

Multiple Cause Markov Analysis with Applications to User Activity Profiling

Ata Kaban

University of Birmingham
School of Computer Science

- Joint work with Mark Girolami (University of Glasgow)
- Paper available:

*Simplicial Mixtures of Markov Chains:
Distributed Modelling of Dynamic User Profiles.*
Advances in Neural Information Processing
(**NIPS'03**).

Overview

- Introduction
- Simplicial Mixtures of Markov Chains
- Inference & Estimation
- Prediction
- Applications
- Conclusions

Introduction

- Symbolic sequences over time
 - E.g. Traces of user log activity in electronic environment – cheap to acquire
 - Heterogeneous behaviour
 - Need of efficient profiling
 - To infer user behaviour preferences
 - Possibly to provide personalised environments based on history of activity

- Finding common dynamic causes
 - Existing models are either global or single cause mixtures
 - Global models cannot capture heterogeneity
 - Single cause mixture models assume homogeneous prototypical behaviour within groups so they cannot capture multiple relationships
 - Common behavioural patterns are the basis of multiple relationships between individuals and groups of individuals, which may yield a more realistic model exhibited by the population as a whole.
 - Dynamic activity requires us to devise dynamic models

Related work

- Various monolithic (global) dynamical models
 - E.g. mixture-transition model (Raftery, Saul&Jordan)
- Modelling heterogeneity of ordered sequences by single cause mixtures (i.e. clustering model)
 - Mover-Stayer model (Frydman '84)
 - Visualising clusters of users (Cadez&Smyth '03)
- ICA of sequences
 - Mannila&Rusakov '01
- New developments in multiple cause modelling for memoryless vector-space document modelling
 - Hofmann (PLSA), Blei&Jordan (LDA), Lee&Seung (NMF)

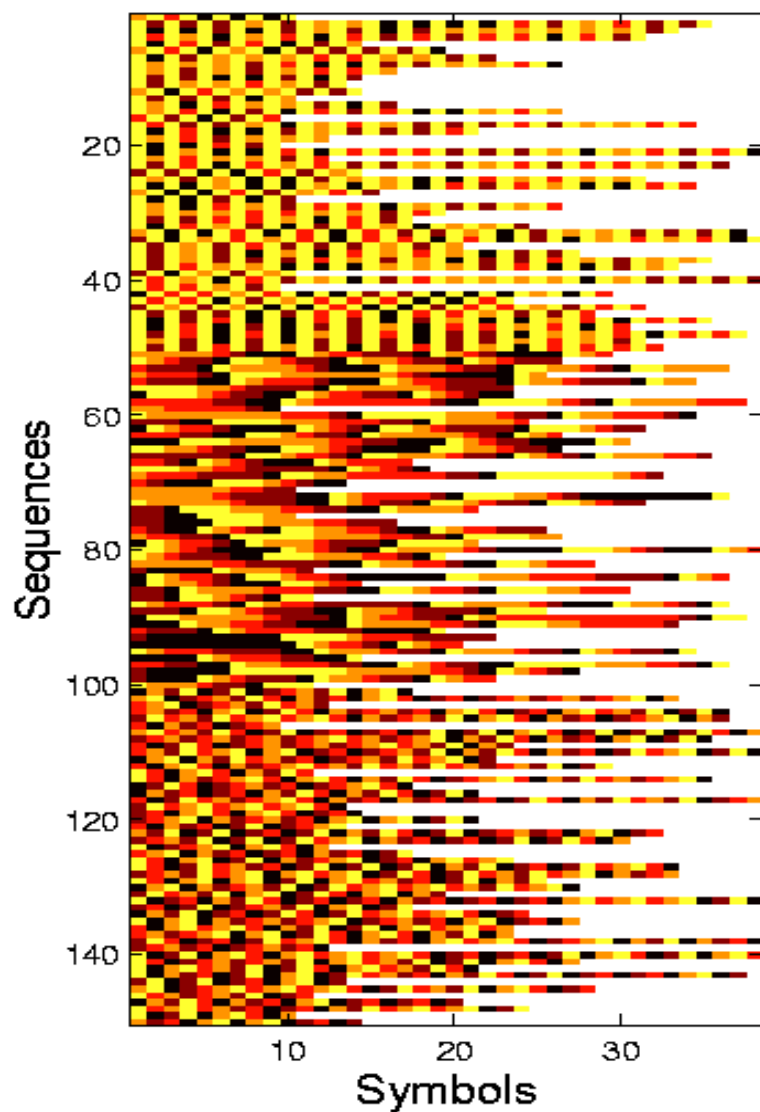
The basic setting

- Observation: Heterogeneous and apparently complex individual behaviour over a common state space
- Desired explanation through a model:
 - Existence of a set of simple basis-generators
 - Interactions between them (mixing)
- Tasks:
 - Estimate the basis-dynamics
 - Infer the interactions
- Further assumptions:
 - On the form of the interactions
 - On the form of dynamics considered

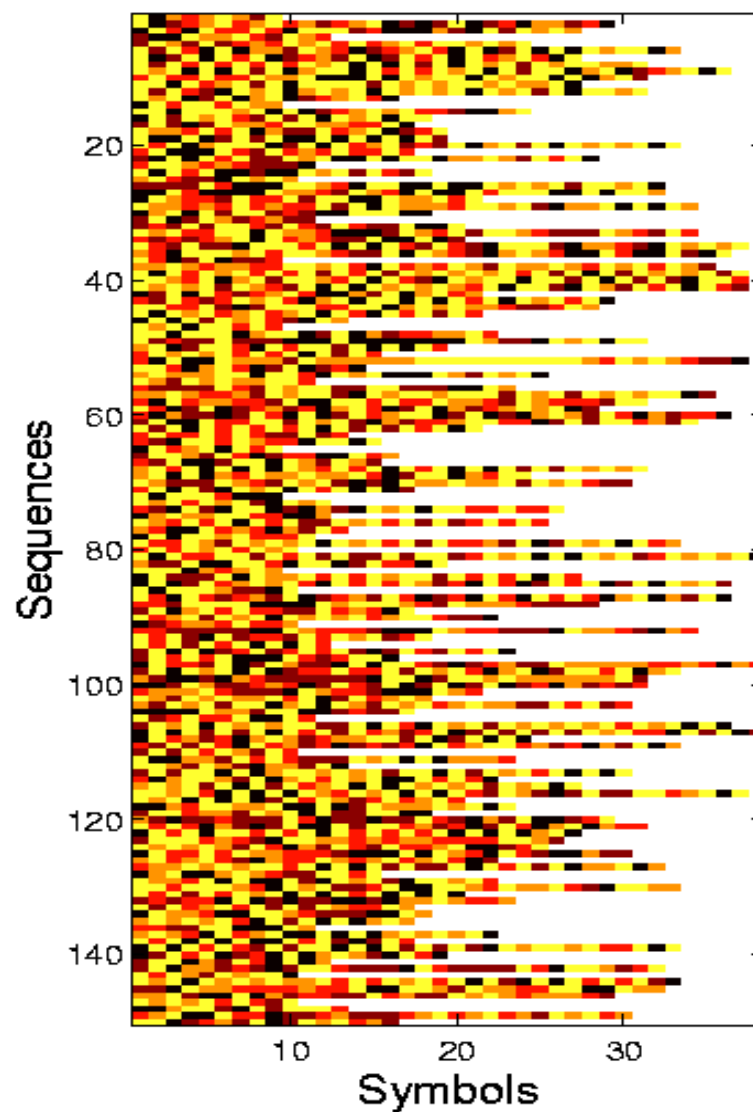
Example

- A set of 1-st order Markov Chains combine to generate sequences by interleaving in various proportions of participation
- Task:
 - Estimate the generator-chains
 - Infer the proportions of participation

