

A Scalable Generative Topographic Mapping for Sparse Data Sequences

Ata Kabán

School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, UK

E-mail: A.Kaban@cs.bham.ac.uk

Abstract

We propose a novel, computationally efficient generative topographic model for inferring low dimensional representations of high dimensional data sets, designed to exploit data sparseness. The associated parameter estimation algorithm scales linearly with the number of non-zero entries in the observations while still learning a truly nonlinear generative mapping. The latent variables of the model lie in a 2D space that can be used for visualisation. We discuss related work and we provide experimental results on text based documents visualisation as well as the exploratory analysis of web navigation sequences.

1 Introduction

Understanding high dimensional data through manifold learning has received a great deal of interest due to its practical importance in many fields. Methods fall in two main categories: density based approaches and spectral methods. While density-based approaches potentially provide a flexible and technically sound framework that can be employed in suitably formulated manifold learning models [2, 7, 9], much of the recent advances have been concentrating on non-probabilistic formulations, such as spectral methods [1], due to their appealing computational advantages. However, the locally linear structural assumption and moreover the lack of a clear density formulation deprives these methods from the flexibility of probabilistic model formulations [9, 2]. For example, many timely applications, ranging from text and user modelling [3, 6, 5] to various scientific data mining problems, involve types of data other than real valued vectors. In the simplest case these may take the form of histograms over a set of indicators, although more structured observation objects that may have their own dependency structure [5] are also possible. In such cases the application of non-probabilistic manifold learning methods is problematic. However, to date, the computational com-

plexity of existing density-based approaches [2, 9, 7] makes them impractical to use with large amounts of high dimensional data. This weakness is more pronounced in the case of non-Gaussian noise models [2, 7] such as those that are statistically appropriate for symbolic sequences.

In this paper, we propose a novel, computationally efficient generative model for inferring a 2D representation of high dimensional data sets, designed to scale to large data sets by exploiting data sparseness. The associated parameter estimation algorithm scales linearly with the number of non-zero entries in the observations while still learning a truly nonlinear generative mapping. We briefly discuss the relation of the proposed approach to previous work. Experiments on both static text collections and a large dynamic web browsing behaviour data set indicate that in addition to its advantages in visualisation and data exploratory tasks, the proposed approach is also competitive as a predictive model, in comparisons with related methods.

2 The model

Consider a set of independent (unnormalised) histograms with the n -th instance denoted as d_n . Our aim is to devise a both principled and scalable method for representing this data in 2D in a maximally information-preserving manner. As in [2, 7], we begin with defining a uniform prior density over a latent space $x \in [-1, 1]^2$ and define a generative model as the inverse mapping of x through a set of smooth nonlinear basis functions $\{\phi_l(\cdot)\}_{l=1:L}$. Contrarily to [2, 7] however, our basis functions are designed to perform the transformation from the Euclidean latent space into the space of L -dimensional multinomial distributions, i.e. the $(L - 1)$ -dimensional simplex. This can be achieved simply as the following:

$$\phi_l(x) = \frac{\exp(-\frac{1}{2\sigma^2}|x_l - x|^2)}{\sum_{l'} \exp(-\frac{1}{2\sigma^2}|x_{l'} - x|^2)} \quad (1)$$

where $x_l, l = 1 : L$ lie on a regular grid in $[-1, 1]^2$ and σ is set to twice the smallest distance between two neighboring

points \mathbf{x}_l .

Then we define the log conditional probability of a datum instance as a parameterised multinomial:

$$\log p(\mathbf{d}_n|\mathbf{x}) = \sum_t d_{tn} \log \sum_l a_{tl} \phi_l(\mathbf{x}) \quad (2)$$

where $a_{tl} \geq 0, \forall t = 1 : T, l = 1 : L$ and $\sum_t a_{tl} = 1, \forall l = 1 : L$. Up to a constant, (2) is equivalent to the negative Kullback-Leibler divergence between the observation \mathbf{d}_n and the model $\sum_l a_{tl} \phi_l(\mathbf{x}) = \mathbf{a}_t \phi(\mathbf{x})$. So the parameter vectors \mathbf{a}_l define projections of the data \mathbf{d}_n , in the Kullback-Leibler sense, onto the information space spanned by the basis set $\{\phi_l\}_{l=1:L}$, and these will be estimated from the data through the generative model.

