

A Scalable Generative Topographic mapping for sparse Data Sequences

Ata Kaban

<http://www.cs.bham.ac.uk/~axk>

a.kaban@cs.bham.ac.uk

School of Computer Science

The University of Birmingham

ITCC'05, April 2005

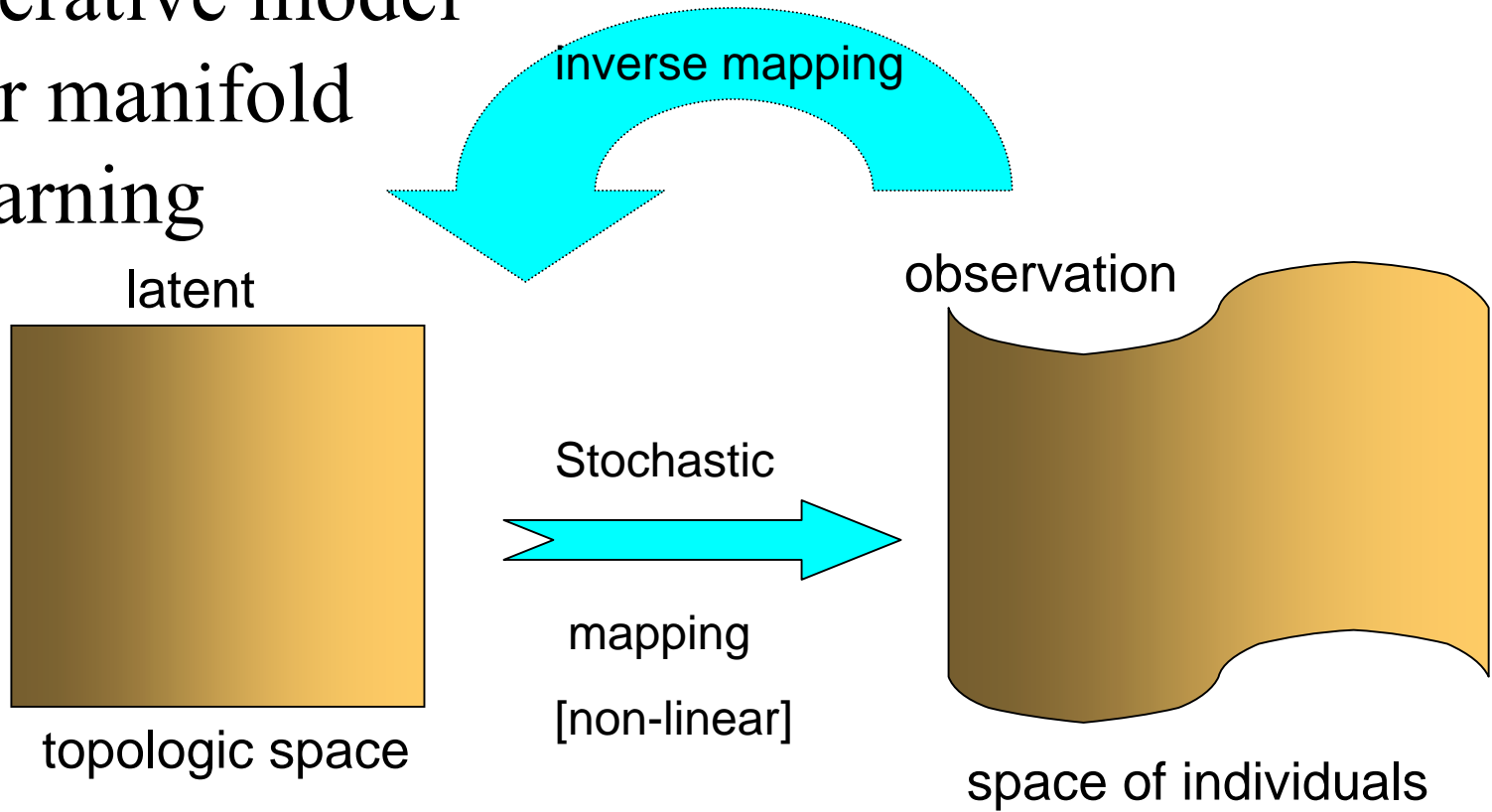
Overview

- Introduction
- A scalable topographic model for symbolic sequences
- Prediction & data exploration with the model
- Applications & experiments
- Conclusions

Introduction

- Example1:
‘... a more appropriate model should consider some *conceptual* dimensions instead of words.’ (Gardenfors)
- Example2:
Collection of 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
 - To provide personalised environments based on history of activity
 - To provide informative overview of the collection of activity

Generative model for manifold learning

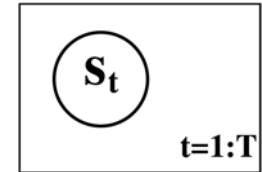


Aim: Infer the inverse stochastic mapping from a latent topological space into the more complex observation-space

