Data Mining and Knowledge Discovery, 10, 175–196, 2005

Sequential Activity Profiling: Latent Dirichlet Allocation of Markov Chains

MARK GIROLAMI girolami@dcs.gla.ac.uk
Bioinformatics Reseach Centre, Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK

ATA KABÁN a.kaban@cs.bham.ac.uk
School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK

Editor: Usama Fayyad

Received July 15, 2003; Accepted January 27, 2005

Abstract. To provide a parsimonious generative representation of the sequential activity of a number of individuals within a population there is a necessary tradeoff between the definition of individual specific and global representations. A linear-time algorithm is proposed that defines a distributed predictive model for finite state symbolic sequences which represent the traces of the activity of a number of individuals within a group. The algorithm is based on a straightforward generalization of latent Dirichlet allocation to time-invariant Markov chains of arbitrary order. The modelling assumption made is that the possibly heterogeneous behavior of individuals may be represented by a relatively small number of simple and common behavioral traits which may interleave randomly according to an individual-specific distribution. The results of an empirical study on three different application domains indicate that this modelling approach provides an efficient low-complexity and intuitively interpretable representation scheme which is reflected by improved prediction performance over comparable models.

Keywords: Markov chains, mixture models, user profiling

1. Introduction

The now commonplace ability to accurately and inexpensively log the activity of individuals in a digital environment makes available log files of user activity which may be employed in characterizing individual specific behavioral profiles. To achieve this it is necessary to induce space efficient representations, or profiles, of individuals from the available traces of each individuals logged activity. Most often, such recordings take the form of streams of discrete symbols ordered in time, for example a web-site log-file stores the time-ordered sequence of site specific web-pages (a finite set) visited by individuals having entered the site.

The modelling of time dependent sequences of discrete symbols from a dictionary $\mathcal{S}$ employing m'th order Markov chains (typically $m = 1$ or $m = 2$) has been extensively studied in a number of domains, most notably in statistical language modelling (Manning and Schütze, 1999). Recent attention has turned to modelling web browsing behavior (Borges and Levene, 1999; Deshpande and Karypis, to appear; Anderson et al., 2001) and somewhat related, web-page pre-fetch prediction (Sarukkai, 2000) as well as bio-sequence analysis

(Krogh,). If there is a requirement to capture long range temporal dependencies within the sequences observed then a higher-order m'th order Markov model can be employed, however these suffer from an $\mathcal{O}(|\mathcal{S}|^{m+1})$ growth in the number of model parameters which require to be estimated from the available data. In statistical language modelling the size of the symbol dictionary (number of unique words in language) may be of the order of tens of thousands elements. As an example if there are 5×10^3 unique words defined in the language (a very small number in language modelling terms), reliably estimating the $(5 \times 10^3)^3 = 12.5 \times 10^{10}$ parameters which define a 2'nd order Markov model becomes a formidable challenge. Many methods have been developed to effectively deal with this exponential rise in the number of free parameters, such as linearly interpolating higher order with lower order models (Manning and Schütze, 1999). In addition approximating the full long term dependencies with linear mixtures of pairwise lower order transition models have been developed in Raftery (1985) and subsequently employed in Saul and Pereira (1997) and Saul and Jordan (1999). Other approaches to capturing longer term dependencies have been presented in Tino and Dorffner (2001) and functions of first order Markov chains such as the Hidden Markov Model (HMM) (Rabiner, 1989; Krogh,) successfully capture longer term dependencies, however, inevitably these come with an increased computational cost.

The representation provided by such models is global in the sense that a single monolithic generating process is assumed to underlie all observed sequences. However, to capture the possibly heterogeneous nature of a set of observed sequences a model with a number of differing generating processes needs to be considered. This is particularly important in user or customer behavior modelling, where the sequential activity of a number of individuals within a group needs to be efficiently modelled. Indeed the notion of a heterogeneous population, characterized for example by occupational mobility and consumer brand preferences, has been captured in the *Mover-Stayer* model (Frydman, 1984). This model is a discrete time stochastic process that is a two component mixture of first-order Markov chains, one of which is degenerate and possesses an identity transition matrix characterizing the *stayers* in the population. The original notion of a two-component mixture of Markov chains has recently been extended to the general form of a mixture model of Markov chains in Cadez et al. (2003). The main motivation in developing this mixture model was the visualization of the class structure inherent in the browsing patterns of visitors to a commercial web-site (Cadez et al., 2003). In such a mixture representation each class of users is characterized by their shared common prototypical behavior, and therefore such mixture models will not be appropriate for identifying the shared behavioral patterns which are the basis of multiple relationships between users and groups of users which may yield a more realistic model of the behaviors exhibited by the population as a whole.

In this paper we propose a dynamic user[1] model, for individuals within a group, that explicitly captures the assumption that there exists a common set of behavioral traits which can be estimated from all observed user activity. In addition each user is defined by a personalized distribution of the probability of exhibiting these traits and each of these forms the individual user profiles within the group. This is a computationally attractive model, as relatively simple structural characteristics may be assumed at the generative level. For example consider a small set of simple first-order Markov Chains (MC) which combine to generate sequences by interleaving in various proportions of participation. Clearly the sequences that result from this combined interleaving will be more complex than sequences generated by any of the chains taken individually. This is the case as the overall sequences may exhibit transitions that are present in any of the available generators in the set.

We employ this construction as a generative model to 'explain' complex heterogeneous user behavior of a number of individuals in terms of a compact set of structurally simple common behavioral patterns along with their user-specific proportions of participation (interpreted as the users' individual profiles over the basis set) and propose to estimate both of these from sets of user trace recordings. This is much more parsimonious than creating separate models for each individual, a task which may be beset with statistical estimation problems if there is only a small amount of available logged activity for an individual. At the same time such a representation can possibly account for more complex behavior at the level of each individual than any single global model of the same order. The resulting model is thus a distributed dynamic model which represents an effective tradeoff between individual-specific and a general group-level behavior model.

The technical aspects of defining such a model, benefit from the recent developments in distributed parts based modelling of static vectorial data (Lee and Seung, 2001; Ross and Zemel, 2003; Hofmann, 2001; Blei et al., 2003; Minka and Lafferty, 2002; Hofmann, 2001), with various applications including image decomposition (Lee and Seung, 2001), document modelling, information retrieval (Hofmann, 2001; Blei et al., 2003; Minka and Lafferty, 2002) and collaborative filtering (Hofmann, 2001). The consistent generative semantics of the recently introduced latent Dirichlet allocation (LDA) (Blei et al., 2003) will be adopted and by analogy with (Minka and Lafferty, 2002) the resulting model will be referred to as a simplicial mixture of Markov chains. A somewhat related idea of decomposing event sequences has been proposed within the independent component analysis framework in Mannila and Rusakov (2001), however the independence assumption is not made here.