MMC Data



SMMC Data



$$p(\mathbf{s}^{(n)} | \mathbf{T}^{(n)}) = \prod_{l=1}^{L_n} p(s_l^{(n)} | s_{l-1}^{(n)}, \mathbf{T}^{(n)})$$

$\mathbf{T}^{(n)} = \mathbf{T}$ global model

$\mathbf{T}^{(n)} = \mathbf{T}^{(n)}$ individual models

$\mathbf{T}^{(n)} \in \{\mathbf{T}_1, \dots, \mathbf{T}_K\}$ single - cause mixture

$\mathbf{T}^{(n)} = \sum_{k=1}^K \lambda_{kn} \mathbf{T}_k$ simplicial mixture

- Single cause prior:

$$p(\boldsymbol{\lambda}) = \sum_k p(\boldsymbol{\lambda}_k^{(0)}) \delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}_k^{(0)})$$

- Multiple cause prior:

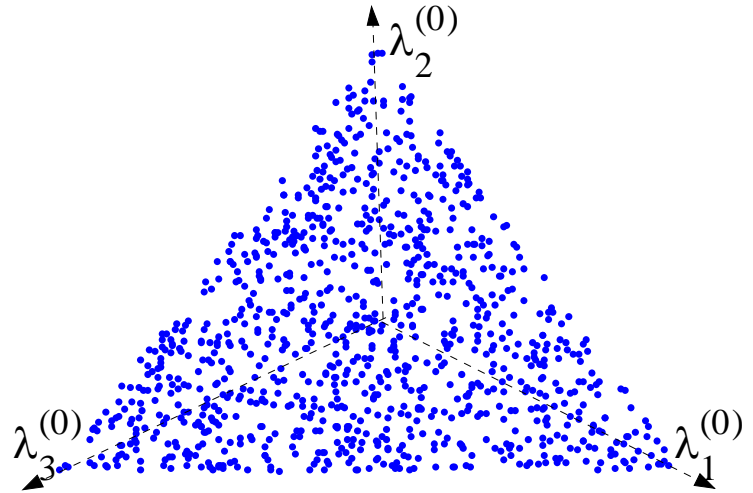
$p(\boldsymbol{\lambda}) = \text{Dirichlet}(\mathbf{1})$ (as implicitly assumed in PLSA)

this is constant for any values of $\boldsymbol{\lambda}$,

so a MAP/ML estimate can be easily obtained

$p(\boldsymbol{\lambda}) = \text{Dirichlet}(\boldsymbol{\alpha})$ (as explicitly formulated in LDA)

the variational estimation procedure developed by Blei et al for LDA can be employed.



Simplicial mixtures of Markov Chains

$$p(\mathbf{s}^{(n)}) = \int d\boldsymbol{\lambda} \text{Dir}(\boldsymbol{\lambda} | \boldsymbol{\alpha}) \prod_{l=1}^{L_n} \left\{ \sum_{k=1}^K T_{s_{l-1}^{(n)} s_l^{(n)} k} \lambda_k \right\}$$

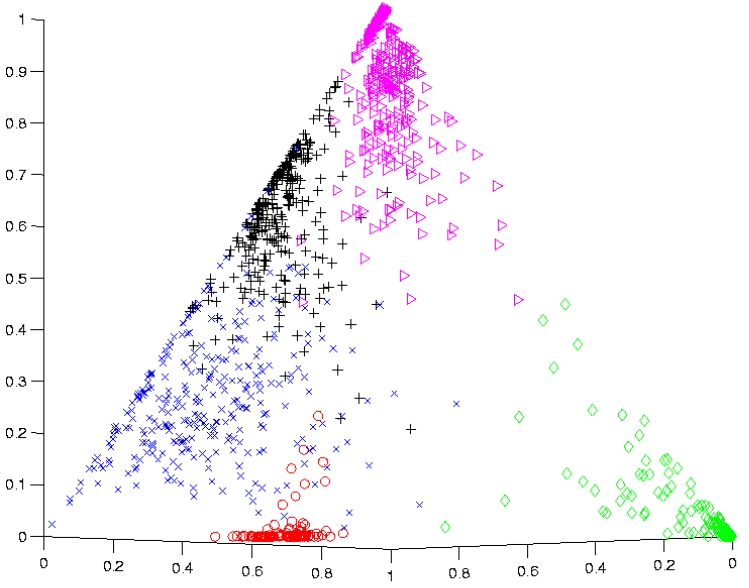
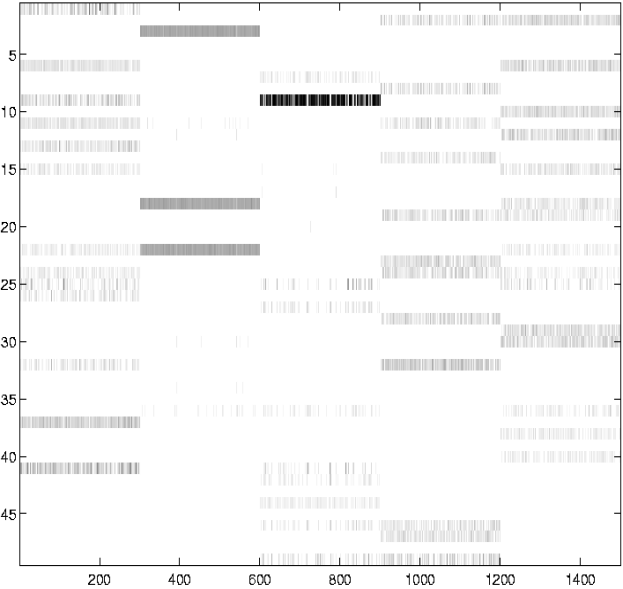
$$P(s_l^{(n)} | s_{l-1}^{(n)}, k)$$

$$P(k | \boldsymbol{\lambda})$$

- Generate $\boldsymbol{\lambda}$ from $\text{Dir}(\boldsymbol{\alpha})$
- For $l=1:L_n$
 - generate k with probability $P(k | \boldsymbol{\lambda})$
 - generate the next symbol s_l^n from the k -th basis transition \mathbf{T}_k i.e. with probability $P(s_l^{(n)} | s_{l-1}^{(n)}, k)$

explanatory variable

Illustration: 5 clusters on a 3D simplex



Inference & Parameter estimation

- Exact inference is not possible, approximation has to be done
- The maths follows Blei et al's and Hofmann's work
- Notice, the above two methods define the same model but differ in the estimation procedure
 - pLSA does *maximum a posteriori* (MAP) estimation for λ and keeps the Dirichlet parameters fixed to 1.
 - LDA does *Variational Bayesian* (VB) estimation for λ and considers a parameterised Dirichlet prior.