Some simple sequence models

- The Random sequence model

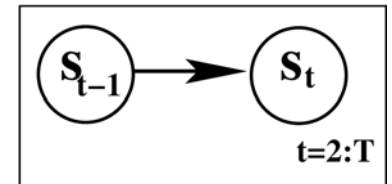
- Symbol at time t does not depend on any previous ones
- A model of a die rolling process
- Nos of parameters: $|W|$
- E.g. ‘bag of words’



$$P_{rand}(Seq) = \prod_{t=1}^T P(s_t)$$

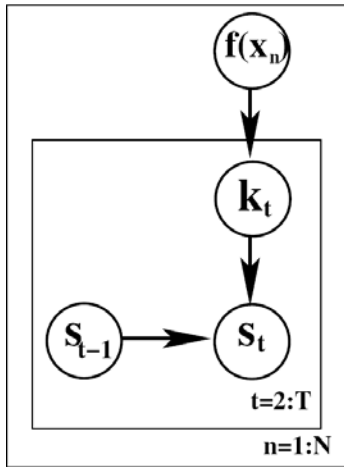
- 1st order Markov chain

- Symbol at time t depends only on the previous ones
- Simple yet effective temporal model
- Nos of parameters: $|W|^2 + |W|$
(initial state prob $P(s_1)$ combined into state transitions, so we can say $|W|^2$)



$$P_{MC(1)}(Seq) = P(s_1) \prod_{t=2}^T P(s_t | s_{t-1})$$

A Scalable Generative Topographic Mapping



$$P(Seq^{(n)}) = \int d\mathbf{x} \text{Uniform}_{[-1,1]^2}(\mathbf{x}) \prod_{t=1}^{T_n} \sum_{k=1}^K P(s_t | s_{t-1}, k) \varphi_k(\mathbf{x})$$

where $\varphi_k(\mathbf{x}) = P(k | \mathbf{x}) \equiv \frac{\exp(-\frac{1}{2\sigma^2} \|\boldsymbol{\mu}_k - \mathbf{x}\|)}{\sum_{k'} \exp(-\frac{1}{2\sigma^2} \|\boldsymbol{\mu}_{k'} - \mathbf{x}\|)}$ is a **smooth function**

The generative process of Seq_n :

- Generate \mathbf{x} from $\text{Uniform}([-1,1]^2)$
- For $t=1:T_n$
 - generate k with probability $P(k | \mathbf{x})$
 - generate the next symbol s_t^n from the k -th 'basis transition model' \mathbf{T}_k i.e. with probability $P(s_t^{(n)} | s_{t-1}^{(n)}, k)$

$E_{P(\mathbf{x}|Seq_n)}[\mathbf{x}]$ can be used to visualise each seq as a point on Euclidean 2D space

Model estimation

- Exact estimation is intractable
- A sampling-based solution developed

$$U(\mathbf{x}) \approx \frac{1}{M} \sum_{m=1}^M \delta(\mathbf{x} - \mathbf{x}_m)$$

- The resulting algorithm scales linearly in the number of non-zero transition counts in the data – this is in sharp contrast to previous topographic models for discrete data
→ this is in sharp contrast with the computational scaling of existing generative topographic models in general and those appropriate for discrete data in particular.

The estimation algorithm

Iterate until convergence :

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

$$\mathbf{A}^{(new)} \propto \mathbf{A}^{(old)} \otimes \{ [\mathbf{D} \mathbf{R}^T] \div [\mathbf{A}^{(old)} \mathbf{\Phi}] \} \mathbf{\Phi}^T$$

where $\mathbf{R} = \{r_{mn}\}$ posteriors of the samples \mathbf{x}_m

\mathbf{A} : parameter matrix with columns having parameters
of the same structure as the observation models

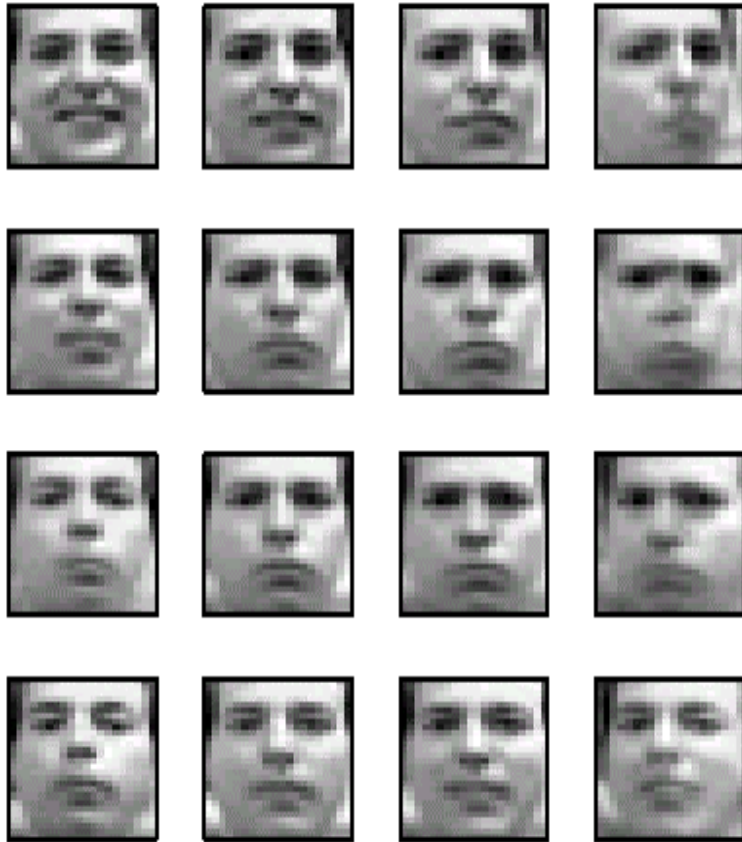
$\mathbf{\Phi} = \{\varphi_k(\mathbf{x}_m)\}$ images of latent space samples