As mentioned, a uniform prior is desirable over \mathbf{x} . It is convenient to discretise the latent space into a regular grid of K points $\mathbf{x}_1, \dots, \mathbf{x}_K$, in which case the latent prior becomes a mixture of Dirac delta functions $p(\mathbf{x}) = \frac{1}{K} \sum_k \delta(\mathbf{x} - \mathbf{x}_k)$.

Integrating over \mathbf{x} , the model likelihood is now the following.

$$p(\mathbf{d}_n) = \frac{1}{K} \sum_k \prod_t \left\{ \sum_l a_{tl} \phi_l(\mathbf{x}_k) \right\}^{d_{tn}} \quad (3)$$

To maximise (3) with the parametrisation proposed, an auxiliary function Q is constructed in the standard way (by employing Jensen's inequality).

$$Q = \sum_n \sum_k r_{kn} \sum_t d_{tn} \log \sum_l a_{tl} \phi_{lk} \quad (4)$$

This is then optimised subject to the mentioned constraints (using Lagrange's method) to provide the generalised E-M algorithm below.

$$\mathbf{R} \propto \exp \left\{ \log(\mathbf{A}^{(old)} \Phi)^T \mathbf{D} \right\} \quad (5)$$

$$\mathbf{A}^{(new)} \propto \mathbf{A}^{(old)} \odot \left\{ [\mathbf{D} \mathbf{R}^T] \oslash [\mathbf{A}^{(old)} \Phi] \right\} \Phi^T \quad (6)$$

Here, \odot denotes element-wise matrix multiplication and \oslash denotes element-wise division and the matrix \mathbf{R} is formed from the posterior probabilities $r_{kn} = P(\mathbf{x}_k|\mathbf{d}_n)$.

A strength of this approach is that the resulting algorithm can exploit data sparseness: If \mathbf{D} is sparse, then the matrix multiplication in the numerator takes $\mathcal{O}(PK)$ where P denotes the number of nonzero elements in \mathbf{D} and K is the number of samples used for approximating the uniform latent space. The remaining matrix multiplications are then able to exploit the sparsity of Φ . As in [2, 7], the posterior means $E[\mathbf{x}|\mathbf{d}_n] = \sum_k \mathbf{x}_k r_{kn}$ are then employed to obtain 2D visualisation plots.

There are two equivalent views on the generation process associated with this model, both are quite intuitive. These

highlight unified advantages of the proposed approach over previous work. In the first view, \mathbf{d}_n is generated as follows: One latent point \mathbf{x} is first generated with uniform probability from the bounded Euclidean latent space. This is then nonlinearly projected into a point from an $(L - 1)$ -dimensional simplex. Finally, this undergoes a convex linear transformation to generate the observed histogram. This interpretation is closer to the Generative Topographic Mapping (GTM) [2], essentially transposed from an Euclidean space into the space of probability distributions. It has the advantage of formulating an explicit continuous nonlinear mapping from a continuous Euclidean latent space to a probability space. However, unlike our algorithm, in GTM for discrete data [2, 7] we are required to solve a nonlinear matrix equation in each M-step, without the possibility of taking advantage of the data sparseness.

In a second view, we can follow the generation process on the level of individual symbols that make up the observed histogram: A latent point index k is first generated uniformly with probability $1/K$. This is global for the whole sequence. Then for each individual symbol, different situated component-indices l are generated with probability defined by the neighbourhood definition $p(l|k) = \phi_l(\mathbf{x}_k)$. Finally, a symbol t is generated with probability $a_{tl} = p(t|l)$. This view is closer to the aspect models or simplicial mixtures [3, 6], which are powerful in high dimensions and scale nicely for symbolic data.

2.1 Prototypes versus aspects in the model

From (3), the formulated model can essentially be seen as a constrained mixture, where the mean parameter of the k -th mixture component is the following.

$$\mathbf{m}_{tk} = \sum_l a_{tl} \phi_l(\mathbf{x}_k) \quad (7)$$

Indeed, these are proper distributions, due to the constraints we imposed, now we have that $\sum_t m_{tk} = 1$. Thus, once the parameters are estimated, prototypical representations as local averages can be obtained from \mathbf{m}_k . As an illustration, the left hand plots of Figure 1 show the prototypes (analogous to reference-vectors in the SOM [8]) created by the proposed algorithm from a set of grey-scale face images [9]. Here each image is taken as an unnormalised histogram over the pixel locations. A 10×10 latent grid has been utilised and subsequently sub-sampled to show each third prototype. The right-hand plots of Figure 1 show the parameters \mathbf{a}_l of the model (sub-sampled from an $L = 5 \times 5$ grid). These are also probability distributions, somewhat analogous to parameters of the so-called aspect models [6, 3]. They are much sparser compared to \mathbf{m}_k -s, and in the face image example, they seem to retain the main characteristics of the face expressions only. To see why this is so, we

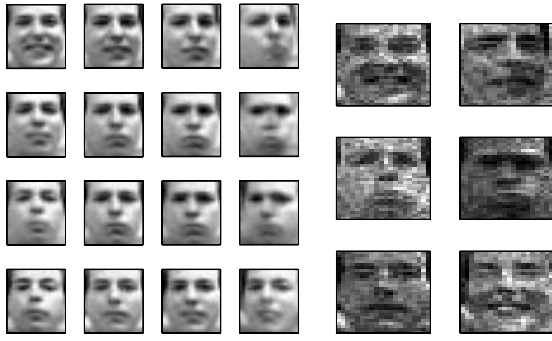


Figure 1. Prototype versus aspect views.

analyse the auxiliary function of the likelihood that is maximised.

Let us observe that in addition to the auxiliary function (4), the following is also an auxiliary function of the same likelihood:

$$Q' = \sum_n \sum_t \sum_k \sum_l r_{kn} r_{lkt} d_{tn} \log a_{tl} \phi_{lk} \quad (8)$$

simply obtained by applying Jensen's inequality to Q in (4), where $r_{lkt} \propto a_{tl} \phi_{lk}$ is an exact posterior. The maximisation of (8) leads to the update equation below.

$$a_{tl} \propto \sum_n \sum_k r_{kn} r_{lkt} d_{tn} \quad (9)$$

This of course, after some manipulations, can be rewritten in the form given in (5)-(6). However, by replacing (9) into (8), the term that depends on \mathbf{A} , becomes the following.

$$Q'_{\mathbf{A}} = \sum_l \sum_t w_l a_{tl} \log a_{tl} = - \sum_l w_l H[a_l] \quad (10)$$

where the weights $w_l = \sum_n \sum_t \sum_k r_{kn} r_{lkt} d_{tn}$ denote the expected total number of symbols that are explained by the l -th aspect. In other words, it turns out that the weighted sum of entropies of \mathbf{a}_l is minimised. So it is now clear that the proposed approach provides both local averages \mathbf{m}_k and low entropy aspects \mathbf{a}_l and they are linked nonlinearly by the neighbourhood-encoding relation ϕ_{lk} .

3 Topical analysis of text

Organising text based repositories is an important practical issue. The vector-space representation of text over a dictionary of words typically provides high dimensional and substantially sparse histogram data to deal with [6, 3, 7]. The use of probabilistic topical manifold modelling has previously considered in [7, 6].

In the first experiment we have taken 10,000 text documents from the benchmark Newsgroups collection, over a dictionary of 9,496 terms that remained after removing terms that occurred less than 10 times, removing stop words and using standard stemming preprocessing performed with the freeware RainBow utility. The visualisation result, as produced by the proposed method, selected according to the highest out of sample log likelihood is shown on Figure 3. The data has been split in two halves, the first half

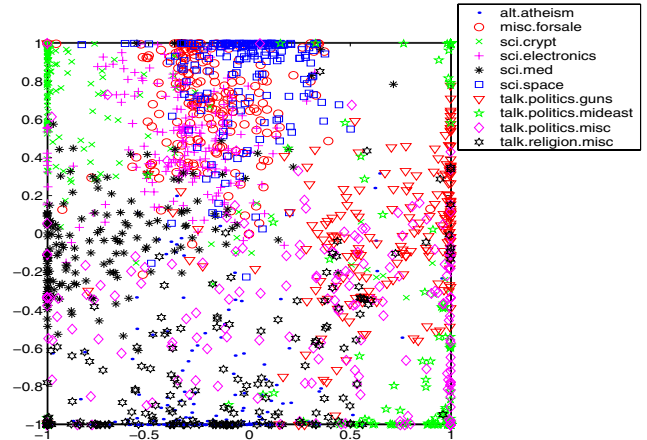


Figure 2. Posterior mean mapping of the 10 Newsgroups text collection obtained with the proposed method.

being used for training while the second half served as an independent test set. The model complexity L has been varied between 9 and 100 while keeping a fixed resolution at a 10×10 regular grid over the latent space. To alleviate the effects of being trapped in local optima, 10 independent, randomly initialised runs have been performed for each of these model settings and only the run that achieved the highest local optimum has been retained for measuring and reporting the out of sample log likelihood on the test set. These values are shown on Figure 3, comparatively with the out of sample log likelihoods obtained from the multinomial PCA [4], also known as the Latent Dirichlet Allocation (LDA) [3], which is one of very few existing dimensionality reduction methods for multinomial data that is also scalable. In both models, the model order refers to the number of aspect components. From this plot, it is clear that the proposed method achieves a comparable performance to multinomial PCA on this data set. As expected, the optimal model order is higher for the topographic model, since it is a constrained model.

Parameter interpretability is now demonstrated on the same experiment. On a lower level, we can list the most

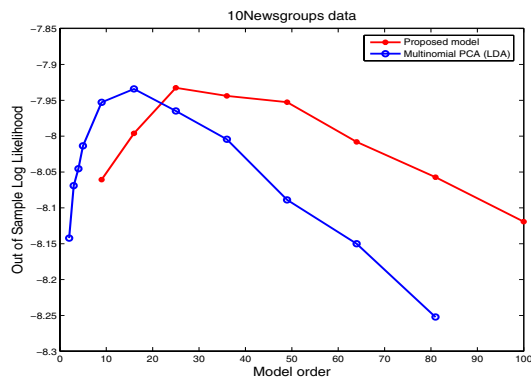


Figure 3. Out of sample log likelihood

probable words associated to each point on the 2D latent space — these are local averages. On a higher level, the list of the most probable words associated with each nonlinear aspect (the discrete index of which is now an explanatory variable of the model) can be computed. Both these levels have provided a coherent topical organisation. The lists of the most informative aspect-level words, as constructed from the parameters a_l are provided in Table 3. Each cell of the table lists the first ten most probable words in that aspect. The topographic organisation in terms of topics is most apparent.

| | | | | |
|---|---|--|---|--|
| kei, encrypt, secur, privati, de, public, escrow, algorithm, chip, govern | system, comput, technologi, program, includ, softwar, data, list, inform, version | space, orbit, nasa, launch, earth, moon, mission, center, lunar, flight | work, year, peopl, russian, land, world, studi, soviet, nation, page | armenian, turkish, armenia, turk, greek, genocid, turkei, soviet, muslim, azerbaijan |
| kei, anonym, chip, govern, protect, user, law, copi, number, clipper | mail, system, phone, program, work, inform, data, time, compani, email | ca, work, sale, sell, price, drive, write, offer, power, interest | time, don, year, peopl, make, write, rate, thing, articl, dai | gun, peopl, db, kill, govern, crimin, crime, handgun, mov, murder |
| wire, patient, medic, effect, diseas, servic, ground, peopl, doctor, drug | write, articl, work, don, year, time, ve, good, apr, post | write, articl, thing, apr, don, ve, good, ac, mine, time | fire, articl, write, don, make, peopl, ga, time, state, thing | law, gun, weapon, govern, arm, drug, polic, firearm, state, constitut |
| insur, health, food, dr, result, report, msg, state, studi, human | articl, write, peopl, don, make, time, good, homosexu, person, question | write, articl, don, peopl, claim, post, make, point, thing, love | peopl, fbi, write, koresh, articl, don, make, bd, start, apr | state, govern, american, nation, countri, war, polit, senat, hous, year |
| parti, abort, frank, convent, valu, polit, dwyer, object, price, state | moral, object, law, scienc, system, human, truth, word, de, life | god, christian, exist, religion, jesu, atheist, belief, argum, faith, bibl | peopl, jew, kill, jewish, peopl, live, world, work, question, nazi, don | israel, isra, arab, presid, myer, jew, palestinian, mr polici, lebanes |

Table 1. The aspect-level semantic space extracted from 10 Newsgroups, as represented by ordered lists of the most probable words that characterise equidistant locations of the latent space

The computation time requirements are now assessed. Subsets of the 20Newsgroups collection have been created with increasing size and the CPU time has been measured against the number of non-zero entries in the data. Figure 4

shows the results comparatively, for our topographic model, multinomial PCA and multinomial GTM. In all cases, two model orders (9 and 64) have been tested.

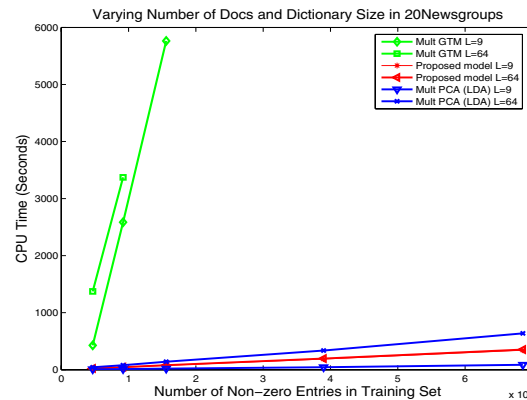


Figure 4. Comparative assessment of computation time requirements.

As expected, our method, similarly to multinomial PCA, scales indeed linearly with the number of non-zero entries in the data. The plot also shows that there is little increase in computation time when increasing the model order L in our model. This is because there is little variation in the level of sparsity of the matrix Φ when L is varied. Overall, at both model orders considered, the scaling of our method lies between that of multinomial PCA with 9 and 64 aspects respectively. We can also see from the plot, as expected, that multinomial GTM is computationally infeasible on large data sets, as it is unable to take advantage on the sparsity of the data.

4 Exploratory analysis of web navigation sequences

The organisation and exploratory analysis of dynamic behaviour of individuals in the context of web environments is a major challenge for automated data analysis research. Such investigations are quite recent [5] and motivated by the availability of vast quantities of user traces together with the need for creating predictive profiles as well as creating tools that allow e.g. a site administrator to explore large sets of navigation sequences. The possibility of visual exploration in this context has been proposed in [5], where an approach employing mixture based clustering of first order Markov chains has been explored.

However, in a mixture model, the relation between clusters is not modelled, and in the case of a large site collection, with several thousands of browsing users, it would be impractical to expect the site administrator to examine all

clusters individually in order to obtain an overview of the ongoing activity or locate behaviours of interest. In addition, browsing behaviours that are common to all clusters of users — and therefore uninteresting — will end up being present on all cluster prototypes, which makes the visual analysis difficult. Indeed, in the mentioned work, such problems have been noticed and the ad-hoc constraint of fixing the initial state in each cluster has been proposed in order to aid visual inspection. This however came at the expense of a suboptimal predictive model as reflected by the out of sample log likelihood.

Here we investigate our approach for organising the same set of web navigation sequences¹ as those used in [5]. The topographical principle induced, originally proposed by Kohonen [8], provides a proximity constraint that has proved intuitive and useful in hundreds of applications in the past [8]. Indeed, our eyes is sensitive not just to individual colours but also to reasonably low-entropy patterns or textures, therefore our hope is that visualising temporal activity in terms of proximity structures may be useful.

In addition to data explanatory capabilities, we also aim at a good predictive model, therefore we begin with assessing the generalisation performance of our model in terms of the out of sample log likelihood. The training set consisted of 100,000 sequences drawn at random from the entire data set, totalling 801,745 page requests and an independent test set consisted of 88,181 sequences totalling 714,280 page requests. All sequences share a common state space of 17 page categories. These are: 'frontpg', 'news', 'tech', 'local', 'opinion', 'onair', 'misc', 'weather', 'msnnews', 'health', 'living', 'business', 'msnsport', 'sports', 'summary', 'bbs' and 'travel'. The vast majority of these sequences is very short so that a simplicial mixture of Markov chains, that would try to model interleaving dynamic primitives, overfits immediately. Figure 5 shows the out of sample log likelihood as obtained on the independent test set comparatively for the proposed model, the constrained mixture of [5], an unconstrained mixture and a baseline global first order Markov model. Clearly, our method outperforms the mixture with fixed initial states and approaches an unconstrained mixture in predictive performance. We can thus be confident that the advantages of our model in terms of visualisation and parameter interpretability do not produce a limitation of its predictive power on this data.

Figure 6 shows a fragment (three columns) of the aspect-level representation created. On the right, the probability transitions associated with the aspects are shown. As we know, these are low-entropy prototypes of behaviour, different from cluster centres or local averages. On the left, the top matching 15 actual user sequences are listed for each aspect. Although the full grid of aspects is not shown, we can still follow a gradual shift of interest on this fragment

¹<http://kdd.ics.uci.edu/databases/msnbc/msnbc.html>

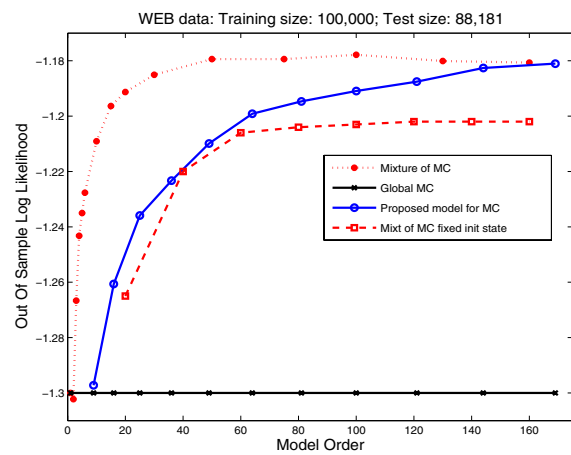


Figure 5. Out of sample log likelihood against model order as obtained on the msnbc.com web navigation data set.

of the representation. E.g. from top toward the bottom, the strong interest in the 'frontpage' of the site (1-st page category) shifts through a repetitive user behaviour toward a pronounced interest in 'news' (2-nd page category), corroborated with a more dynamic browsing activity. On the horizontal axis in the first row the interest shifts from 'frontpage' to 'sports' and 'health'. These trends can be more easily followed by looking at the transition plots. However, from the listing of the actual sequences we can see how represented, how homogeneous or inhomogeneous these aspects are, we can recognise groups of similar behaviours by the specific combinations of patterns and colours. The topographic organisation is most apparent in both views.

It is also worth noticing that, as expected, the prototype-level transition behaviours are far not as informative in the case of browsing behaviours. This is simply because behavioural patterns that are common to all clusters appear on all prototypes, making it difficult to distinguish the distinctive features. It is a unique feature of the proposed model that it is able to produce these low-entropy projections (aspects) of the data, simultaneously with a 2D nonlinear projection, while additionally also being a competitive density estimator, able to generalise on new, previously unseen data (as demonstrated by the out of sample likelihood values).

Finally, our model also induces probabilistic profiles for each individual sequence, in the form of two posterior distributions (over the grid of aspects and over the grid of mixture prototypes). Figure 7 shows three examples. These could be used to understand individual user interests as well as relationships between users. In the first example, a fairly long activity sequence induces a relatively sharp cluster posterior. In the second example we have a multi-modal profile

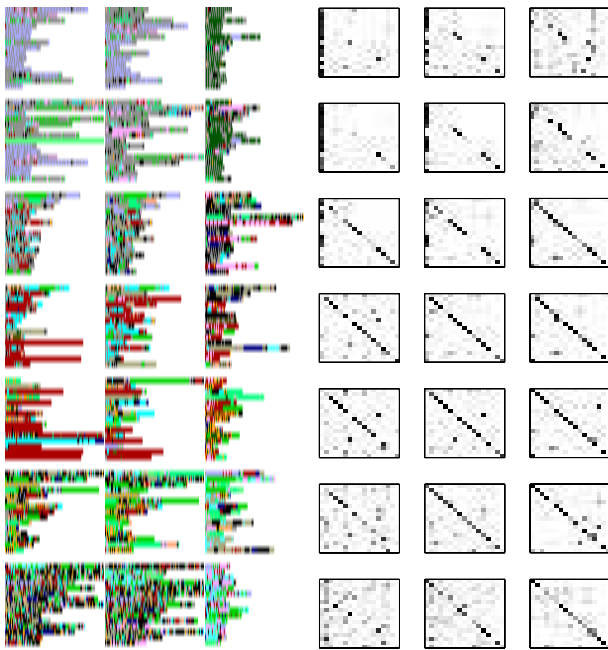


Figure 6. A fragment from an aspect-level topographical display of the msnbc.com web navigation data set. *Left:* The top matching sequences for each aspect. Each row is a sequence, colors encode page categories. *Right:* Transition probabilities of the same aspects; darker means higher probability.

that encodes more interesting relationships based on multiple interests. In both these examples, the aspect posterior produces smoothed versions of the cluster posterior. In the third example in turn we have a very common activity, therefore the cluster posterior is very broad. By contrary, the aspect posterior shrinks, as the aspect parameters tend to separate common behaviours into fewer aspects.

5. Conclusions

This paper has presented a both computationally efficient and technically principled generative model for the analysis of sparse sequences. This has been demonstrated in realistic applications.

References

[1] Y. Bengio, J-F. Paiement, and P. Vincent, Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. Neural Information Processing Systems 16 (NIPS'03), 2003.

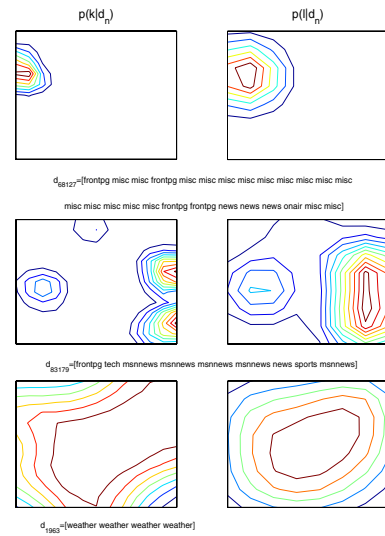


Figure 7. Three individual user profiles over clusters (left) and aspects (right) respectively.

[2] C.M. Bishop, M. Svensen, and C.K.I. Williams, Developments of the Generative Topographic Mapping, Neurocomputing, vol. 21, pp. 203-224, 1998.

[3] D.M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, 3(5):993-1022, 2003.

[4] W. Buntine, Variational Extensions to EM and Multinomial PCA, ECML 2002.

[5] I. Cadez, D. Heckerman, C. Meek, P. Smyth and S. White, Model-based Clustering and Visualisation of Navigation Patterns on a Web Site, Data Mining and Knowledge Discovery, 7(4), 2003.

[6] T. Hofmann. Probmap - a probabilistic approach for mapping large document collections. Journal for Intelligent Data Analysis, 4:149-164, 2000.

[7] A. Kabán and M. Girolami, A Combined Latent Class and Trait Model for the Analysis and Visualisation of Discrete Data, IEEE Transactions on Pattern Analysis and Machine Intelligence 23(8), pp. 859-872, 2001.

[8] S. Kaski, J. Kangas, and T. Kohonen. Bibliography of Self-Organizing Map (SOM) Papers: 1981- 1997. In Neural Computing Surveys, volume 1, pages 102-350, 1998.

[9] S. Roweis, L.K. Saul and G. Hinton, Global Coordination of Local Linear Models, NIPS vol. 14, pages 889-896, 2002.