naturally merge clusters in the previous level of the hierarchy. This hierarchical approach will lead to models which scale well to handle large similarity matrices.

$$\begin{array}{ccc}
 Y_{\mathcal{S}^{n_1}} & \xrightarrow{\tilde{p}(x|y)} & X_{\mathcal{S}^{n_2}} \\
 p(b|y) \downarrow & & \uparrow p(x|a) \\
 B_{\mathcal{S}^{k_1}} & \xrightarrow{p(a|b)} & A_{\mathcal{S}^{k_2}} \\
 p(d|b) \downarrow & & \uparrow p(a|c) \\
 D_{\mathcal{S}^{r_1}} & \xrightarrow{p(c|d)} & C_{\mathcal{S}^{r_2}}
 \end{array}$$

Figure 9: Approximate 2-layer commutative diagram motivating a latent variable model for hierarchical clustering.

References

- Agresti, A. (1990) *Categorical Data Analysis*. New York: Wiley.
- Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) The PageRank Citation Ranking: Bringing Order to the Web. unpublished manuscript:
<http://dbpubs.stanford.edu/pub/1999-66>.
- Cramer, E. (2000) Probability measures with given marginals and conditionals: I -projections and conditional iterative proportional fitting. *Statistics and Decisions* 18, 311-329.
- Csiszar, I. (1989) A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. *Annals of Statistics*, Vol.17, No.3, 1409-1413.
- Darroch, J. and Ratcliff, D. (1972) Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, Vol.43, No.5, 1470-1480.
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journ. Amer. Soc. for Inf. Science*. 41 391-407.
- Friedman, N., Mosenzon, O., Slonim, N. and Tishby, N. (2001) Multivariate information bottleneck. *Uncertainty in Artificial Intelligence 2001*.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.
- Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196.
- Hofmann, T. and Puzicha, J. (1998) Statistical Models for Co-occurrence Data. Technical Report, Artificial Intelligence Laboratory Memo AIM-1625.
- Larsen, L., Hansen, L.K., Szymkowiak, A., Christiansen, T., Kolenda, T. (2000) Webmining: Learning from the World Wide Web, special issue of *Computational Statistics and Data Analysis*, vol. 38, pp. 517-532.
- Lee, D. and Seung, S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(675), 788-791.
- Lin, J.K. (2003) Reduced rank approximations of transition matrices. In C. M. Bishop and B. J. Frey (eds), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Jan 3-6, 2003, Key West, FL.
- Meila, M. and Shi, J. (2001) A random walks view of spectral segmentation. in *Proc. International Workshop on AI and Statistics (AISTATS), 2001*
- Meng, X.L. and van Dyk, D. (1997) The EM algorithm - an old folk-song to a fast new tune. *J. Roy. Statist. Soc. Ser. B*, 59, 511-67.
- Ng, A., Jordan, M. and Weiss, Y. (2002) On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems 14*.
- Saul, L. and Pereira, F. (1997) Aggregate and mixed-order Markov models for statistical language processing. In *Proceedings of the second conference on empirical methods in natural language processing*, 81-89.

Shi, J. and Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22(8), 888-905.

Vardi, Y. and Lee, D. (1993) From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, Volume 55, Issue 3, 569-612.

Welling, M. and Weber, M. (2001) Positive tensor factorization. *Pattern Recognition Letters* 22 (12), pp. 1255-1261.

Weiss, Y. (1999) Segmentation using eigenvectors: a unifying view. In *International Conference on Computer Vision 1999*.