\mathbf{D} : observatin data, having frequency counts

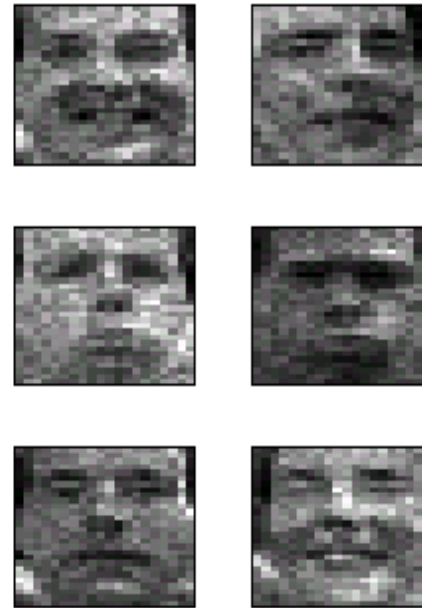
\otimes : element - wise multiplication

\div : element - wise division

Representation properties: Prototypes vs. aspects in the model



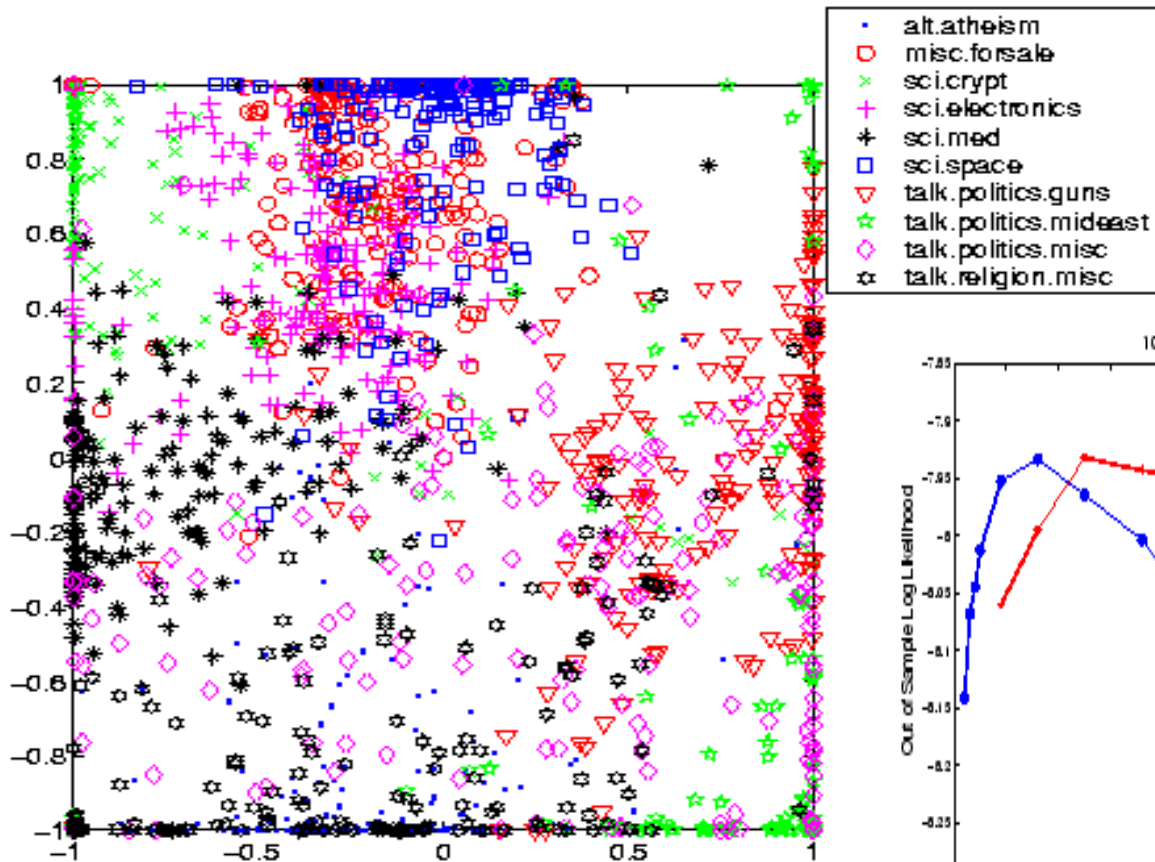
$$\mathbf{m}_{.m} = \sum_k \mathbf{a}_{.k} \varphi_k(\mathbf{x}_m)$$