2. Simplicial mixtures of markov chains

We define a sequence of L symbols $s_L, s_{L-1}, \ldots, s_1, s_0$, such the symbol emitted at time t is s_0, symbol s_1 is emitted at the previous time $t-1$ and s_L is observed at time $t = 0$. This sequence of symbols, denoted by $\mathbf{s}$, can be generated from a dictionary $\mathcal{S}$ by an m'th order discrete time invariant Markov chain k which has initial state probability $P_1(k)$ and has $|\mathcal{S}|^{m+1}$ state transition probabilities denoted by $T(s_m, \ldots, s_1 \to s_0 \mid k)$. The number of times that the symbol s_0 follows from the state defined by the m-tuple of symbols $s_m, \ldots, s_1$ within the sequence is given as $\mathcal{N}(s_m, \ldots, s_1 \to s_0)$ and so the probability of the sequence of symbols under the k'th Markov process of order m is $P(\mathbf{s} \mid k) = P_1(k) \prod_{s_m=1}^{|\mathcal{S}|} \cdots \prod_{s_0=1}^{|\mathcal{S}|} T(s_m, \ldots, s_1 \to s_0 \mid k)^{\mathcal{N}(s_m,\ldots,s_1 \to s_0)}$. We employ `Start` and `Stop` states in each symbol sequence $\mathbf{s}_n$ and incorporate the initial state distribution of the `Start` state as the transition probabilities from this state within the $|\mathcal{S}|^m \times |\mathcal{S}|$ dimensional state transition matrix T_k. We denote the set of all state transition matrices $\{\mathrm{T}_1, \ldots, \mathrm{T}_k, \ldots, \mathrm{T}_K\}$ as $\mathbf{T}$. Suppose that we are given a set of symbolic sequences $\{\mathbf{s}_n\}_{n=1:N}$ over a common finite state space, each having different length L_n. In contrast to cluster models for sequences which try to model inter-sequence heterogeneities, our intuition is that in sequences over a common finite state space, provided they are sufficiently long it is sensible to look for several randomly interleaved generating processes, some of which might be common to several sequences. To account for this idea, we will adopt the LDA (Blei et al., 2003) modelling strategy. We will employ general m'th-order Markov models here, however other appropriate models would equally be possible to assume. The complete generative semantics of LDA allows us to describe the process of sequence generation where mixing components $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_k, \ldots, \lambda_K]$ are K-dimensional Dirichlet random variables and so are drawn from the $K-1$ dimensional simplex defined by the Dirichlet distribution $\mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\alpha})$ with parameters $\boldsymbol{\alpha}$. These are then combined with the individual state-transition probabilities T_k, which are model parameters to be estimated, and yield the symbol transition probabilities $T(s_m, \ldots, s_1 \to s_0 \mid \boldsymbol{\lambda}) = \sum_{k=1}^{K} T(s_m, \ldots, s_1 \to s_0 \mid k)\lambda_k$. The overall probability for a sequence $\mathbf{s}_n$ under such a mixture, which we shall now refer to as a simplicial mixture of Markov chains (Minka and Lafferty, 2002), denoted as $P(\mathbf{s}_n \mid \mathbf{T}, \boldsymbol{\alpha})$ is equal to

$$\int_{\triangle} P(\mathbf{s}_n \mid \mathbf{T}, \boldsymbol{\lambda})\mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\alpha})d\boldsymbol{\lambda} = \int_{\triangle} d\boldsymbol{\lambda}\mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\alpha}) \prod_{s_m=1}^{|\mathcal{S}|} \cdots \prod_{s_0=1}^{|\mathcal{S}|} \left\{ \sum_{k=1}^{K} T(s_m, \ldots, s_1 \to s_0 \mid k)\lambda_k \right\}^{\mathcal{N}_n(s_m, \ldots, s_1 \to s_0)} \tag{1}$$

Each sequence will have its own expectation under the Dirichlet mixing coefficients and so the ability of such a representation to model intra-sequence heterogeneity emerges naturally. It should be noted here that in the case where no memory is assumed in the generating process, sometimes referred to as a *zero*'th order Markov model, then (1) reduces to the multinomial LDA model as originally detailed in Blei et al. (2003).

It may be interesting to observe that if in (1) the mixing coefficients were constrained to be drawn exclusively from the vertices of the simplex then the summation within (1) becomes a selector for the k'th generator and the expectation with respect to the Dirichlet distribution becomes an expectation over the distribution of probability mass allocated to each vertex of the simplex, i.e. the integral over the simplex reduces to a weighted summation over the number of possible vertices

$$P(\mathbf{s}_n) = \sum_{k=1}^{K} P(k) \prod_{s_m=1}^{|\mathcal{S}|} \cdots \prod_{s_0=1}^{|\mathcal{S}|} T(s_m, \ldots, s_1 \to s_0 \mid k)^{\mathcal{N}_n(s_m, \ldots, s_1 \to s_0)} \tag{2}$$

For the case where $m = 1$ then the mixture of Markov chains proposed in Cadez et al. (2003) is recovered. Indeed, we now see that for the model represented in (2), for each observed sequence, only one Markov process will be responsible for the generation of a whole sequence.

2.1. *Inference and parameter estimation*

The detailed derivation of the inference and parameter estimation algorithm for the case where $m = 0$ i.e. a multinomial distribution over a *bag-of-words* can be found in Blei et al.

(2003). Extending the detailed derivation developed in Blei et al. (2003) to arbitrary order Markov chains now requires multiple indices, despite this the generalization is straightforward. As detailed in Blei et al. (2003) exact inference within the LDA framework is not possible , however the likelihood can be lower-bounded by introducing a sequence specific parameterised variational posterior $q_n(\boldsymbol{\lambda})$ whose parameters will depend on n, and applying Jensen's inequality such that

$$\log P(\mathbf{s}_n \mid \mathbf{T}, \boldsymbol{\alpha}) \geq \mathsf{E}_{q_n(\boldsymbol{\lambda})}\left[\log\left\{P(\mathbf{s}_n \mid \mathbf{T}, \boldsymbol{\lambda})\frac{\mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\alpha})}{q_n(\boldsymbol{\lambda})}\right\}\right] \tag{3}$$

where $\mathsf{E}_{q_n(\boldsymbol{\lambda})}$ denotes expectation with respect to the variational posterior. In this case, a parameterised Dirichlet form of variational posterior will be adopted: $q_n(\boldsymbol{\lambda}) = \mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\gamma}_n)$. Employing the following abbreviated notation $\mathcal{N}_n^{m\ldots 0} \equiv \mathcal{N}_n(s_m, \ldots, s_1 \to s_0)$, $T_{m\ldots 0,k} \equiv T(s_m, \ldots, s_1 \to s_0 \mid k)$, and introducing the additional variational parameter $Q_n^{m\ldots 0,k}$ then the above (3) can be further lower-bounded by noting that

$$\log P(\mathbf{s}_n \mid \mathbf{T}, \boldsymbol{\lambda}) \geq \sum_{s_m=1}^{|\mathcal{S}|} \cdots \sum_{s_0=1}^{|\mathcal{S}|} \sum_{k=1}^{K} \mathcal{N}_n^{m\ldots 0} Q_n^{m\ldots 0,k} \log\left\{\lambda_k \frac{T_{m\ldots 0,k}}{Q_n^{m\ldots 0,k}}\right\} \tag{4}$$

where $Q_n^{m\ldots 0,k} \geq 0$, $\sum_{k=1}^{K} Q_n^{m\ldots 0,k} = 1$. Alternatively, it is enlightening to think of $Q_n^{m\ldots 0,\cdot}$ as a variational distribution on a discrete hidden variable with K possible outcomes that selects which transition matrix is active at each time step of the generative process. As in Blei et al. (2003) by employing (4) in (3), expanding and evaluating $\mathsf{E}_{q_n(\boldsymbol{\lambda})}[\log \lambda_k] = \psi(\gamma_k) - \psi(\sum_{k'} \gamma_{k'})$, where ψ denotes the digamma function, then solving for $Q_n^{m\ldots 0,k}$ and γ_{kn} and finally combining yields the following multiplicative iterative update for the sequence specific variational free parameter $\boldsymbol{\gamma}_n$,

$$\gamma_{kn}^{t+1} = \alpha_k + \exp\left\{\psi\left(\gamma_{kn}^t\right)\right\} \sum_{s_m=1}^{|\mathcal{S}|} \cdots \sum_{s_0=1}^{|\mathcal{S}|} \mathcal{N}_n^{m\ldots 0} \frac{T_{m\ldots 0,k}}{\sum_{k'=1}^{K} T_{m\ldots 0,k'} \exp\left\{\psi\left(\gamma_{k'n}^t\right)\right\}} \tag{5}$$

Solving for the transition probabilities and combining with the fixed point solutions for each $Q_n^{m\ldots 0,k}$ yields the following iteration

$$\tilde{T}_{m\ldots 0,k} = T_{m\ldots 0,k}^t \sum_{n=1}^{N} \mathcal{N}_n^{m\ldots 0} \frac{\exp\left\{\psi\left(\gamma_{kn}^t\right)\right\}}{\sum_{k'=1}^{K} T_{m\ldots 0,k'}^t \exp\left\{\psi\left(\gamma_{k'n}^t\right)\right\}}; \quad T_{m\ldots 0,k}^{t+1} = \frac{\tilde{T}_{m\ldots 0,k}}{\sum_{s_{0'}=1}^{|\mathcal{S}|} \tilde{T}_{m\ldots 0',k}} \tag{6}$$

The parameters of the prior Dirichlet distribution $\boldsymbol{\alpha}$ given the variational parameters $\boldsymbol{\gamma}_n$ are estimated using standard methods (Ronning, 1989; Blei et al., 2003).

Note that both (5) and (6) require an elementwise matrix multiplication and division so these iterations will scale linearly with the number of non-zero state-transition counts.