MAP estimation

$$\begin{aligned}\boldsymbol{\lambda}_n^{MAP} &= \arg \max_{\boldsymbol{\lambda}} \log P(\boldsymbol{\lambda} | \mathbf{s}_n, \mathbf{T}, \boldsymbol{\alpha}) \\ &= \arg \max_{\boldsymbol{\lambda}} \log P(\boldsymbol{\lambda} | \boldsymbol{\alpha}) + \log P(\mathbf{s}_n | \mathbf{T}, \boldsymbol{\lambda})\end{aligned}$$

$$\log P(\mathbf{s}^{(n)} | \mathbf{T}, \boldsymbol{\lambda}) = r_{ij}^{(n)} \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \log \left\{ \sum_{k=1}^K T_{ijk} \lambda_k \right\} \geq \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \sum_{k=1}^K r_{ij}^{(n)} Q_{ijk}^{(n)} \log \left\{ \lambda_k \frac{T_{ijk}}{Q_{ijk}^{(n)}} \right\}$$

$$\text{where } Q_{ijk}^{(n)} \geq 0, \sum_k Q_{ijk}^{(n)} = 1, \forall i, j, k, n.$$

Solving for \mathbf{T} , $\boldsymbol{\lambda}^{MAP}$, $\boldsymbol{\alpha}$, \mathbf{Q} , then replacing \mathbf{Q} in the updates which contain it yields simple multiplicative updates similar to NMF.

Algorithm

$$\lambda_{kn}^{MAP} := \lambda_{kn}^{MAP} \sum_{i=1}^{|S|} \sum_{j=1}^{|S|} \frac{r_{ij}^{(n)}}{\sum_{k'=1}^K T_{ijk} \lambda_{kn}^{MAP}} T_{ijk}; \lambda_{kn}^{MAP} = \frac{\lambda_{kn}^{MAP}}{\sum_{k'=1}^K \lambda_{k'n}^{MAP}}$$

$$T_{ijk} := T_{ijk} \sum_{n=1}^N \frac{r_{ij}^{(n)}}{\sum_{k'=1}^K T_{ijk} \lambda_{kn}^{MAP}} \lambda_{kn}^{MAP}; T_{ijk} := \frac{T_{ijk}}{\sum_{j'=1}^{|S|} T_{ij'k}}$$

- Linear in the number of observed transitions
- if $\alpha=1$ fixed and memoryless case considered then this is identical to PLSA

VB estimation

$$\log P(\mathbf{s}^{(n)} | \mathbf{T}, \boldsymbol{\alpha}) \geq E_{q_n(\boldsymbol{\lambda})} [\log \{ P(\mathbf{s}^{(n)} | \mathbf{T}, \boldsymbol{\lambda}) \frac{D(\boldsymbol{\lambda} | \boldsymbol{\alpha})}{q_n(\boldsymbol{\lambda})} \}]$$

$$\log P(\mathbf{s}^{(n)} | \mathbf{T}, \boldsymbol{\lambda}) = r_{ij}^{(n)} \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \log \{ \sum_{k=1}^K T_{ijk} \lambda_k \} \geq \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \sum_{k=1}^K r_{ij}^{(n)} Q_{ijk}^{(n)} \log \{ \lambda_k \frac{T_{ijk}}{Q_{ijk}^{(n)}} \}$$

where $Q_{ijk}^{(n)} \geq 0, \sum_k Q_{ijk}^{(n)} = 1, \forall i, j, k, n.$

$q_n(\boldsymbol{\lambda}) := D(\boldsymbol{\lambda} | \boldsymbol{\gamma}_n)$ parameterised approximate posterior

Solving for \mathbf{T} , $\boldsymbol{\chi}$, $\boldsymbol{\alpha}$, \mathbf{Q} , then replacing \mathbf{Q} in the updates which contain it yields simple multiplicative updates similar to NMF.

Algorithm

$$\gamma_{kn} := \alpha_k + \exp(\psi(\gamma_{kn})) \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} \frac{r_{ij}^{(n)}}{\sum_{k'=1}^K T_{ijk} \exp(\psi(\gamma_{kn}))} T_{ijk}$$

$$T_{ijk} := T_{ijk} \sum_{n=1}^N \frac{r_{ij}^{(n)}}{\sum_{k'=1}^K T_{ijk} \exp(\psi(\gamma_{kn}))} \exp(\psi(\gamma_{kn})); T_{ijk} := \frac{T_{ijk}}{\sum_{j'=1}^{|\mathcal{S}|} T_{ij'k}}$$

- Linear in the number of observed transitions
- If memoryless case considered, this is LDA

- Remember, that MAP estimation can also be seen as defining a lower bound approximation on the log likelihood, with expectations now taken wrt.

$$q_n(\boldsymbol{\lambda}) := \delta(\boldsymbol{\lambda} - \boldsymbol{\lambda}_n^{MAP})$$

- Remember also that although MAP estimators are usually simpler to derive, they are notoriously prone to overfitting

Prediction

$$\begin{aligned} P(s_{next} | \mathbf{s}_n) &= \int d\boldsymbol{\lambda} P(s_{next} | s_{L_n}, \boldsymbol{\lambda}) P(\boldsymbol{\lambda} | \mathbf{s}_n) \\ &= \int d\boldsymbol{\lambda} \sum_{k=1}^K T_{next, L_n, k} \lambda_k P(\boldsymbol{\lambda} | \mathbf{s}_n) \\ &= \sum_{k=1}^K T_{next, L_n, k} E_{P(\boldsymbol{\lambda} | \mathbf{s}_n)}[\lambda_k] \end{aligned}$$

- Combines basis-wise predictions in proportions specified by the posterior expectation
- User-specific deeper past is embodied in the posterior

SM of 1-st order MCs is not 1-st order

$$E_{P(\boldsymbol{\lambda}|\boldsymbol{\gamma}_n)}[\boldsymbol{\lambda}_k] \left\{ \begin{array}{l} \approx E_{D(\boldsymbol{\lambda}|\boldsymbol{\gamma}_n)}[\boldsymbol{\lambda}_k] = \frac{\gamma_{kn}}{\sum_{k'=1}^K \gamma_{k'n}} \text{ if VB estimation employed} \\ \approx E_{\delta(\boldsymbol{\lambda}-\boldsymbol{\lambda}_n^{MAP})}[\boldsymbol{\lambda}_k] = \boldsymbol{\lambda}_{kn}^{MAP} \text{ if MAP estimation employed} \\ = P(\boldsymbol{\lambda} = \boldsymbol{\lambda}_k^{(0)} | \mathbf{s}_n) \text{ for single - cause mixtures} \end{array} \right.$$

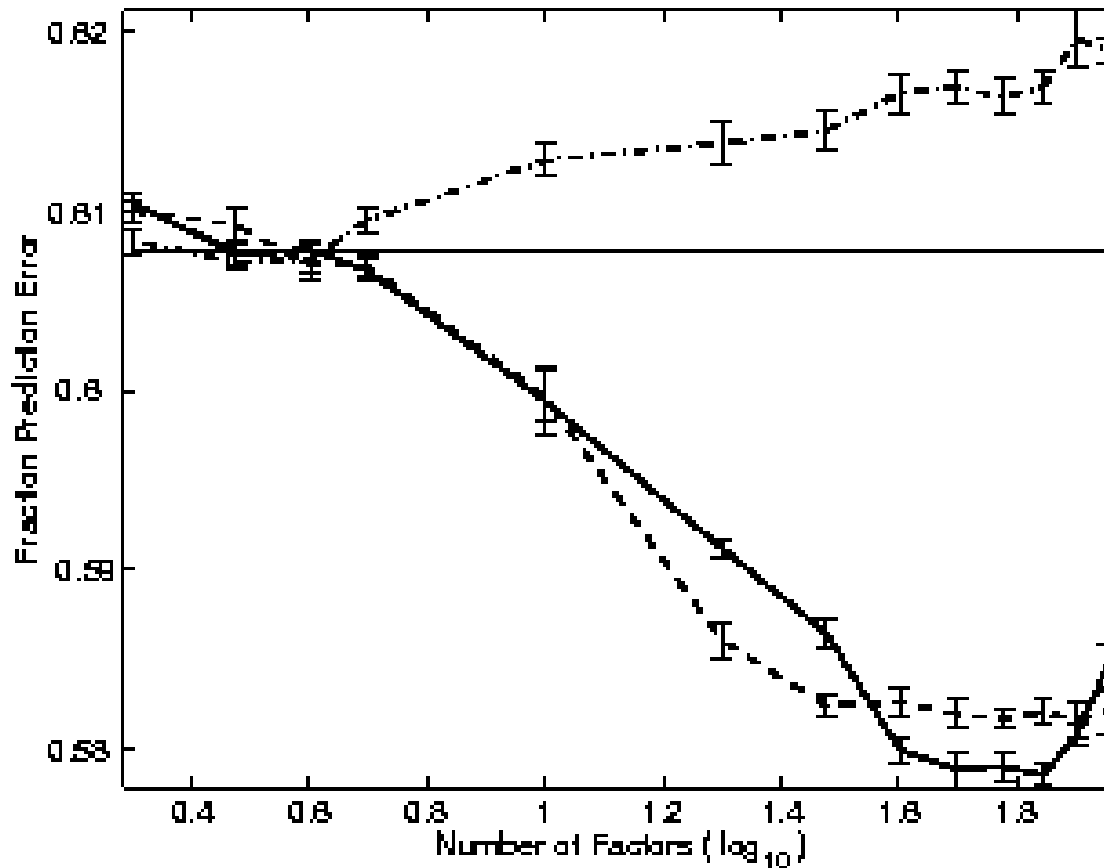
Applications & experiments

- Sequential traces logged from user activity in three different domains considered
- 1) modelling the usage and interaction of a number of individuals with a word-processor software package
- 2) modelling the sequential usage of a telephone service by a large group of individuals
- 3) modelling the web browsing activity of visitors to a commercial website

Telephone Usage Modelling

- 1,172,578 calls in week1
- 1,753,304 calls in week 2
- Destination numbers mapped to 87 geographic regions & mobile operators
- Week1 activity employed for estimation
- Week2 activity used for testing
- Performance measures considered:
 - Predictive perplexity on unseen sequences
 - Percentage of symbols correctly predicted on unseen s
 - Out of sample log likelihood

Prediction error on transactions in Week2



Solid straight line:
global 1-st order
MC

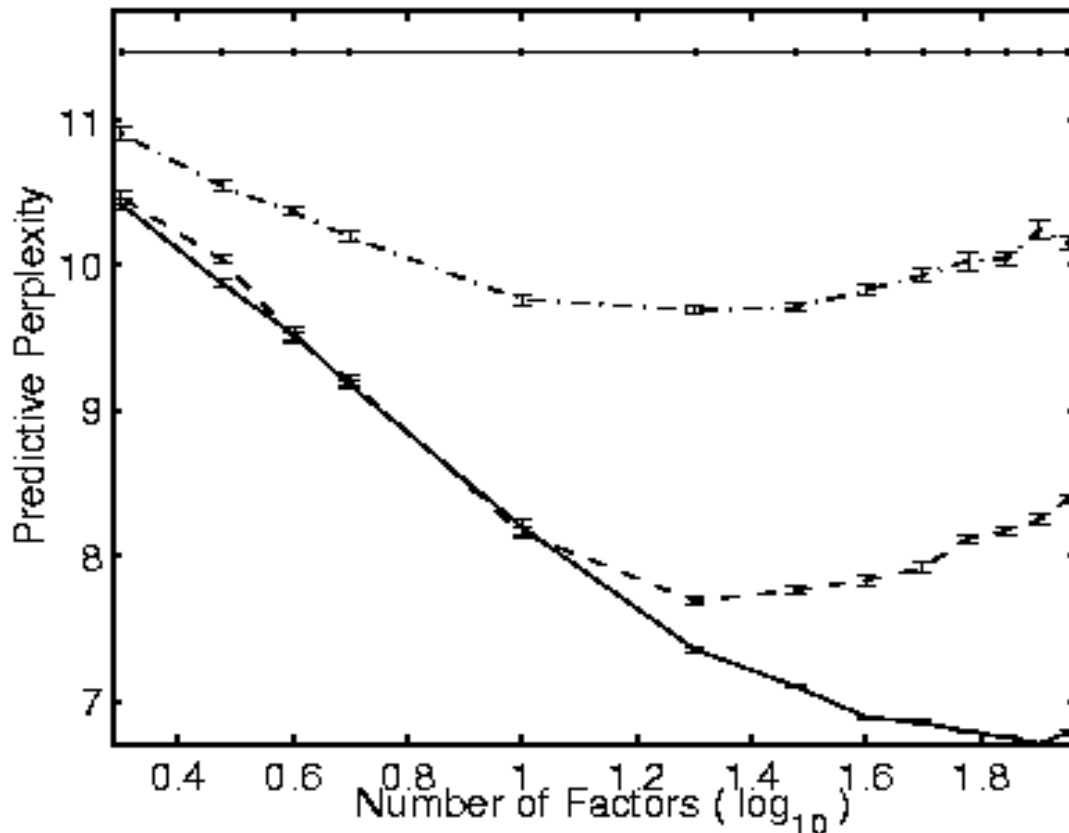
Solid line: SMMC
(estimated with VB)

Dashed line:
SMMC (estimated
with MAP)

Dash-dot line: MMC

Predictive Perplexity on Week2 data

$$\exp\left\{-\frac{1}{N_{test}} \sum_{m=1}^{N_{test}} \log P(s_{next} | \mathbf{s}^{(n)})\right\}$$