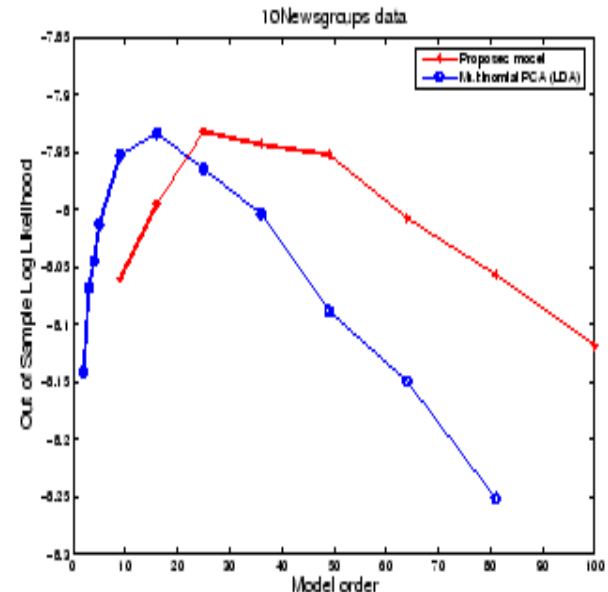
It can be shown that the model estimation algorithm minimises a weighted sum of entropies of the parameters.

Application 1: Predictive modelling and visualisation of large document collections (bag of words version)

$\{E[\mathbf{x}|\text{doc}]\}$ obtained from the 10Newsgroups text collection



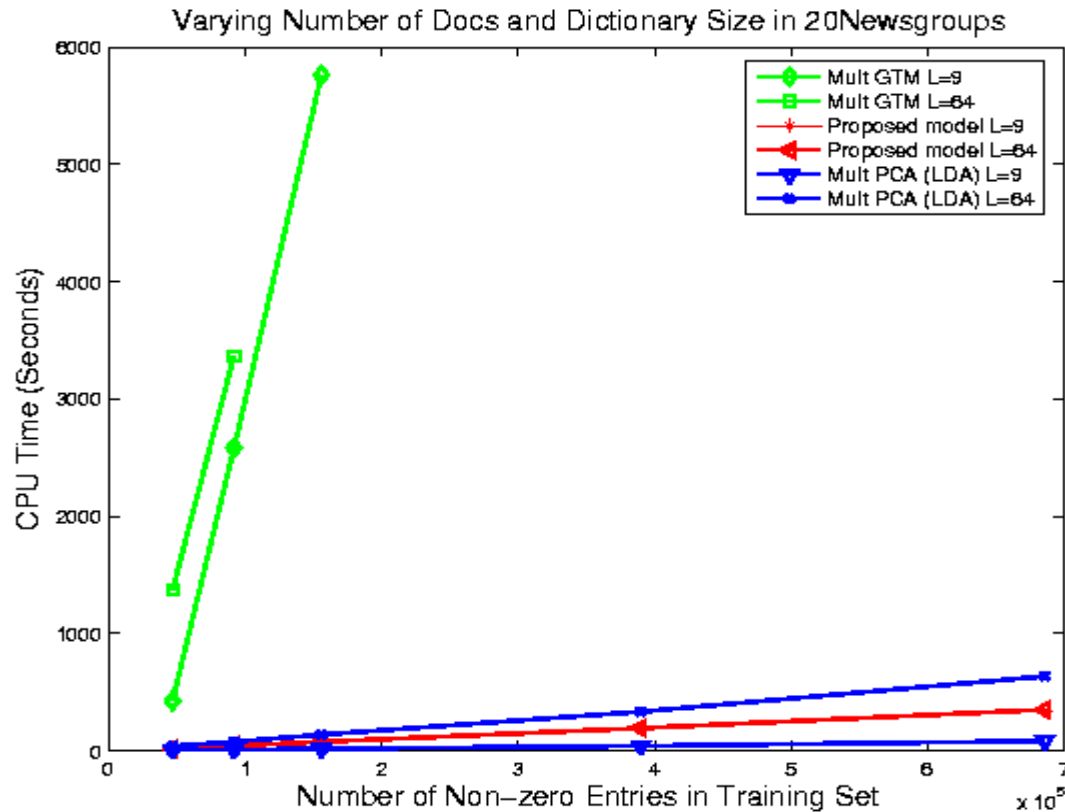
Predictive performance comparable with best state-of-the-art document model



kei encrypt secur privaci 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 insur	write articl work don year time ve good apr post articl	write articl thing apr don ve good ac mine time	fire articl write don make peopl ga time state thing peopl	law gun weapon govern arm drug polic firearm state constitut
health food dr result report msg state studi human	write peopl don make time good homosexu person question	write articl don peopl claim post make point thing love	fbi write koresh articl don make bd start apr	state govern american nation countri war polit senat hous year
parti abork frank convent valu polit dwyer object state price	moral object law scienc system human truth word de life	god christian exist religion jesu atheist belief argum faith bibl	jew kill jewish peopl live world work question nazi don	israel isra arab presid myer jew palestinian mr polici lebanes

Aspect-level map
of the estimated
topical
components at
equidistant
locations of the
latent space