The next section will detail the relationship of the presented variational Bayes estimation scheme with a simpler Maximum A Posteriori point estimation, highlighting in the same time the close relationship of the LDA modelling framework with Probabilistic Latent Semantic Analysis (PLSA) (or the so called aspect model) (Hofmann, 2001). We will show that these two methods are instances of the same theoretical model and differ only in the estimation procedure adopted. Detailed derivations are given in the Appendix for completeness.

2.2. *Relation with the aspect model*

While approximation methods can only guarantee local maximization of a lower bound on the true likelihood, it is however relatively simple to compute the maximum argument of the true posterior without actually computing the posterior density. This is the so called *maximum a posteriori* (MAP) estimation technique, frequently employed in latent variable models and their network implementations (Attias, 2001). MAP estimators are notoriously prone to overfitting, especially where there is a paucity of available data (Lappalainen and Miskin, 2000). However, MAP estimators are useful e.g. when the task is simply to analyze a given data set, as they provide the most probable hypothesis given the data (Mitchell, 1996). It is also known that if sufficient data is available, then the MAP estimate reaches the Maximum Likelihood estimate (Attias, 2001). We will show that a MAP estimate of LDA under a uniform Dirichlet prior yields exactly PLSA (Hofmann, 2001) (for the zero-th order case), both being instances of the same theoretical model. As an additional insight, we will also highlight the similarity of these two methods at the algorithmic level, both yielding iterations of multiplicative form similar to the 'parts based modelling' technique of Non-negative Matrix Factorisation (Lee and Seung, 2001).

The posterior probability of the random variable $\boldsymbol{\lambda}$ given the observed sequence $\mathbf{s}_n$ and current parameters is $P(\boldsymbol{\lambda} \mid \mathbf{s}_n, \mathbf{T}, \boldsymbol{\alpha})$ so the MAP estimate for $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda}_n^{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{\lambda}} \log\{P(\boldsymbol{\lambda} \mid \mathbf{s}_n, \mathbf{T}, \boldsymbol{\alpha})\} = \operatorname*{argmax}_{\boldsymbol{\lambda}} \log\{P(\mathbf{s}_n \mid \boldsymbol{\lambda}, \mathbf{T})\} + \log\{\mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\alpha})\}$$

Adding a Lagrange multiplier to enforce the constraint that $\boldsymbol{\lambda}_n^{\mathrm{MAP}}$ is a sample point from a Dirichlet random variable, then solving for each $\boldsymbol{\lambda}$ yields the following convergent series of updates for λ_{kn}^t where the superscript denotes the t'th iteration, and as in Lee and Seung (2001), for each observed sequence in the sample a MAP value for the variable $\boldsymbol{\lambda}$ is iteratively estimated by the following multiplicative updates

$$\begin{aligned} \tilde{\lambda}_{kn} &= \lambda_{kn}^t \sum_{s_m=1}^{|\mathcal{S}|} \cdots \sum_{s_0=1}^{|\mathcal{S}|} \mathcal{N}_n^{m\ldots 0} \frac{T_{m\ldots 0,k}}{\sum_{k'=1}^K T_{m\ldots 0,k'} \lambda_{k'n}^t} + (\alpha_k - 1) \\ \lambda_{kn}^{t+1} &= \frac{\tilde{\lambda}_{kn}}{L_n + \sum_k (\alpha_k - 1)} \end{aligned} \tag{8}$$

where $L_n = \sum_{s_m \ldots s_0} \mathcal{N}_n^{m\ldots 0}$ denotes the length of the sequence s_n. Once the MAP values $\boldsymbol{\lambda}_n^{\mathrm{MAP}}$ for each $\mathbf{s}_n$ are obtained then the maximum likelihood estimation of the transition probabilities yields the multiplicative iteration

$$\tilde{T}_{m\ldots 0,k} = T^{t}_{m\ldots 0,k} \sum_{n=1}^{N} \mathcal{N}_n^{m\ldots 0} \frac{\lambda_{kn}^{\mathrm{MAP}}}{\sum_{k'=1}^{K} T^{t}_{m\ldots 0,k'} \lambda_{k'n}^{\mathrm{MAP}}} \; ; \quad T^{t+1}_{m\ldots 0,k} = \frac{\tilde{T}_{m\ldots 0,k}}{\sum_{s_{0'}=1}^{|\mathcal{S}|} \tilde{T}_{m\ldots 0',k}}. \tag{9}$$

Observe that as a special case of the 0-th order model, specifically if employing the maximum entropy Dirichlet prior (i.e. when $\alpha_k = 1, \forall\, k = 1 : K$) we recover exactly the PLSA algorithm. As each $\boldsymbol{\lambda}_n^{\mathrm{MAP}}$ is a Dirichlet sample point, it then defines a multinomial distribution, the k-th dimension of $\boldsymbol{\lambda}_n$ is viewed in PLSA as $P(k \mid n)$. To make the relation to the previously outlined variational approach more evident on the algorithmic level, note that the MAP estimation can be seen as defining the bound (3) using the MAP estimator, such that $q_n(\boldsymbol{\lambda}) = \delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}_n^{\mathrm{MAP}})$, in which case (3) is equal to $\log P(\mathbf{s}_n \mid \mathbf{T}, \boldsymbol{\lambda}_n^{\mathrm{MAP}}) + \log \mathcal{D}(\boldsymbol{\lambda}_n^{\mathrm{MAP}} \mid \boldsymbol{\alpha}) + \mathcal{H}^{\delta}$ where $\mathcal{H}^{\delta}$ denotes the entropy of the delta function around $\boldsymbol{\lambda}_n^{\mathrm{MAP}}$ (which can be discarded in this setting as it does not depend on the model parameters, although it amounts to minus infinity).

2.3. *Prediction with simplicial mixtures*

The predictive probability of observing symbol s_{next} given the n'th sequence of L symbols $\mathbf{s}_n = \{s_{Ln}, \ldots, s_1\}$, generated by an individual, based on a simplicial mixture of m'th order Markov chains is given as

$$P(s_{\mathrm{next}} \mid \mathbf{s}_n) = \int_{\triangle} P(s_{\mathrm{next}} \mid s_m, \ldots, s_1, \boldsymbol{\lambda}) P(\boldsymbol{\lambda} \mid \mathbf{s}_n)\, d\boldsymbol{\lambda} \tag{10}$$

$$= \sum_{k=1}^{K} T(s_m, \ldots, s_1 \to s_{\mathrm{next}} \mid k) \mathsf{E}_{P(\boldsymbol{\lambda}|\mathbf{s}_n)}\{\lambda_k\} \tag{11}$$

hence it is achieved by performing prediction on each 'basis'-transition separately and then combining the results in a user-specific manner as defined by the expectation $\mathsf{E}_{P(\boldsymbol{\lambda}|\mathbf{s}_n)}\{\lambda_k\}$. Note also that from (10) despite m-th order Markov chains forming the basis of the representation, the resulting simplicial mixture is not m-th order Markov with any global transition model. Rather it approximates the individual specific m-th order models whilst keeping the generative parameter set compact. A simplicial mixture of m-th order Markov chains embodies the m-th order information of each individual's past behavior in the user-specific latent variable estimate.

Employing the variational Dirichlet approximation then the following approximation can be employed in the above predictive distribution

$$\mathsf{E}_{P(\boldsymbol{\lambda}|\mathbf{s}_n)}\{\lambda_k\} \approx \mathsf{E}_{\mathcal{D}(\boldsymbol{\lambda}|\gamma_n)}\{\lambda_k\} = \frac{\gamma_{kn}}{\sum_{l=1}^{K} \gamma_{ln}} \tag{12}$$

If we employ the MAP approximation for the Dirichlet distribution then the required expectation can be approximated as

$$\mathsf{E}_{P(\boldsymbol{\lambda}|\mathbf{s}_n)}\{\lambda_k\} \approx \mathsf{E}_{\delta(\boldsymbol{\lambda}-\boldsymbol{\lambda}_n^{\mathrm{MAP}})}\{\lambda_k\} = \lambda_{kn}^{\mathrm{MAP}} \tag{13}$$

where $\lambda_{kn}^{\mathrm{MAP}}$ is the k-th dimension of $\boldsymbol{\lambda}_n^{\mathrm{MAP}}$.