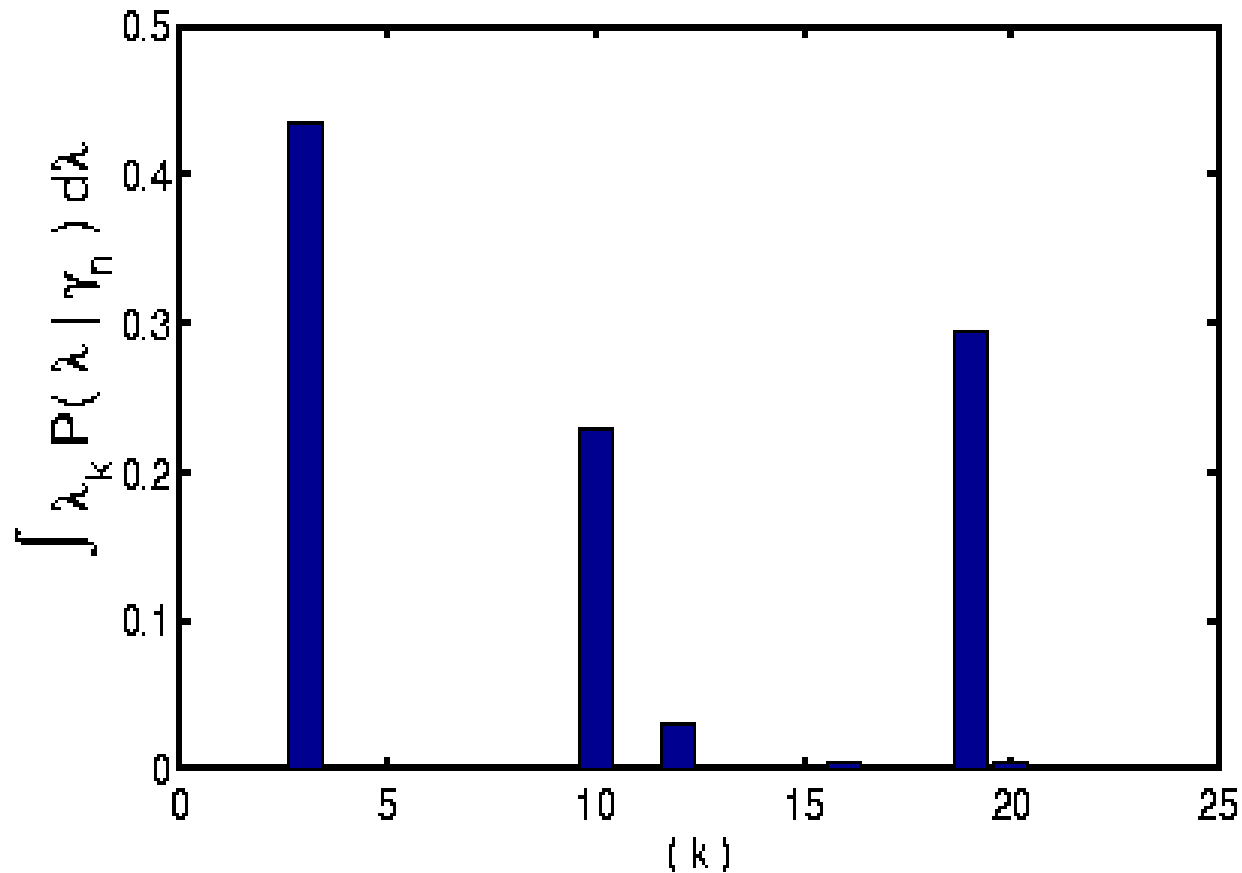
Solid straight line:
global 1-st order
MC

Solid line: SMMC
(estimated with VB)

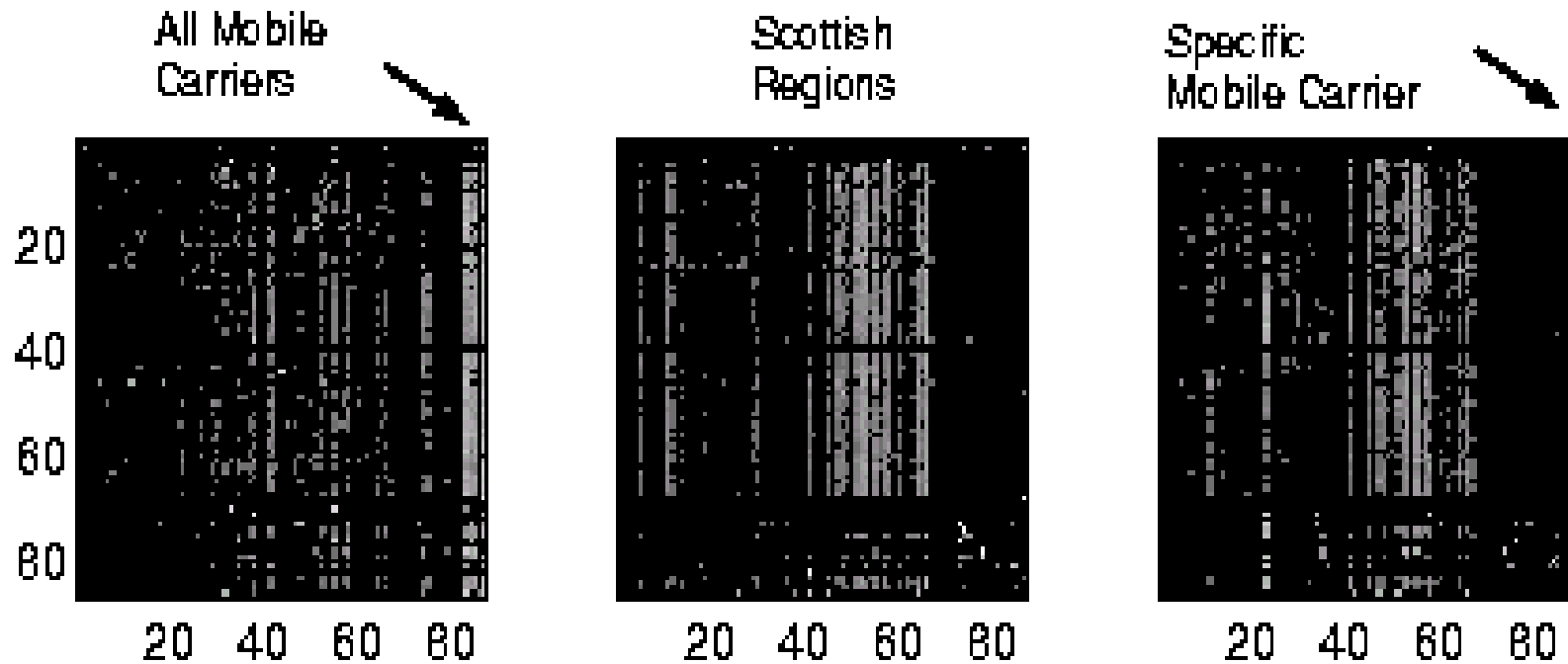
Dashed line:
SMMC (estimated
with MAP)