CPU Time

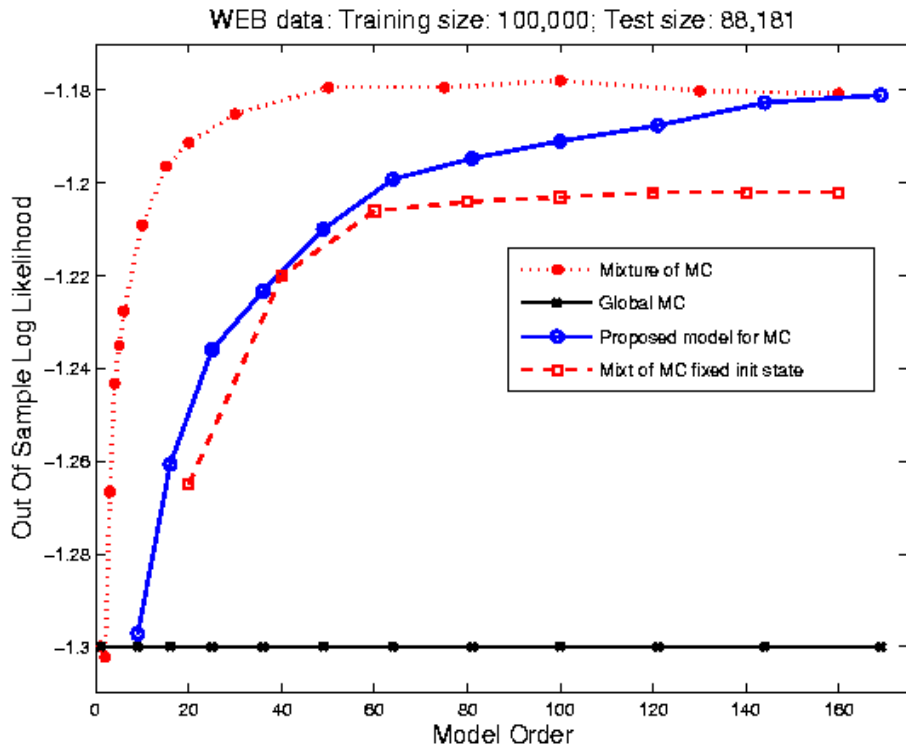


Computational demand drastically reduced in comparison with existing probabilistic topographic models for discrete data!

Application 2: Predictive modelling and exploratory analysis of dynamic user behaviour from a large web log collection

- Using the big mnhc.com web log sequence collection previously used in Cadez et al.
- Training on randomly chosen 100,000 user traces, totalling 801,745 page requests
- Testing on further, previously unseen 88,181 user trances, totalling 714,280 page requests
- Evaluation criteria used:
 - Generalisation (out of sample log likelihood)
 - Prediction (out of sample predictive perplexity) – varying sample size issues studied
 - Visualisation and exploratory analysis

Out of sample log likelihood (the higher the better): $\frac{1}{N_{test_set}} \sum_{n=1}^{N_{test_set}} \log P(Seq)$



Dash: Constrained mixture of MCs proposed in Cadez et al. for model-based visualisation

Solid line: our Topographic Mixture model

Dotted line: unconstrained mixture of MCs

SMMC not appropriate here (overfits) due to the extremely short sequence lengths (not shown)

Straight line: Global MC (1st order)

Although designed for exploratory purpose, the Topographic Mixture is a proper generative model – it outperforms the model constrained in an ad-hoc manner and approaches the unconstrained mixture probability model in terms of out of sample log likelihood. → We do not need to trade off prediction for exploratory abilities!