In a mixture model, due to the delta function prior, Eq. (2) , the predictive distribution is

$$\begin{aligned} P(s_{\mathrm{next}} \mid \mathbf{s}_n) &= \sum_{k=1}^{K} P(s_{\mathrm{next}} \mid s_m, \ldots, s_1, k) P(k \mid \mathbf{s}_n) \\ &= \sum_{k=1}^{K} T(s_m, \ldots, s_1 \to s_{\mathrm{next}} \mid k) P(k \mid \mathbf{s}_n) \end{aligned} \tag{14}$$

Note that the posteriors $P(k \mid \mathbf{s}_n)$ are typically sharp, moreover the model insists that only one component is responsible for symbol emission and sequence generation. In consequence, a mixture of m-th order Markov models is not m-th order at the global level, as is noted in Cadez et al. (2003), however at the level of each cluster or prototypical behavior the representation is still m-th order.

In all cases, given a new sequence $\boldsymbol{s}_{new}$, the symbol s_{next} which is most likely to be predicted from the model as a suggested continuation of the sequence, is the maximum argument of $P(s_{\mathrm{next}} \mid \mathbf{s}_n)$. In the next section we consider practical examples where it is reasonable to assume Markovian dynamics at the level of individual behavior, however the heterogeneity of individuals makes even sophisticated global prediction models inefficient.

3. Experiments : Distributed modelling of sequential activity

Three different types of sequential activity are now modelled in the following sections. The first illustrates the utility of the simplicial mixture of Markov chains in modelling the usage and interaction of a number of individuals with a wordprocessor software package. The second example considers modelling the sequential usage of a telephone service by a large group of individuals and finally the web browsing activity of visitors to a commercial website is studied. A brief description of the three collections of sequences is now provided.

3.1. Collections of sequences considered

3.1.1. Word processor command usage. The first collection of sequential activity used in this study consists of the sequences of wordprocessor commands which were issued by a number of individual users during daily working sessions over a period of time. During a session of wordprocessor usage it is possible that a user-specific number of distinct tasks requiring particular sequences of commands may be undertaken, for example the creation and formatting of a table or the insertion and editing of an embedded object. In such a case there may be intra-sequence heterogeneity over interleaved common dynamic patterns which may not be modelled adequately by a mixture model. The sequences were acquired by monitoring the day to day usage of a wordprocessor package by more than 20 individuals

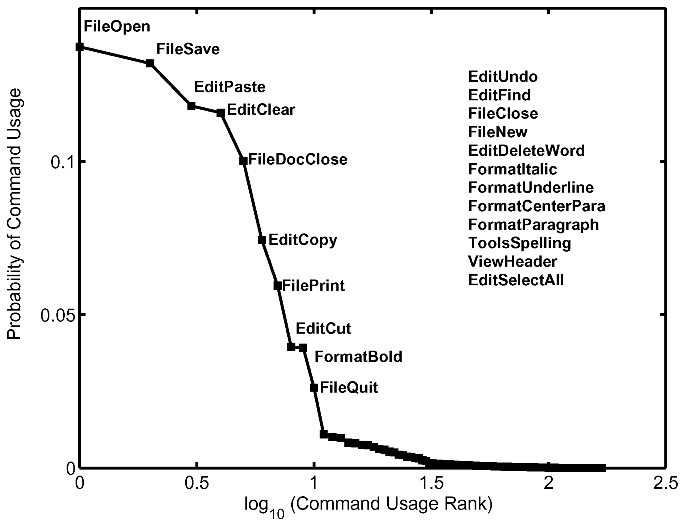

Figure 1. The ranked normalized frequency of usage of the 169 available wordprocessor commands. Each of the ten most frequently used commands is listed at the relevant rank-point on the graph, with the remaining twelve most frequently issued commands being listed in the chart body.

at the MITRE corporation (Linton et al., 2000) over a period of 12 months.[2] Sequences of interactions that were logged during each session having less than three commands issued were discarded and after this trimming there remained 1,460 individual sequences. There are a total of 169 unique commands, however if we observe the ranked distribution of usage of the 169 commands on a log scale we can observe that there is a steep drop in the frequency of usage of certain commands ranked lower than 22 of 169, figure 1. Due to this the twenty-two most commonly used commands are retained, as shown in figure 1, and the remaining 147 are grouped together and listed as `other`. For the purposes of modelling there is a total of 23 symbols within the dictionary which correspond to the set of twenty two most frequently used commands with one symbol representing all those commands which are rarely or never issued. This data set will be referred to as WORD from now on.

3.1.2. Telephone usage modelling. The ability to model the usage of a telephone service[3] is of importance at a number of levels, e.g. to obtain a predictive model of customer specific activity and service usage for the purposes of service provision planning, resource management of switching capacity, identification of fraudulent usage of services. A representative description can be based on the distribution of the destination numbers dialled and connected by the customer, in which case a multinomial distribution over the dialling codes can be employed. One method of encoding the destination numbers dialled by a customer is to capture the geographic location of the destination, or the mobile service provider if not a land based call. This is useful in determining the potential demand placed on telecommunication switches which route traffic from various geographical regions on the service providers network. Two weeks of transactions from a UK telecommunications operator were logged during weekdays, amounting to 36,492,082 and 45,350,654 transactions in each week respectively. All transactions made by commercial customers in the Glasgow region of the UK were considered in this study. This amounts to 1,172,578 transactions from 12,202 high usage customers in the first week considered and 1,753,304 transactions being made in the following week. The mapping from dialling number to geographic region or mobile operator was encoded with 87 symbols amounting to a possible 7,569 symbol transitions. Each customers activity is defined by a sequence of symbols defining the sequence of calls made over each period considered and these are employed to encode activity in a customer specific generative representation.

3.1.3. Web page browsing. The third data set used in this study is a selected subset of the msnbc.com user navigation collection employed in Cadez et al. (2003). Sequences of users who visited at least 9 of the overall 17 page categories (frontpage, news, tech, local, opinion, on-air, misc, weather, msn-news, health, living, business, msn-sports, sports, summary, bbs, travel) have been retained, this selection criteria is motivated by the observation that there would be little scope in trying to model interleaved dynamic behavior in observables which are too short to reveal any intra-sequence heterogeneity. The resulting data set, referred to as WEB, totals 119,667 page requests corresponding to 1,480 web browsing sessions, thus being comparable in size to WORD however having fewer states and state transitions.

3.2. Results

3.2.1. Word processor command usage. In this experiment a range of global and mixture models were assessed for predictive performance. The most basic representation was a zero'th-order Markov chain, in short a MAP estimated [4] multinomial distribution over the twenty three commands. First, second and third order Markov chains (global models) were then assessed for predictive performance by computing the out-of-sample perplexity. In this experiment perplexity is measured, under each model, in the standard manner, computed as the exponential of the negative normalized (normalized by the number of observed symbols) log-likelihood obtained on out-of-sample sequences i.e.

$$\exp\left\{-\frac{1}{\sum_{m=1}^{N_{\text{test}}} L_m}\sum_{m=1}^{N_{\text{test}}}\log P(\mathbf{s}_m)\right\} \tag{15}$$

and due to the small number of available sequences this was estimated using ten-fold cross validation.

From the first row of Table 1 it can be observed how the estimated perplexity varies under differing orders of global Markov models. Moving from a zero'th order to a first-order model accounts for a halving (from 14.09 to 6.70) of the achievable perplexity under the model. It is therefore clear that taking into account the temporal nature of the sequences has a substantial effect on the predictive description of the data. Looking further we observe that a second-order model delivers a very slightly lower perplexity than the first-order model, however it is not statistically significant at the 5% level, as tested using a parametric t-test and a non-parametric Wilcoxon Rank-Sum test. The third-order model exhibits a degree of overfitting which is somewhat expected given that 12,167 state transition probabilities

Table 1. The performance of various models in terms of perplexity. The rows indicate the type of model, either Global, Mixture or Simplicial mixture. The columns indicate the order of the Markov chain, 0—zero'th order, 1—first-order, 2–second-order, 3–third-order. The entries give the mean value $\pm$ one standard error of the perplexity computed over ten-folds and the value in brackets corresponds to the number of components in each of the mixture models. The best result is highlighted in bold.

	0	1	2	3
Global	14.09 ± 0.35	6.70 ± 0.24	6.36 ± 0.23	7.07 ± 0.26
Mixture	9.52 ± 0.48(90)	6.38 ± 0.26 (20)	6.49 ± 0.24 (2)	–
Simplicial	9.13 ± 0.45(90)	$\mathbf{5.91 \pm 0.24}$ (80)	6.04 ± 0.2 (10)	–
HMM	–	6.80 ± 0.20 (50)	–	–

require to be estimated (in comparison to 529 state transition probabilities in the first-order model).