Dash-dot line: MMC

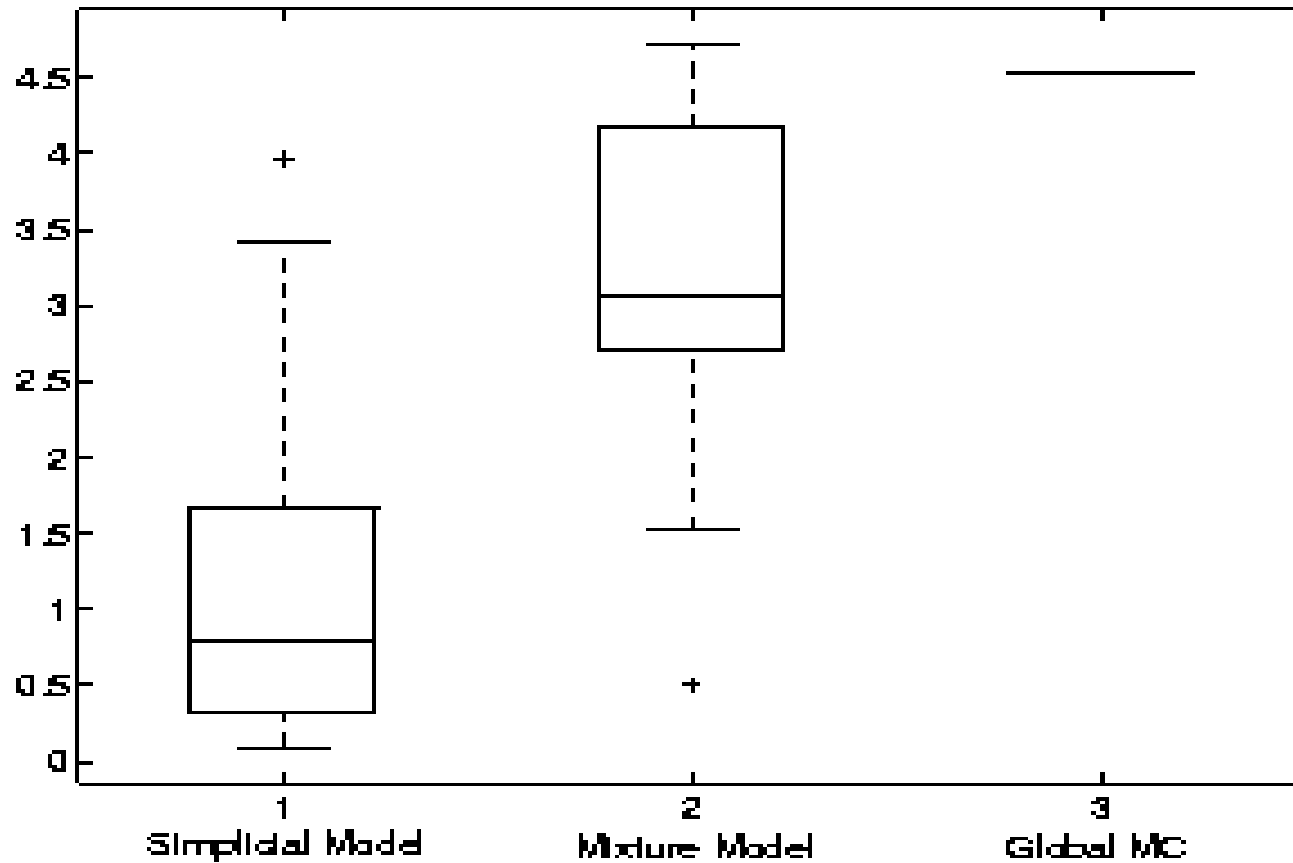
Profile of one of the customers over
a 20-basis-transitions SMMC
(one point on a 19D latent simplex)



The three dominant transition matrices in the previous profile



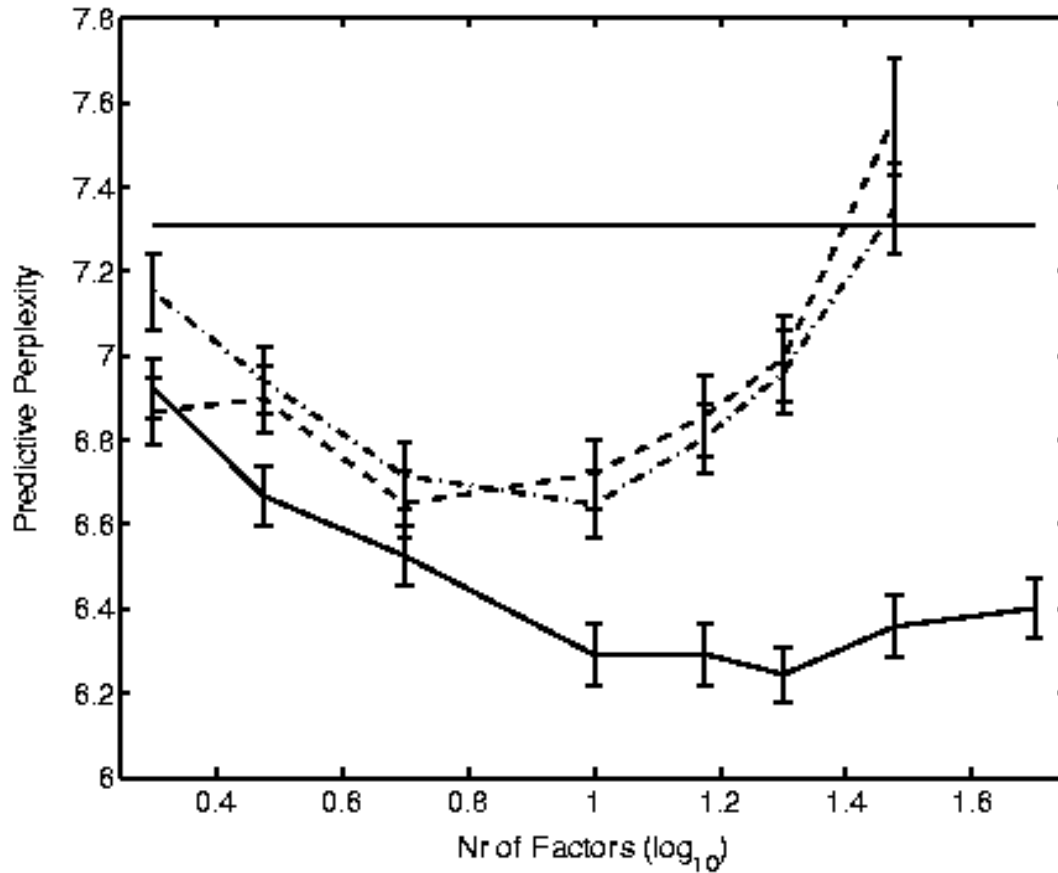
Compression efficiency: Entropy rates of the bases



Web page browsing behaviour modelling

- Dataset used in Cadez&Smyth paper
- 17 page categories from MSN website form the common state space
- Users who visited at least 9 out of 17 page categories selected
- Total 119,667 page requests over 1,480 web browsing sessions (small data set)

10-fold cross-validated predictive perplexity



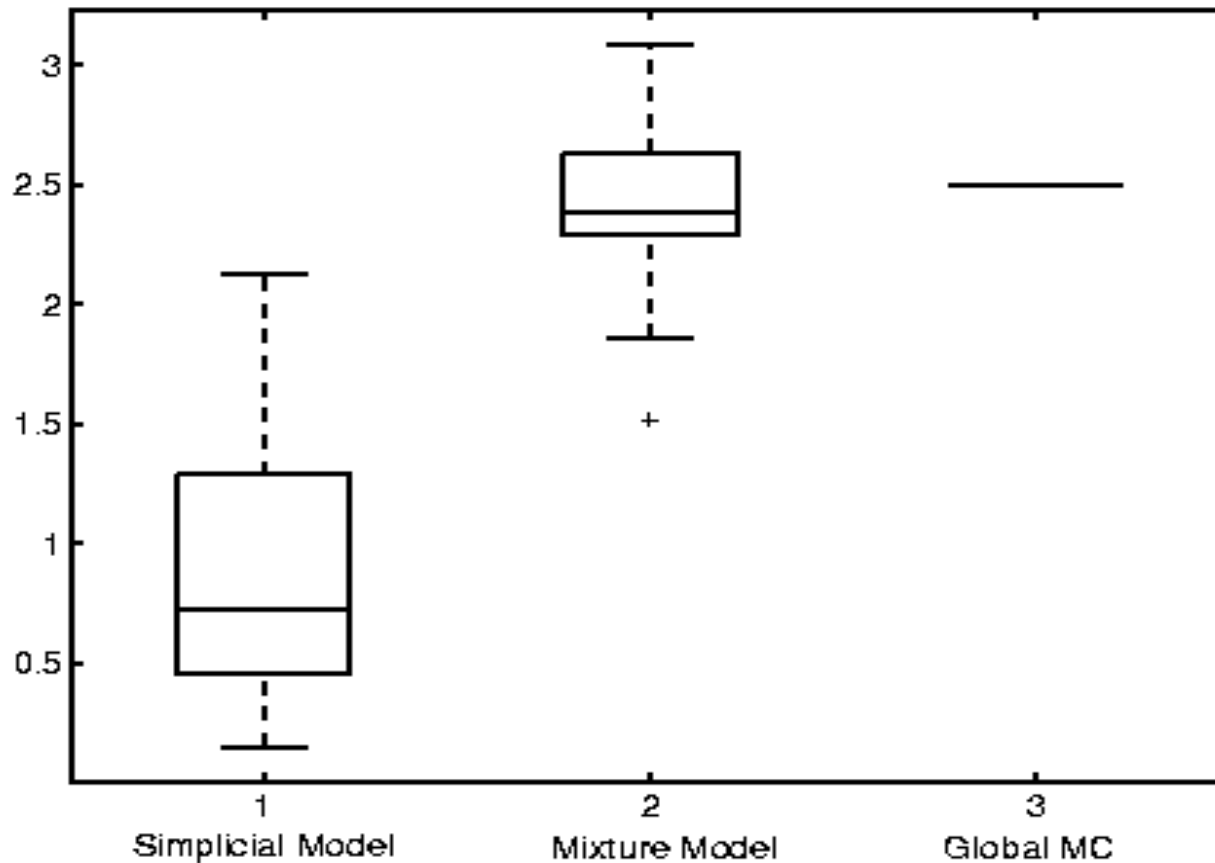
Solid straight line:
global 1-st order
MC

Solid line: SMMC
(estimated with VB)

Dashed line:
SMMC (estimated
with MAP)

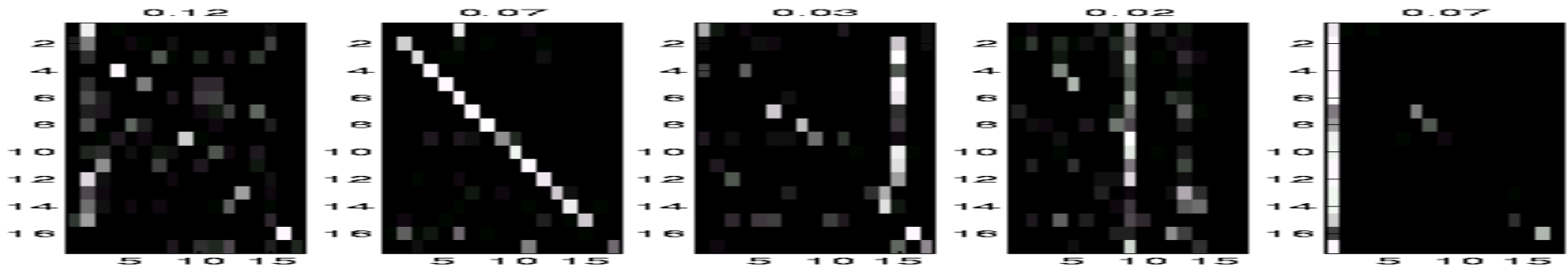
Dash-dot line: MMC

Compression efficiency: entropy rates of the bases

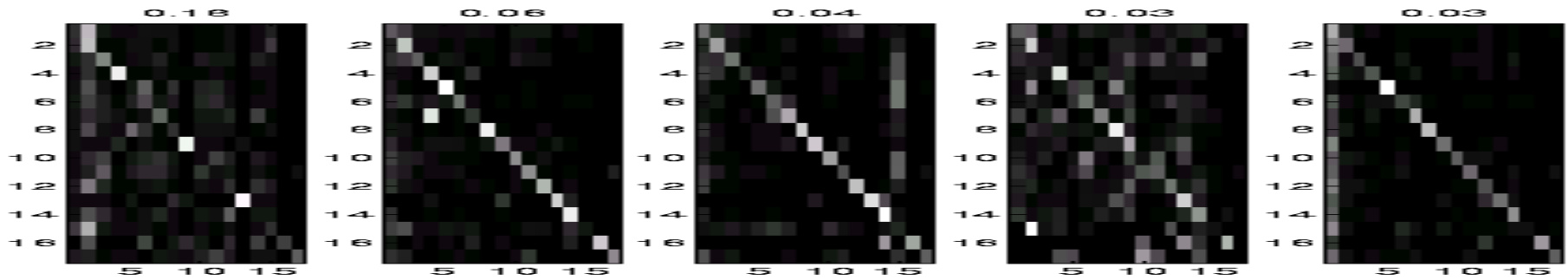


5 selected basis-transitions

SMMC features (common multiple causes)



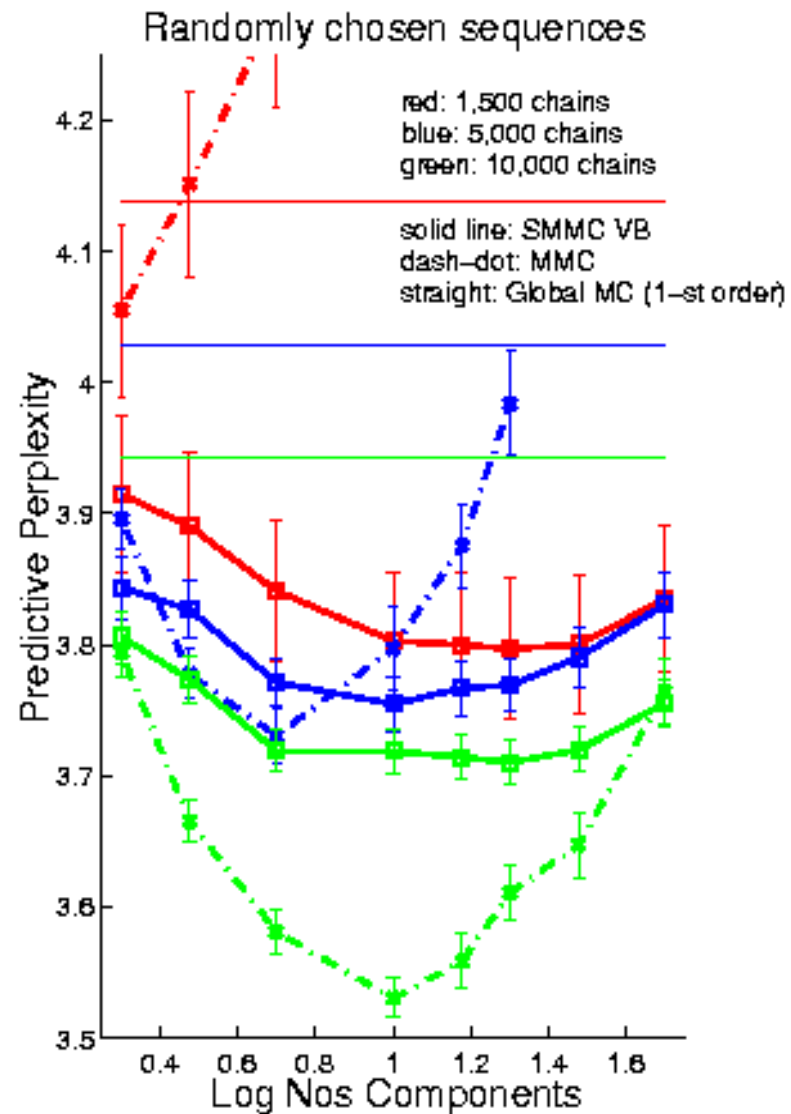
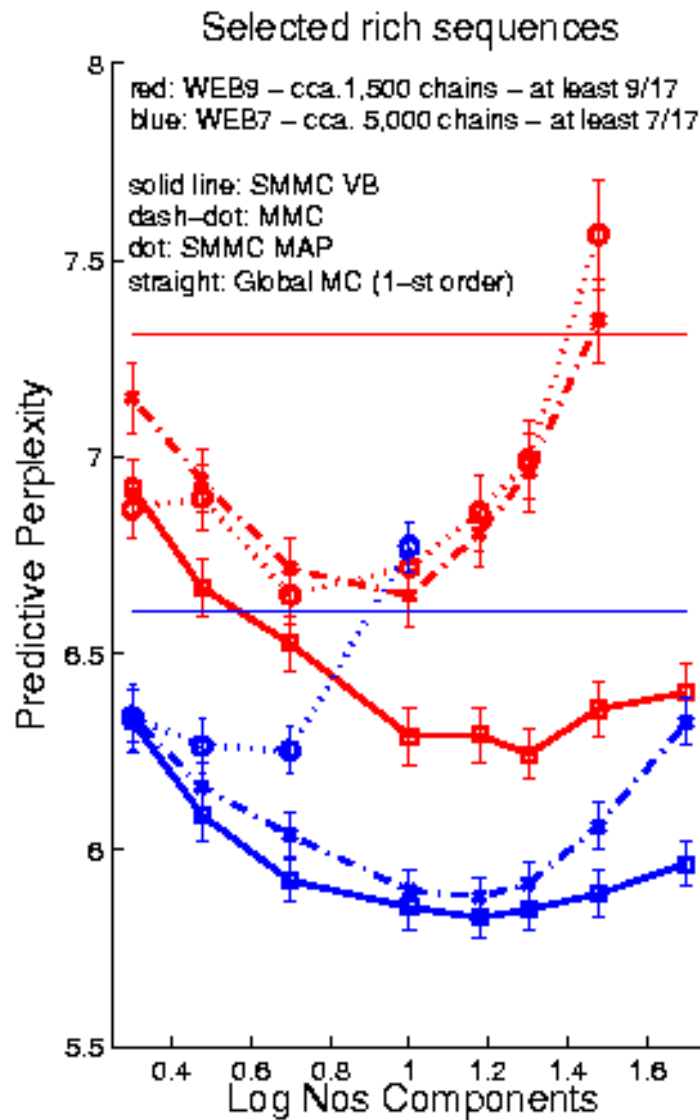
MMC cluster-prototypes



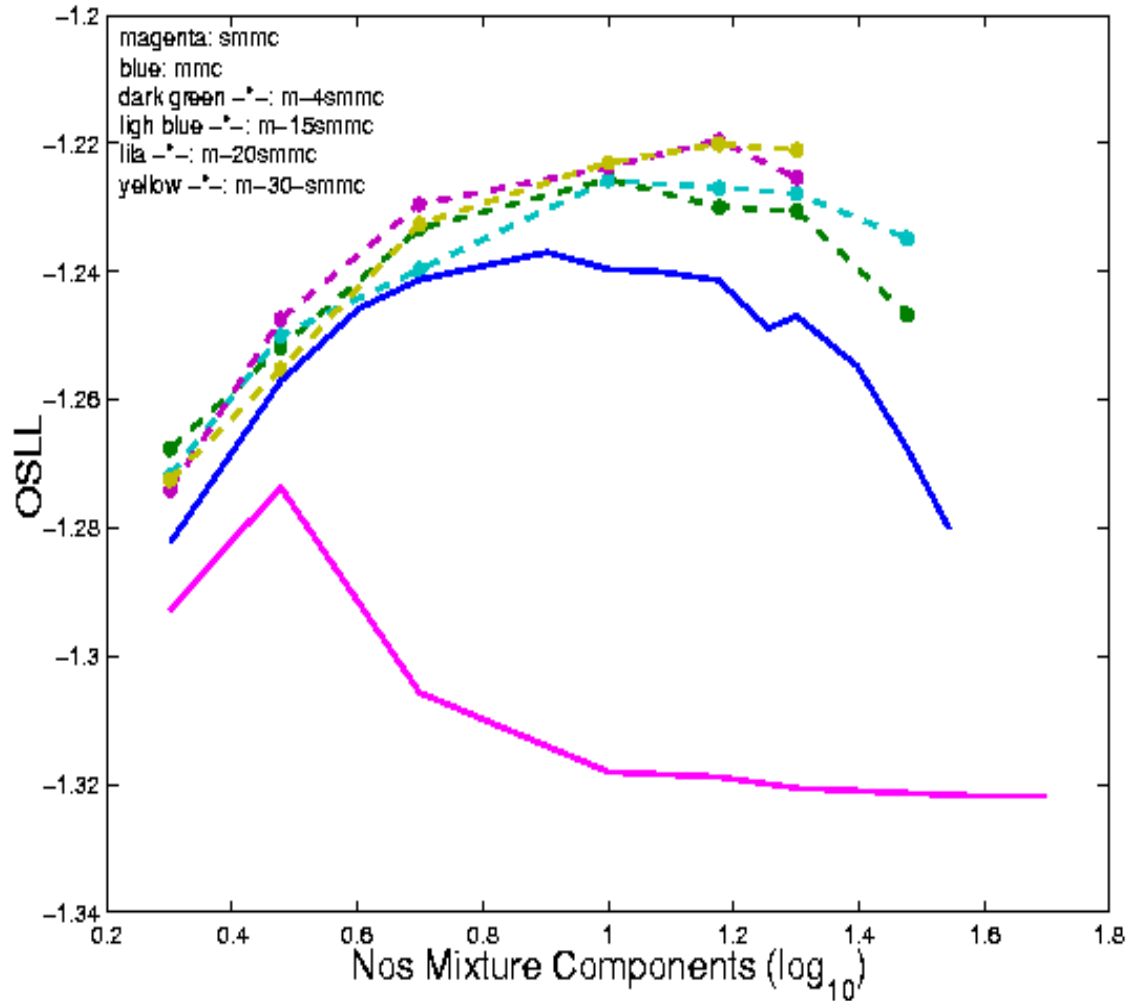
Conclusions

- Complex heterogeneous behaviour is assumed as a convex combination of simple behaviour patterns
- Principled (MAP and VB) inference & estimation employed, following Hofmann's pLSA and Blei et al.'s LDA
- Simple linear time algorithms obtained
- Tested on three applications
 - Efficient compression
 - Interpretable parameters
 - Improved prediction on new, unseen sequences
- Limitations?
 - Need 'rich enough' sequences
 - Assume 'sparse enough' generators
 - Other types of interaction between dynamic generators?

What has not been said



A solution?



Random subset
of size 18,000:
half used for
training, the other
half as an
independent test
set

The End
for Now