Using the model for prediction

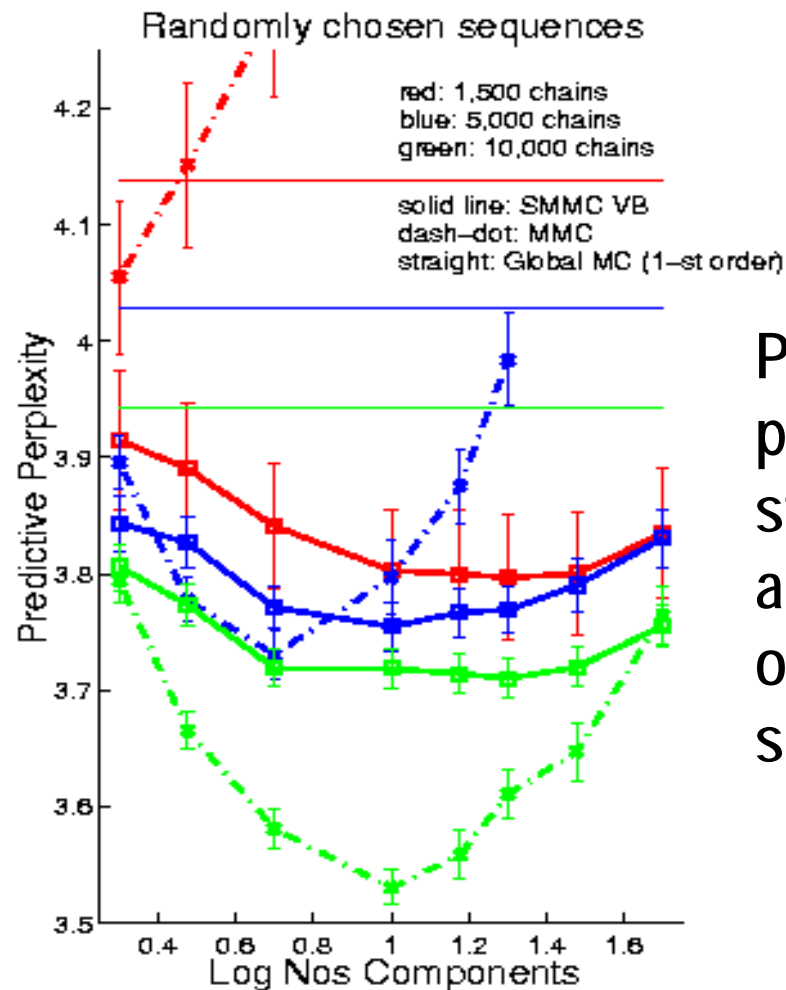
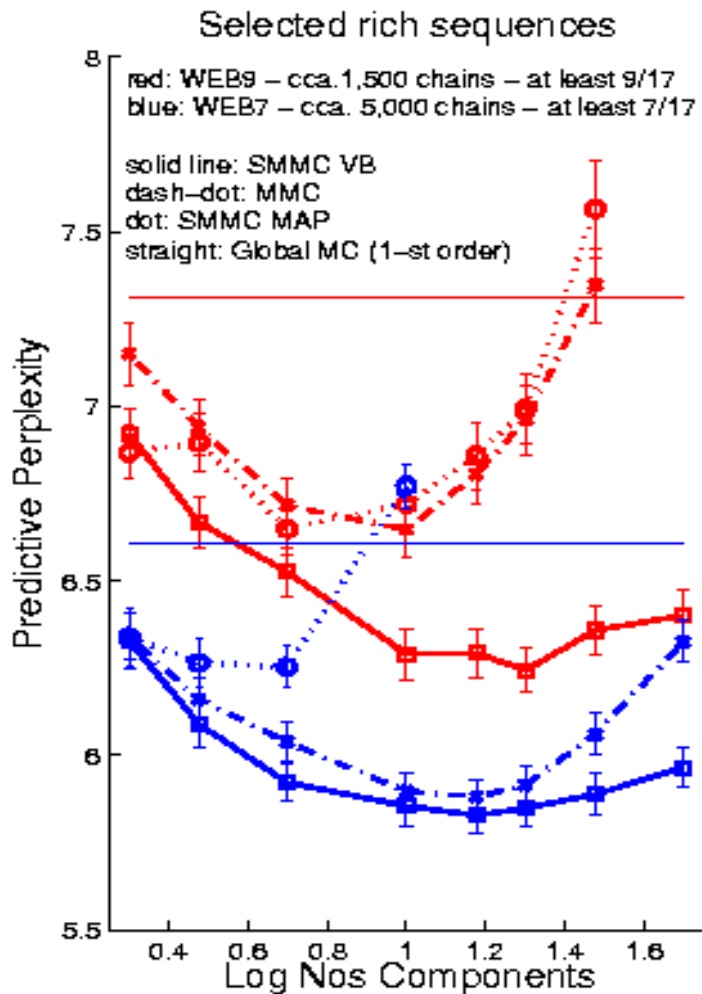
$$\begin{aligned} P(s_{next} | Seq_n) &= \int d\mathbf{x} P(s_{next} | s_{T_n}, \mathbf{x}) P(\mathbf{x} | Seq_n) \\ &= \int d\mathbf{x} \sum_{k=1}^K P(s_{next} | s_{T_n}, k) \varphi_k(\mathbf{x}) P(\mathbf{x} | Seq_n) \\ &= \sum_{k=1}^K P(s_{next} | s_{T_n}, k) E_{P(\mathbf{x} | Seq_n)}[\varphi_k(\mathbf{x})] \end{aligned}$$

- Combines basis-wise predictions in proportions specified by the posterior expectation
- User-specific deeper past (w.r.t. the global trait) is embodied in the posterior expectation!

Predictive perplexity (the lower the better): $\exp\left\{-\frac{1}{N_{test_set}} \sum_{n=1}^{N_{test_set}} \log P(s_{T+1}^{(n)} | Seq_{1:T}^{(n)})\right\}$

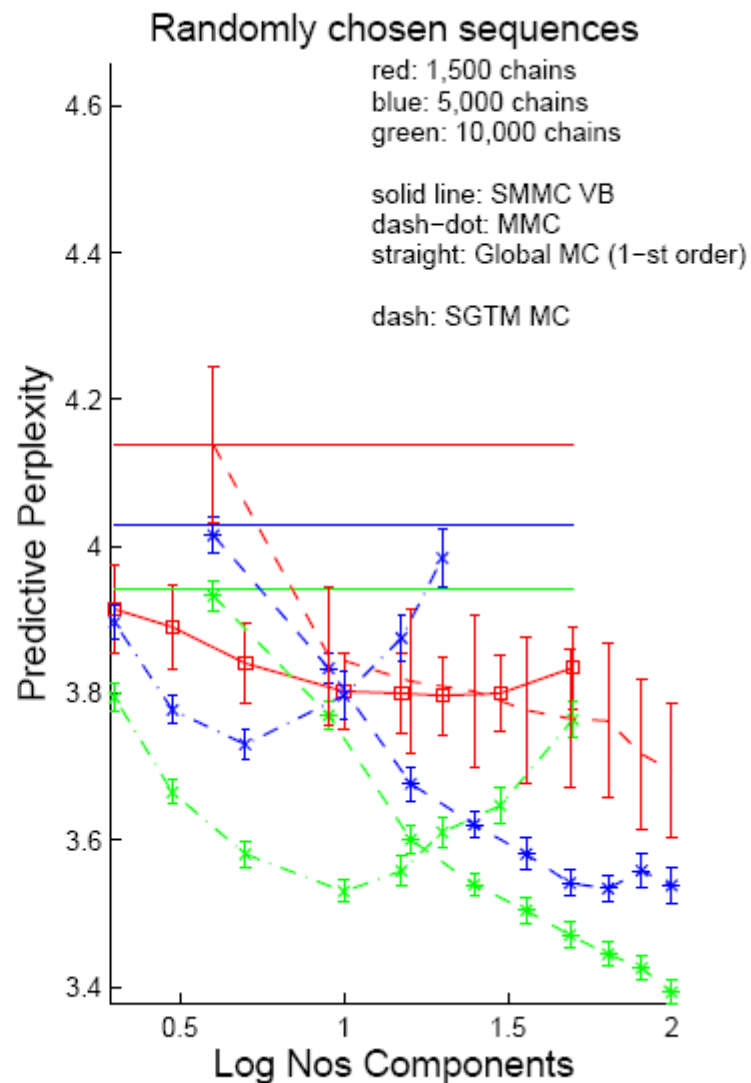
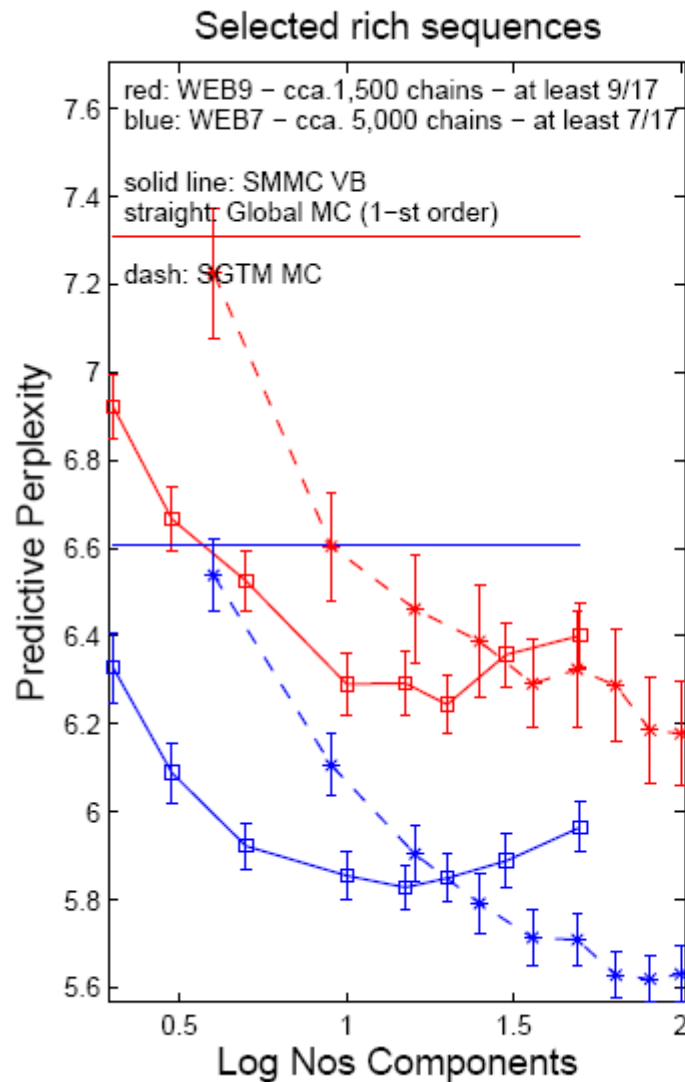
Sample size issues with existing methods

- The estimation of mixtures needs a large number of sequences
- The estimation of simplicial mixtures needs long (rich) sequences

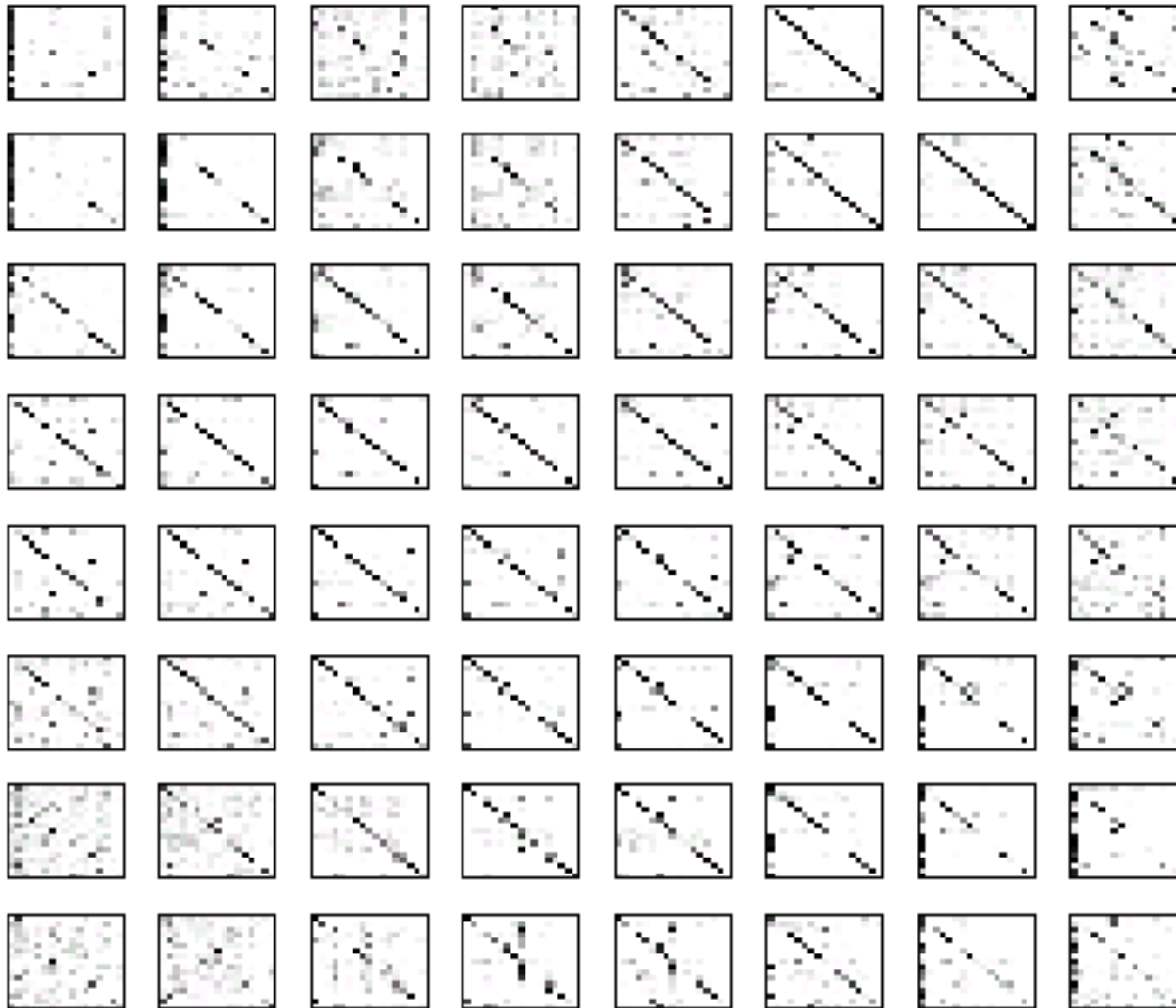


Prediction
perplexity of
state-of-the-
art models
of multiple
sequences

The proposed SGTm is robust to sample size issues & outperforms the state-of-the-art



Topographic organisation for exploratory analysis



**Map of state
transition
components
estimated from
the browsing
sequence data
set**

white=0
black=1

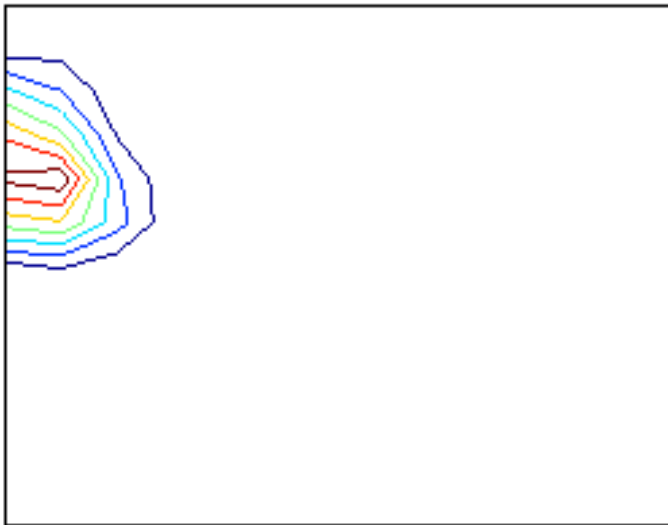
A summary overview of the large sequence collection in terms of lists of most probable sequences at equal locations of the map



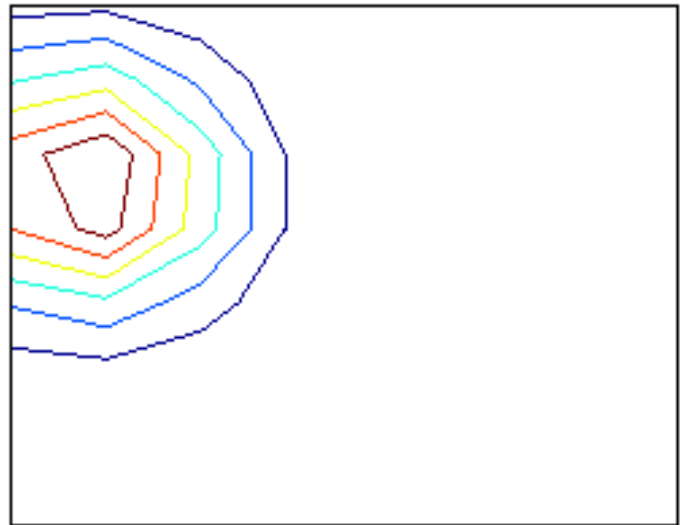
User profiles extracted from the same model

User Profile 1

$$\{P(\mathbf{x}_m | