Reduced Rank Approximations of Transition Matrices

Juan K. Lin
Department of Statistics
Rutgers University
Piscataway, NJ 08855
jklin@stat.rutgers.edu

Abstract

We present various latent variable models for the reduced rank approximation of transition matrices. Two main categories of models, termed Latent Markov Analysis(LMA) models, are introduced. We first address the case where the transition matrix is consistent with a reversible random walk. A more general case is subsequently addressed. Iterative EM-type algorithms are presented for all models. LMA is applied to clustering based on pairwise similarities, where similarities between objects are described probabilistically. In the model, 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 a prior specification of the transition between latent cluster states is incorporated. This provides an algorithm which searches for clusters satisfying pre-specified relationships.

1 Introduction

Latent variable models fitted with the EM algorithm, along with their non-negative matrix formulations have been applied to many machine learning tasks. Vardi and Lee(1993) formulated linear inverse problems with positivity constraints probabilistically and applied it to image reconstruction in tomography and portfolio optimization. Saul et al.(1997), Lee et al.(1999) and Hofmann(2001) presented latent variable models with applications ranging from statistical language processing and image processing to information retrieval. More recently, Welling et al.(2001) presented a positive tensor factorization model which has structural similarities to a naive Bayes model. The conditional independence assumptions in these models are all expressible as graphical models. In contrast, we investigate latent variable models specified explicitly through sets of conditional independence and exchangeability assumptions. Some of the resulting models presented are not graphical in nature. EM-type iterative I-projection based algorithms are formulated for fitting the models.

In this paper we present various latent variable models, termed Latent Markov Analysis (LMA), for the reduced rank approximation of a transition matrix. The LMA model is applied to clustering based on pairwise similarities. Following the treatment in graph partitioning and spectral clustering communities (eg. Shi et al. 2000, Weiss 1999, Meila et al. 2001, Ng et al. 2002), we normalize the affinity matrix into a transition matrix, and re-formulate the clustering problem as one of finding cluster groupings with similar within cluster and between cluster transitions.

2 Latent Markov Analysis

The Latent Markov Analysis approach to finding reduced rank approximations of transition matrices is summarized 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. For added clarity, both the random variables and the corresponding simplices representing all possible distributions of the random variables are used as labels in the diagram in Fig. 1. We seek 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 states and H with k_2 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.

$$\begin{tikzcd}
Y_{\mathcal{S}^m} \arrow[r, "\tilde{p}(x|y)"] \arrow[d, "p(h|y)"'] & X_{\mathcal{S}^n} \\
H_{\mathcal{S}^{k_2}} \arrow[r, "p(g|h)"'] & G_{\mathcal{S}^{k_1}} \arrow[u, "p(x|g)"']
\end{tikzcd}$$

Figure 1: Approximate commutative diagram showing the relation between the specified high dimensional mapping from the simplex $\mathcal{S}^n$ to $\mathcal{S}^m$, and its reduced rank counterpart which maps $\mathcal{S}^{k_1}$ to $\mathcal{S}^{k_2}$.

The LMA model is applicable to the analysis of two-way contingency table data (eg. Agresti 1990) consisting of either co-occurence (Hofmann 2001) or conditional-occurence data. For co-occurence data, the data is sampled from the full joint distribution over two random variables X and X'. Conditional-occurence 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-occurence and conditional-occurence data specify the empirical joint and conditional distributions respectively. We consider the more general case of conditional-occurence data since the joint distribution fully specifies the conditional distribution.

2.1 Model and Algorithm: Reversible Case

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 is applicable to the analysis of symmetric affinity matrices. The more general model without the reversibility assumption is presented in the Section 3. 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:

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

Structurally, this model is analogous to the spectral decomposition of a symmetric matrix with non-negative constraints on the the eigenvalues and eigenvectors. Here the distributions over the hidden states $g(x|h)$ take the place of eigenvectors, while probabilities of the hidden states $p(h)$ play the role of the eigenvalues.

2.2 Numerical Examples

We apply symmetric LMA to clustering based on a pairwise affinity matrix. From a set $\mathcal{X} = \{\vec{x}_1, ..., \vec{x}_n\}$, the affinity matrix is defined as $A_{ij} = \exp\left(-||\vec{x}_i - \vec{x}_j||^2/2\sigma^2\right)$, 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 affinity matrix, $p(x_i, x'_j) = kA_{ij}$ for all $i, j \in \{1, ..., n\}$. Since the affinity matrix is symmetric, $\tilde{p}(x'|x)$ is consistent with a reversible random walk.

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 objects is in accordance with the MAP assignment of each object to the states of 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 objects.

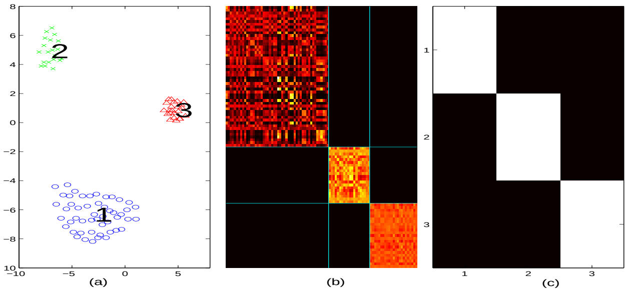

Figure 2: Latent Markov Analysis of self-transition matrix generated from the pairwise similarity matrix as described in the text. The well separated data is plotted on the left. The transition matrix, permuted in accordance with MAP assignment clustering is shown in the center. The reduced rank transition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

The 3 well separated clusters 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,

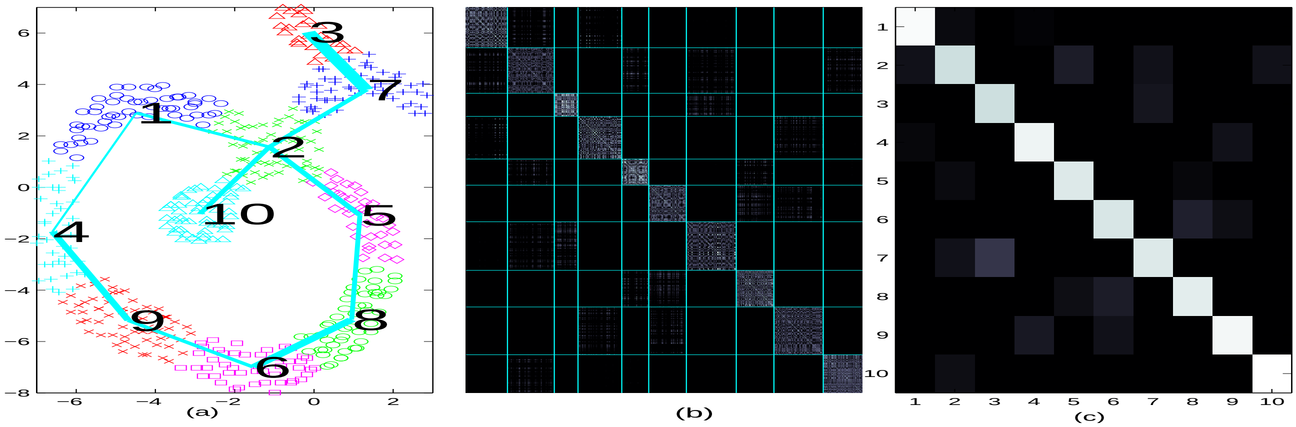

Figure 3: Latent Markov Analysis clustering based on a normalized affinity matrix. The observations are plotted on the left. The transition matrix, permuted according to LMA clustering is shown in the center. The reduced rank transition matrix is shown on the right.

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 $(p(h_i|h_j) + p(h_j|h_i))$.

It should be noted that since the relevant cost function is Kullback-Leibler information divergence instead of least squares, the reduced rank transition matrix is not simply the average of all the corresponding transition probabilities between the respective cluster observations.

3 Latent Markov Analysis: General Case

Here we present 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-occurence table $n(x|y)$, which is specifies the empirical conditional distribution $\tilde{p}(x|y)$. In the bag-of-words model for text information retrieval, this conditional-occurence 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 distribution. In Section 3.1 we present a modified EM algorithm for the fitting of an empirical conditional distribution. Section 3.2 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.

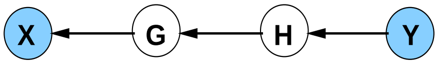

Figure 4: Directed graphical model for the general case LMA model.

3.1 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-occurence instead of a co-occurence table. Equivalently, the constraint is in the form of a specified conditional distribution $\tilde{p}(x|y)$ instead of a joint. A brief motivation of this I-projection step is given in the Appendix.

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 in Section 3.1 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 Cyclic I-projection Algorithm

Since the performance of the modified EM-algorithm was not encouraging, we pursued a more combinatorially motivated model and 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:**
  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 similar to 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, as detailed in the Appendix. Convergence of this cyclic algorithm was much faster than for the algorithm given in Section 3.1.

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 $G \perp Y|H$ and $X \perp \{H, Y\}|G$. 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.

4 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

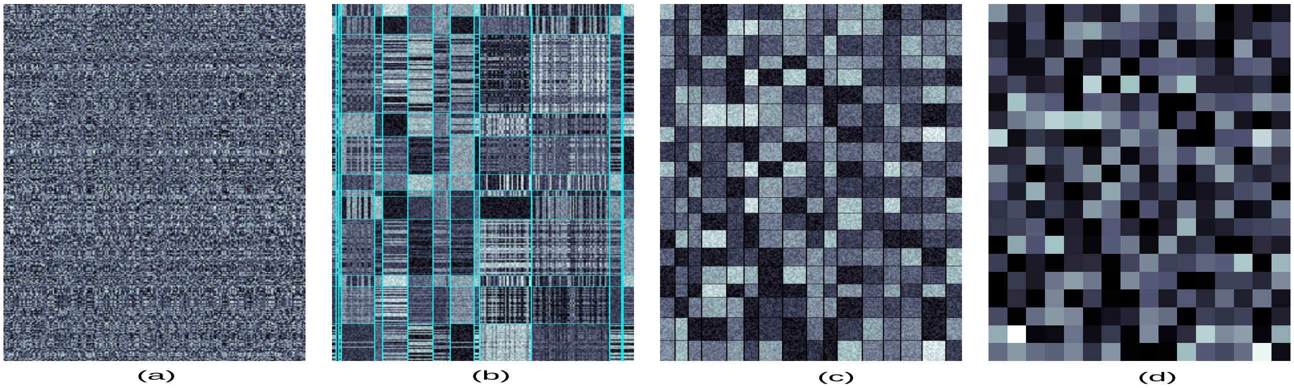

Figure 5: Latent Markov Analysis of a synthetic approximately block uniform transition matrix. (a) Specified transition matrix. (b) Re-permuted transition matrix using modified EM algorithm. (c) Re-permutation using cyclic I-projection algorithm. (d) Reduced rank transition matrix.

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 2.1, 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. 6, numerics for targeted LMA relational clustering is shown. The specified relation $\tilde{p}(h'|h)$ in Fig. 6(c) corresponds to a root cluster attached to three leaf clusters. State 1 for the latent variable corresponds to the root cluster state which has significant transition probabilities to the other three leaf cluster states. Observations in the root cluster are drawn with circles in Fig. 6(a), while the leaf cluster observations are drawn with crosses, triangles and diamonds. In the numerics, we also implemented a second stage clustering where the observations in the leaf cluster drawn with crosses were further decomposed into a root cluster surrounded by three leaf clusters. These subsequent clusters are drawn in Fig. 6(a) in various colors/grey scales, with the root cluster in black.

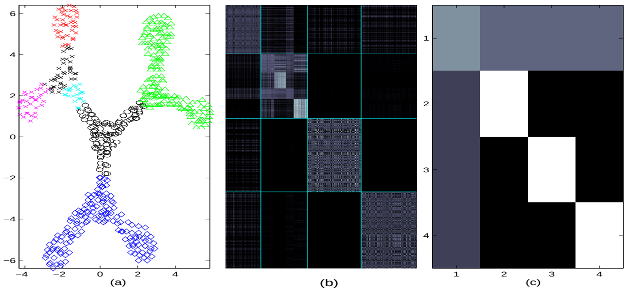

Figure 6: Hierarchical targeted relational clustering. The two-dimensional data is plotted on the left. The target reduced rank transition matrix is shown on the right. The re-permuted transition matrix is shown in the center.

5 Discussion

EM-type iterative I-projection based algorithms (Darroch et al. 1972, Csiszar 1989, Cramer 2000) are presented in this paper for the reduced rank approximation of transition matrices. 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 affinity 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. There are a few contributions of this work 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. In contrast with most other latent variable clustering models, the symmetric LMA model is not a graphical model. We also presented a *targeted* symmetric LMA model which allows for the targeted extraction of clusters with pre-specified inter-cluster relationships. For the general LMA model, we presented fast, iterative I-projection based algorithms for fitting and empirical conditional distribution.

In the LMA formalism, the reduced rank transition matrix naturally arises out of conditional independence and exchangeability assumptions. In the clustering application, since relations between clusters are again described by a transition matrix, clusters can again be clustered in the LMA formalism. The approximate commutative diagram for the latent variable reduced rank model is shown in Fig. 7. Preliminary numerical results show improved clustering performance - in particular, the ability to naturally merge clusters in the previous level of the hierarchy.

$$\begin{tikzcd}
Y_{\mathcal{S}^{n_1}} \arrow[r, "\tilde{p}(x|y)"] \arrow[d, "p(b|y)"'] & X_{\mathcal{S}^{n_2}} \\
B_{\mathcal{S}^{k_1}} \arrow[r, "p(a|b)"'] \arrow[d, "p(d|b)"'] & A_{\mathcal{S}^{k_2}} \arrow[u, "p(x|a)"'] \\
D_{\mathcal{S}^{r_1}} \arrow[r, "p(c|d)"'] & C_{\mathcal{S}^{r_2}} \arrow[u, "p(a|c)"']
\end{tikzcd}$$

Figure 7: Approximate 2-layer commutative diagram motivating a latent variable model for hierarchical clustering.

In summary, the main contributions of this paper are as follows. First, LMA provides a model and algorithm for the reduced rank approximation of transition matrices. The latent variable models are introduced using approximate commutative diagrams which detail the approximations via mappings through lower dimensional simplices. Second, the clustering formalism provided by LMA intuitively groups together states which have similar transitions to other states. Cluster relationships are directly encoded probabilistically by the reduced rank transition matrix. This provides a natural and direct representation of cluster relationships by a weighted graph. Finally, combinatorially-inspired variants of the EM algorithm in the form of iterative I-projections are introduced for LMA. The algorithms are motivated combinatorially and their performance demonstrated numerically. They motivate further research into enhanced convergence properties of combinatorial I-projection based algorithms.

6 Appendix: Modified E-step I-projection

Instead of the conventional E-step I_1-projection consisting minimization with respect to the first argument of the Kullback Liebler information divergence, we use the I_2-projection, minimizing with respect to the second argument. These projections are equivalent when constraints are in the form of specified marginals but different for specified conditional distributions (Cramer 2000).

Let p and q be two joint distributions over the random variables X, Y, Z. Consider the problem of minimizing $D(p||q)$ with respect to q, subject to the constraint $q(y|z) = \tilde{w}(y|z)$. Writing $p = p(z)p(y|z)p(x|y,z)$ and $q = q(z)\tilde{w}(y|z)q(x|y,z)$, we have

$$\begin{aligned}
& D(p||q) \\
= & \int p \log(\frac{p}{q}) dxdydz \\
= & \int p(z) \log(\frac{p(z)}{q(z)}) dz + \int p(z)p(y|z) \log(\frac{p(y|z)}{\tilde{w}(y|z)}) dydz \\
& + \int p(z)p(y|z)p(x|y,z) \log(\frac{p(x|y,z)}{q(x|y,z)}) dxdydz. \\
= & D(p(z)||q(z)) + D(p(y,z)||\tilde{w}(y|z)p(z)) \\
& + D(p(x,y,z)||q(x|y,z)p(y,z))
\end{aligned}$$

All three terms above are non-negative, and only the first and third terms have a dependence on q. Those terms vanish when $q(z) = p(z)$ and $q(x|y,z) = p(x|y,z)$. These are the conditions which minimize $D(p||q)$, with $min(D(p||q)) = D(p(y,z)||\tilde{w}(y|z)p(z))$.

References

[1] 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.

[2] 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.

[3] Lee, D. and Seung, S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(675), 788-791.

[4] Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196.

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

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

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

[8] Meila, M. and Shi, J. (2001) A random walks view of spectral segmentation. in *Proc. International Workshop on AI and Statistics (AISTATS), 2001*

[9] Ng, A., Jordan, M. and Weiss, Y. (2002) On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems 14.*

[10] Agresti, A. (1990) Categorical Data Analysis. New York: Wiley.

[11] Hofmann, T. and Puzicha, J. (1998) Statistical Models for Co-occurrence Data. Technical Report, Artificial Intelligence Laboratory Memo AIM-1625.

[12] Darroch, J. and Ratcliff, D. (1972) Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, Vol.43, No.5, 1470-1480.

[13] Csiszar, I. (1989) A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. *Annals of Statistics*, Vol.17, No.3, 1409-1413.

[14] Cramer, E. (2000) Probability measures with given marginals and conditionals: *I*-projections and conditional iterative proportional fitting. *Statistics and Decisions* 18, 311-329.

[15] Friedman, N., Mosenzon, O., Slonim, N. and Tishby, N. (2001) Multivariate information bottleneck. *Uncertainty in Artificial Intelligence 2001.*