We now consider fitting mixture models to this collection of sequences and assessing their predictive performance. In all mixture models naive random initialization of the parameters was employed and parameter estimation was halted when the in-sample likelihood did not improve by more than 0.001%, no annealing or early stopping was utilized, fifteen randomly initialized parameter estimation runs for each model were performed. In estimating the basis-transitions MAP smoothing with a constant Dirichlet parameter greater than 1 has been utilized, similarly to Cadez et al. (2003) which guarantees that the basis transitions are ergodic.

Initially both zero'th order mixture models (Naive Bayes mixture model) and simplicial mixture models employing the MAP estimator (PLSA) are considered here. The number of factors (dimensionality of the Dirichlet random variable, aspects in PLSA parlance, or classes for the mixture model) in each model ranged from 2 up to 100 elements. The order of the mixture model (number of factors) which gave the lowest out-of-sample perplexity was chosen and the performance along with the corresponding number of factors (in brackets) is listed in Table 1. In the case of the zero'th order model we observe that the mixture model substantially improves over the single global representation and that the MAP estimated simplicial mixture model (PLSA in this case) provides an improvement over the mixture model which is, however, statistically insignificant at the 5% level (employing the t-test). We shall observe in subsequent experiments that employing the variational estimation procedure and relaxing the uniform prior assumption, improved solutions can be obtained as observed in Blei et al. (2003) when modelling text based documents.

If we now consider mixtures of first-order Markov models we note from the second column of Table 1 that the best mixture model achieves a lower perplexity than the global model, the difference however is statistically insignificant at the 5% level. On the other hand the simplicial mixture of first-order Markov chains yields a statistically significant lower value of perplexity than the global and best performing mixture model. A range of hidden Markov models were also assessed on this data and the best performing model achieved similar performance as the global model.

The second-order mixture models performance can be seen to be slightly inferior (though statistically insignificant) to the global model, whilst the simplicial mixture model indicates

Table 2. A listing of five of the most probable transitions from three of the transition matrices of a ten component simplicial mixture of first-order Markov chains.

Component 2	Component 1	Component 8
FileDocClose→FileQuit	FormatItalic→FormatBold	EditClear→EditClear
FilePrint→FileClose	FormatUnderline →FormatItalic	EditCut→EditPaste
FileOpen→FileNew	FormatCenterPara →FormatCenterPara	EditCopy→EditCut
FileNew→FileDocClose	FormatItalic→FormatItalic	EditSelectAll→EditCopy
FileQuit→FileOpen	FormatUnderline →FormatUnderline	EditSelectAll →ToolsSpelling

a robustness to overfitting on this data, which may be improved by employing more efficient estimation methods than the MAP estimator. To further illustrate the manner in which the simplicial representation represents the observed sequences, five of the most probable state transitions are listed for three of the component transition matrices of a ten component first-order model. It is illustrative that the transitions correspond to activities associated with the generic commands `file, format, edit`.

This experiment indicates that the only statistically significant improvement in perplexity over the global first-order Markov model is obtained by a MAP estimated simplicial mixture of Markov chains, thus indicating the potential of such an approach to modelling sequential activity of a group of individuals. This performance can of course be improved by employing more efficient estimation methods such as those developed in Blei et al. (2003) Minka and Lafferty (2002). The following experiment considers a substantially larger collection of logged user activity and assesses whether any practically significant improvement can be achieved when employing simplicial mixtures.

3.2.2. Telephone usage modelling. As with the collection of sequential data of the previous section a reduction in perplexity (measured on the logged activity from the second week) of 53% is achieved when replacing a zero'th order global model with a global first-order model indicating the importance of the temporal content in the sequences. In this experiment the ability of the models to correctly predict the next symbol s_{next} given a sequence $\mathbf{s}_m$ is assessed by employing both the predictive perplexity defined as

$$\exp\left\{-\frac{1}{N_{\text{test}}}\sum_{m=1}^{N_{\text{test}}}\log P(s_{\text{next}} \mid \mathbf{s}_m)\right\} \tag{16}$$

and, in addition, the predictive accuracy under a 0–1 loss, i.e. given a number of previously unobserved truncated sequences, the number of times the model correctly predicts the symbol which follows in the sequence is then counted.

The number of components for the models considered ranged from 2 up to 200. On this data set the parameters of a global first-order Markov chain (bigram), mixtures of

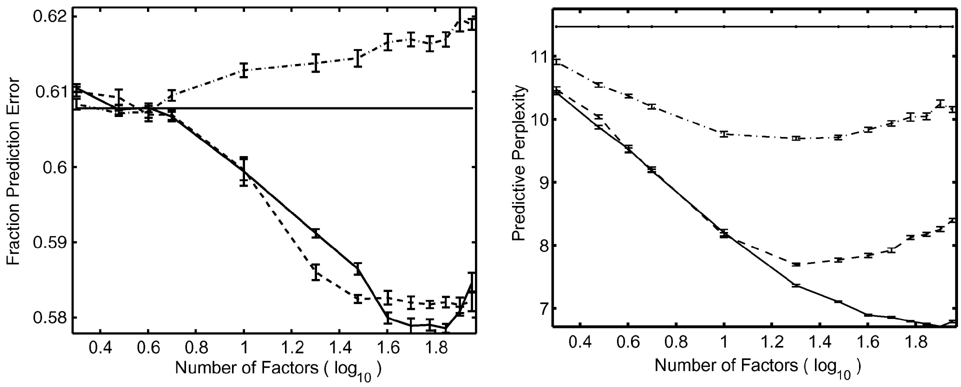

Figure 2. The left hand plot shows the percentage of incorrect predictions against the number of model factors and the right hand plot charts the predictive perplexity of each model against model order for the PHONE dataset. The global first-order Markov chain is represented by a solid straight line, the dashed line represents the MAP estimated simplicial model, the solid line represents the VB estimated simplicial model and the dash-dot line represents the mixture model. The error bars represent one standard error.

first-order Markov chains (Cadez et al., 2003), and simplicial mixtures of first-order Markov chains (using both the MAP and variational (Variational Bayes - VB) estimation procedures) are estimated using the first week of customer transactions and the predictive capabilities of the models are assessed on the transactions from the following week. The results are summarized in figure 2, from the predictive perplexity measures it is clear that the simplicial representation provides a statistically (tested at the 5% level using a t-test) and practically significant reduction in perplexity over the global and mixture models. This is also reflected in the levels of prediction error under each model, however the mixture models tend to perform slightly worse than the global model. As expected the MAP estimated simplicial model performs slightly worse than that obtained using VB (Blei et al., 2003). This also provides an additional insight as to why LDA models improve upon PLSA, as they are in fact both the same model using different approximations to the likelihood, refer to Lappalainen and Miskin (2000) for an illustrative discussion on the weaknesses of MAP estimators. As a comparison to different structural models hidden Markov models with a range of hidden states were also tested on this data set the best results obtained were for a ten state model which achieved a predictive perplexity score of (mean $\pm$ standard-deviation) 11.119 ± 0.624 and fraction prediction error of 0.674 ± 0.959, considerably poorer than that obtained by the models considered here.

In addition to the predictive capability of a simplicial representation of a customers activity the cost of encoding such a representation can be assessed by measuring the entropy rate (Cover and Thomas, 1991) of each of the constituent first-order transition matrices which act as a basis in the representation of the individual specific generative process. The left hand plot of figure 3 shows the distribution of the entropy rates for the transition probabilities in twenty factor simplicial and mixture models, the results are obtained from fifty randomly initialized estimation procedures. The entropy rates for the simplicial mixture are significantly lower than that of a mixture model indicating that the basis of each representation

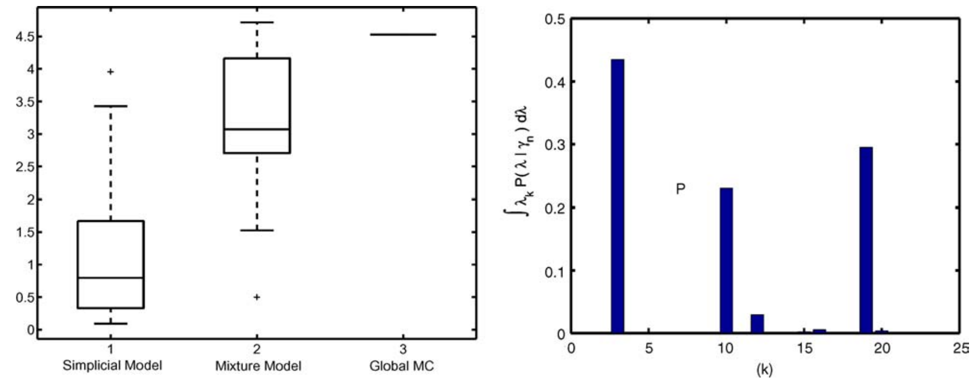

Figure 3. The left hand plot shows the distribution of entropy rates for the transition matrices of a twenty factor mixture and simplicial mixture models (VB). The right hand plot shows the expected value of the Dirichlet variable under the variational approximation for one customer indicating the levels of participation in factor specific behaviors.

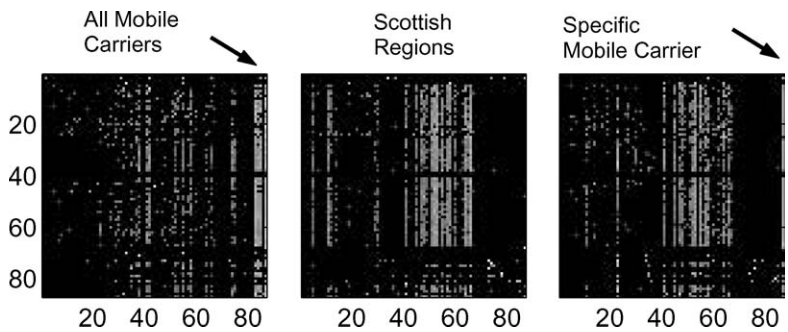

Figure 4. The state transition matrices corresponding to factors 3, 10 & 19 for the customer under consideration.

describes a number of simpler processes. The transition matrices of the three dominant factors defining the behavior of the customer considered in figure 3 are shown in figure 4. These have a clear interpretation in terms of customer activity, the transition matrix corresponding to factor 3, generates sequences which have many transitions from regions 35 & 40 to 75 to 80. Symbols 35 to 40 correspond to locations within Glasgow whilst 75 to 80 correspond to mobile service providers. The second factor shows finite probabilities of transition in region 40 to 60 as well as 10 & 20 corresponding to Scottish regions, whilst the final transition matrix gives high probability of making calls to one specific mobile provider. The profile of each customer can then be defined by the distribution of the expected Dirichlet variable (figure 3) given the 'basis' transition matrices (figure 4).

3.2.3. Web page browsing. The final experiment demonstrated considers the WEB data set. The results of ten-fold cross-validated predictive perplexities again show statistically significant improvement obtained with the VB-estimated simplicial mixture. The results

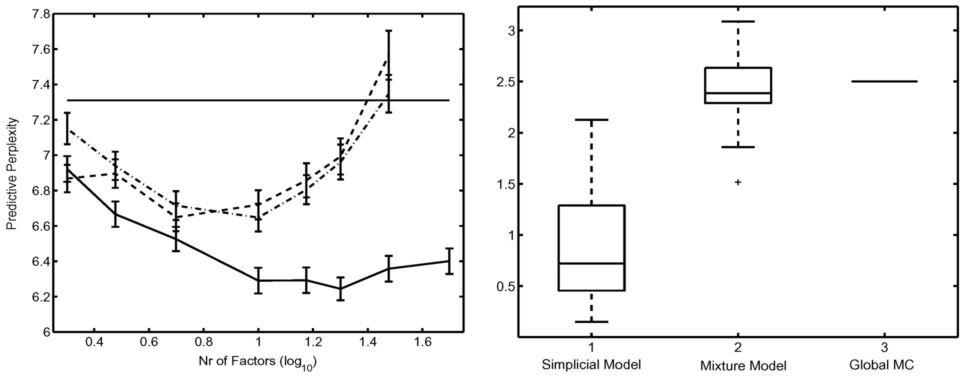

Figure 5. The left hand plot is the predictive perplexity for the WEB data (the straight line corresponds to a global first-order Markov chain. As before, the dashed line represents the MAP estimated simplicial model, the solid line represents the VB estimated simplicial model and the dash-dot line represents the mixture model. The left hand hand plot is the distribution of entropy rates.

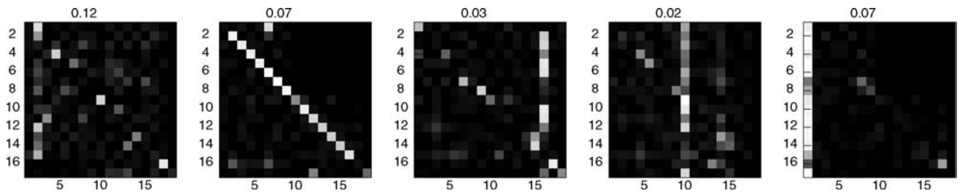

Figure 6. State transition matrices of selected factors from a twenty-factor run produced by Simplicial Mixtures of MCs on the WEB data set.

are summarized in figure 5. Five of the estimated transition factors of a twenty-factor model are shown in figure 6, demonstrating once more that the proposed model creates a low entropy and an easily interpretable dynamic factorial representation. The numbers on the axes on these charts correspond to the 17 page categories enumerated earlier and the average strength of each of these factors amongst the full set of twenty factors computed as $\frac{1}{N}\sum_{n=1}^{N} \mathsf{E}_{\mathcal{D}(\lambda|\gamma_n)}\{\lambda_k\}$ is also given above each chart. We can see that a behavioral feature manifested is a keen interest to visit pages about 'news' along with a quite dynamic transition model (left hand chart) which characterizes around 12% of the behavioral patterns of the entire user population under consideration while static state-repetition (second chart) or an almost exclusive interest in viewing the homepage (last chart) etc represent also relatively strong common characteristics of browsing behavior. The distribution of the entropy rates of the full set of these twenty basis-transitions in comparison to those obtained from the mixture model is given on the right hand plot of figure 5. Clearly, the coding efficiency of a simplicial mixture representation is significantly (statistically tested) superior. Note also these basis-transitions embody correlated transitions (transitions which appear in similar dynamical contexts and so have similar functionality), as can be seen from the multiplicative nature of the equations used for identifying the model. It is not surprising then that state

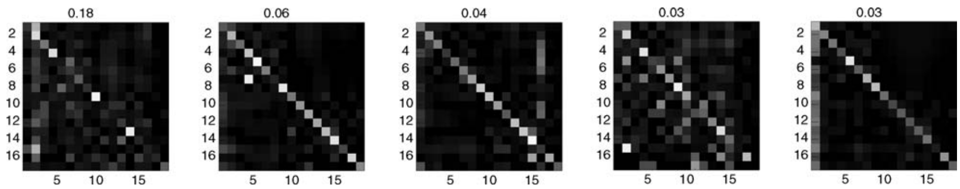

Figure 7. Selected state transition matrices produced by mixture of MCs on a twenty-component run on the WEB data set.

repetitions or transitions which express focused interest in one of the topic categories appear together on distinct factors. We can also see a joint interest in msn-news and msn-sport being present together on the 4-th chart of figure 6—indeed, as the prefix of these page categories also indicates, these are related page categories.

Transitions produced by mixtures of MCs, found to be visually similar to those listed for simplicial mixtures on figure 6, are given on figure 7. The state repetition probabilities are notably high on all parameter transitions. This is because by their construction, mixtures tend to partition the users and represent average behaviors of the identified groups. Clearly, in this case, prototypical users of all groups exhibit a behavioral feature characterized by state repetitions. By contrast, the simplicial mixture does not partition users but extracts behavior features instead. These features may be common to several users or groups of users.

Before concluding, it may also be worth mentioning that our intuition that simplicial mixtures are more appropriate for modelling observation sequences that are sufficiently long and diverse, such that their intra-sequence heterogeneity can be exploited, has also been confirmed in our experiments. While simplicial mixtures perform consistently superior on 'rich' observation sequences, they may become poor when the typical sequence length is very short—in such cases mixtures appear to be more appropriate. We have also found that simplicial mixtures are much more robust against small number of sequences compared to mixtures.

4. Storage and computational scaling

The storage requirements for the sufficient statistics of the N sequences will scale, in the worst case, as $\mathcal{O}(N|\mathcal{S}|^{m+1})$, although for sequential characteristics which do not employ the full dictionary of symbols this storage will scale as the number of non-zero counts. The model parameters will require $\mathcal{O}(NK)$ for the variational parameters, $\mathcal{O}(K|\mathcal{S}|^{m+1})$ for the transition probabilities, with $\mathcal{O}(K)$ storage required for the Dirichlet parameters.

Inspection of the iterations for the variational and transition probabilities shows that they both require elementwise matrix multiplication and so will scale linearly in N, K and with the value of $|\mathcal{S}|^{m+1}$. The estimation routine for the Dirchlet parameters scales linearly in K (see Blei et al., 2003 for details). As in Cadez et al. (2003) plots are provided (figures 8 and 9) which demonstrate the overall linear scaling of the parameter estimation routine run to convergence using sets of sequences from the PHONE dataset.

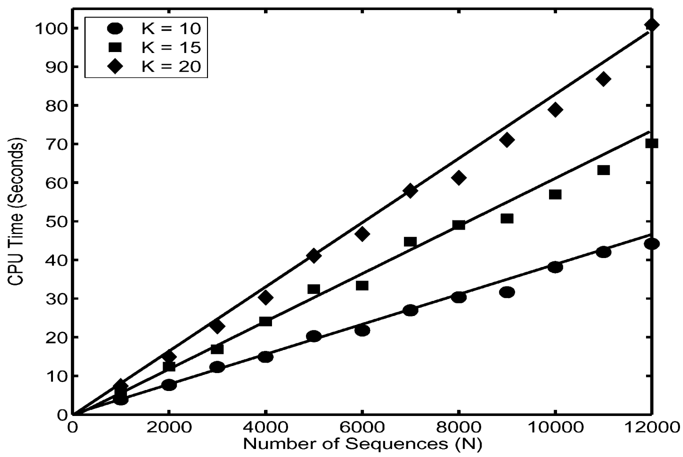

Figure 8. CPU time required (average value from ten randomly initialised runs) for model parameter estimation iterations to converge for a varying number of sequences (N) and transition matrics (K).

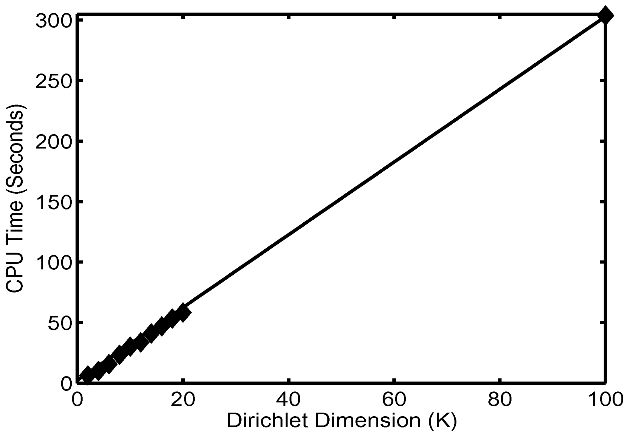

Figure 9. CPU time required for model parameter estimation iterations to converge for a varying Dirichlet parameter dimension (K) using a set of sequences from the PHONE dataset.

5. Conclusions

This paper has presented a linear time method to model finite-state sequences of discrete symbols which may arise from user or customer activity traces. The main feature of the proposed approach has been the assumption that heterogeneous user behavior may be 'explained' by the interleaved action of some structurally simple common generator processes and we have related this representation to several existing models. An empirical study conducted on three collections of logged user activity demonstrated that the proposed approach yields an efficient representation, revealed by both objective measures of prediction performance, low entropy rates, and interpretable representations of the user profiles provided. In spite of its computational simplicity it has been observed that a simplicial mixture of first-order Markov chains is capable of outperforming a global Hidden Markov Model in terms of prediction performance.

Appendix

Here we provide details of the derivation of SMMC update equations. These are obtained from straightforward maximisation of the lower bound likelihood (4) given in the main text and will also highlight the relationship between the VB and MAP estimation procedures considered in this paper.

Obtaining the solution for $Q_n^{m\ldots 0,k}$ and $T_{m\ldots 0,k}$ follows a common route in both the VB and the MAP estimation procedures. As mentioned earlier in the text, the main difference between these two techniques is the form of approximate posterior adopted, therefore the update equation of the variational Dirichlet posterior parameters γ_n is specific to the VB approach only, whereas the update of $\boldsymbol{\lambda}_n^{\mathrm{MAP}}$ is specific to the MAP approach only.

Solving for $Q_n^{m\ldots 0,k}$

The term of the log likelihood bound which contains $Q_n^{m\ldots 0,k}$ is the following:

$$\sum_n \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots 0} \sum_k Q_n^{m\ldots 0,k} \left\{ \log T_{m\ldots 0,k} + E_{q_n(\boldsymbol{\lambda})}[\log \lambda_k] - \log Q_{m\ldots 0,k} \right\} + \beta_{m\ldots 0,n} \left(\sum_k Q_n^{m\ldots 0,k} - 1 \right)$$

$\beta_{m\ldots 0,n}$ are Lagrange multipliers to enforce that $\sum_k Q_n^{m\ldots 0,k} = 1$.

The stationary equation corresponding to $Q_n^{m\ldots 0,k}$ is then following.

$$\mathcal{N}_n^{m\ldots 0} \left\{ \log T_{m\ldots 0,k} + E_{q_n}[\log \lambda_k] - \log Q_n^{m\ldots 0,k} - 1 \right\} - \beta_{m\ldots 0,n} = 0 \tag{18}$$

which provides the update below.

$$Q_n^{m\ldots 0,k} \propto T_{m\ldots 0,k} \exp(E_{q_n}[\log \lambda_k]) \tag{19}$$

with the normalisation factor being $\sum_{k'} T_{m\ldots 0,k'} \exp(E_{q_n}[\log \lambda_{k'}])$.

Note that in Eq. (19), $E_{q_n}[\log \lambda_k] = \psi(\gamma_{kn}) - \psi(\sum_{k'} \gamma_{k'n})$ in the case of VB estimation, whereas in the case of MAP estimation, $E_{q_n}[\log \lambda_k] = \log \lambda_{kn}^{\mathrm{MAP}}$. Hence the difference in the final form of the updates.

Solving for $T_{m\ldots0,k}$

The relevant terms for $T_{m\ldots0,k}$ are the following:

$$\sum_n \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots0} \sum_k Q_n^{m\ldots0,k} \log T_{m\ldots0,k} + \beta_{m\ldots1,k} \left\{ \sum_{s_0=1}^{|S|} T_{m\ldots0,k} - 1 \right\} \tag{20}$$

where $\beta_{m\ldots1,k}$ are Lagrange multipliers which ensure that $\boldsymbol{T}_k$ are stochastic matrices.

Taking derivatives wrt. $T_{m\ldots0,k}$ and equating to zero yields the parameter update below.

$$T_{m\ldots0,k} \propto \sum_n \mathcal{N}_n^{m\ldots0} Q_n^{m\ldots0,k} \tag{21}$$

with the appropriate normalisation factor $\sum_{s_{0'}} \sum_n \mathcal{N}_n^{m\ldots0'} Q_n^{m\ldots0',k}$.

Solving for the variational parameters γ_n

In VB estimation, the posterior Dirichlet is parameterised as $q_n(\boldsymbol{\lambda}) = \mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\gamma}_n)$. The relevant terms (which contain $\boldsymbol{\gamma}_n$) are the following:

$$\begin{aligned} &\sum_k (\alpha_k - 1) \left[\psi(\gamma_{kn}) - \psi\left(\sum_{k'} \gamma_{k'n} \right) \right] - \log \Gamma\left(\sum_k \gamma_{kn} \right) + \sum_k \log \Gamma(\gamma_{kn}) \\ &\quad - \sum_k (\gamma_{kn} - 1) \left[\psi(\gamma_{kn}) - \psi\left(\sum_{k'} \gamma_{k'n} \right) \right] + \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots0} \sum_k Q_n^{m\ldots0,k} \\ &\quad \times \left[\psi(\gamma_{kn}) - \psi\left(\sum_{k'} \gamma_{k'n} \right) \right] \end{aligned} \tag{22}$$

Taking derivatives and equating to zero yields the parameter update

$$\gamma_{kn} = \alpha_k + \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots0} \sum_k Q_n^{m\ldots0,k} \tag{23}$$

Solving for $\boldsymbol{\lambda}^{\mathrm{MAP}}$

In the case of MAP estimation, the the posterior is approximated with a delta function around its maximum, $q_n(\boldsymbol{\lambda}) = \delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}_n^{\mathrm{MAP}})$. The term which contains $\boldsymbol{\lambda}_{kn}^{\mathrm{MAP}}$ is the following

$$\begin{aligned} &\sum_n \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots0} \sum_k Q_n^{m\ldots0,k} \log \lambda_{kn}^{\mathrm{MAP}} + \sum_k (\alpha_k - 1) \log(\lambda_{kn}^{\mathrm{MAP}}) \\ &\quad - \beta_n \left\{ \sum_k \lambda_{kn}^{\mathrm{MAP}} - 1 \right\} \end{aligned} \tag{24}$$

where β_n are Lagrange multipliers that enforce that $\sum_k \lambda_k^{\text{MAP}} = 1$. Taking derivatives and equating to zero yields the update equation.

$$\lambda_{kn}^{\text{MAP}} \propto \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots 0} Q_n^{m\ldots 0,k} + \alpha_k - 1 \tag{25}$$

with the normalisation factor $\sum_{k'} \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots 0} Q_n^{m\ldots 0,k'} + \sum_{k'}(\alpha_{k'} - 1) = \sum_{s_m=1}^{|S|} \cdots \sum_{s_0=1}^{|S|} \mathcal{N}_n^{m\ldots 0} + \sum_{k'}(\alpha_{k'} - 1) = L_n + \sum_{k'}(\alpha_{k'} - 1)$ where L_n is the length of the n-th sequence.

Obtaining multiplicative updates

It is convenient to replace the update expression of $Q_n^{m\ldots 0,k}$ into the updates of the other parameters. This provides multiplicative updates as given in the main text of the paper.

List of symbols: The table below lists the main symbols and nomenclature employed in the paper.

Symbol	Description
$\mathbf{s}_n$	The n'th sequence of observed symbols
L	Length of a sequence
N	Number of observed sequences
K	Dimension of Dirichlet variable which represents the number of *behavioral traits*
$\|\mathcal{S}\|$	Cardinality of symbol dictionary
γ_{kn}	Component k of the Dirichlet posterior parameter for sequence n
$T(s_m, \ldots, s_1 \to s_0 \mid k)$	State transition probabilities
$T_{m\ldots 0,k}$	Shorthand for state transition probabilities
$\mathcal{N}(s_m, \ldots, s_1 \to s_0)$	Number of times symbol s_0 follows the m-tuple $s_m, \ldots, s_1$
$\mathcal{N}_n^{m\ldots 0}$	Shorthand for above definition
$\mathcal{D}(\boldsymbol{\lambda} \mid \boldsymbol{\alpha})$	Dirichlet distribution of $\boldsymbol{\lambda}$ given parameters $\boldsymbol{\alpha}$

Acknowledgments

Mark Girolami is part of the DETECTOR project funded by the Department of Trade and Industry (DTI) Management of Information (LINK) Programme and the Engineering & Physical Sciences Research Council (EPSRC) grant GR/R55184.

Notes

1. The term *user* is employed in this context to mean an individual using a resource, such as, someone who visits and browses a web-site, or someone who regularly uses a telephone service.
2. http://athos.rutgers.edu/ml4um/datasets/owl-data-info.html
3. This data will be made publicly available to allow replication of the reported experiments and enable further investigation.
4. Standard Laplace smoothing was adopted.

References

Anderson, C., Domingos, P., and Weld, D. 2001. Adaptive web navigation for wireless devices. In Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 879–884.

Attias, H. 2001. Learning in high dimension: modular mixture models. In Proc. AI and Statistics.

Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. Latent dirchlet allocation. Journal of Machine Learning Research, 3(5):993–1022.

Borges, J. and Levene, M. 1999. Data mining of user navigation patterns. WEBKDD, pp. 92–111.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. 2003. Model-based clustering and visualisation of navigation patterns on a web site. Journal of data Mining and Knowledge Discovery, 7(4):399–424.

Cover, T.M and Thomas, J.A. 1991. Elements of Information Theory. New York: Wiley.

Deshpande, M. and Karypis. to appear. Selective markov models for predicting web-page accesses. ACM Transactions on Internet Technology.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42:177–196.

Hofmann, T. 2001. Learning What People (Don't) Want. In European Conference on Machine Learning, pp. 214–225.

Frydman, H. 1984. Maximum likelihood estimation in the mover-stayer model. Journal of the American Statistical Society, 79:632–638.

Krogh, A. Hidden Markov Models in computational biology: Applications to protein modelling. Journal of Molecular Biology, 235:1501–1531.

Lee, D. and Sebastian Seung, H. 2001. Algorithms for Non-negative Matrix Factorization. In Advances in Neural Information Processing Systems 13, Todd K. Leen, Thomas G. Dietterich, and Volker. Tresp, (eds.) MIT Press, pp. 556–562.

Linton, F., Joy, D., Schaefer, H.-P. and Charron, A. 2000. OWL: A recommender system for organization-wide learning. Educational Technology & Society, 3.

Mannila, H., and Rusakov, D. 2001. Decomposing event sequences into independent components. In First SIAM Conference on Data Mining.

Manning, C.D. and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.

Minka, T. and Lafferty, J. 2002. Expectation-propogation for the generative aspect model. In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence.

Mitchell, T. 1996. Machine Learning. New York, US: McGraw-Hill.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In Proc. of the IEEE, 77(2):257–285.

Raftery, A. 1985. A model for higher-order Markov chains. Journal of the Royal Statistical Society B, 47(3):528–539.

Ronning, G. 1989. Maximum likelihood estimation of Dirichlet distributions. Journal of Statistical Computation and Simulation, 32(4):215–221.

Ross, D.A. and Zemel, R.S. 2003. Multiple-cause vector quantiztion. In Advances in Neural Information Processing Systems 15, S. Becker, S. Thrun and K. Obermayer (Eds.), MIT Press, pp. 1017–1024.

Sarukkai, R. 2000. Link predicition and path analysis using markov chains. Computer Networks, 33(1–6):377–386.

Saul, L. and Pereira, F. 1997. Aggregate and mixed-order markov models for statistical language processing. In Proceedings of 2nd International Conference on Empirical Methods in Natural Language Processing, pp. 81–89.

Saul, L.K. and Jordan, M.I. 1999. Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones. Machine Learning, 37:75–87.

Tino, P. and Dorffner, G. 2001. Predicting the future of discrete sequences from fractal representations of the past. Machine Learning, 45(2):187–218.

Lappalainen, H. and Miskin, J.W. 2000. Ensemble learning. In Advances in Independent Component Analysis, M. Girolami (Ed.). Springer-Verlag, pp. 75–92.

Mark Girolami is a Reader in the Department of Computing Science at the University of Glasgow. In 2000 he was the TEKES visiting professor at the Laboratory of Computing and Information Science in Helsinki University of Technology. In 1998 and 1999 Dr. Girolami was a research fellow at the Laboratory for Advanced Brain Signal Processing in the Brain Science Institute, RIKEN, Wako-Shi, Japan. He has been a visiting researcher at the Computational Neurobiology Laboratory (CNL) of the Salk Institute. This year (2005) he will take up an MRC funded Discipline Hopping Award in the Department of Biochemistry. Mark Girolami holds a degree in Mechanical Engineering from the University of Glasgow (1985), and a Ph.D. in Computing Science from the University of Paisley (1998). Dr. Girolami was a development engineer with IBM from 1985 until 1995 when he left to pursue an academic career.

Ata Kabán received the B.Sc. degree with honours (1999) in computer science from the University "Babes-Bolyai" of Cluj-Napoca, Romania, and the Ph.D. degree in computer science (2001) from the University of Paisley, UK. She is a lecturer in the School of Computer Science of the University of Birmingham. Her current interests concern probabilistic modelling, machine learning and their applications to automated data analysis. She has been a visiting researcher at Helsinki University of Technology (June–December 2000 and in the summer of 2003). Prior to her career in Computer Science, she received the B.A. degree in musical composition (1994) and the M.A. (1995) and Ph.D. (1999) degrees in musicology from the Music Academy "Gh. Dima" of Cluj-Napoca, Romania.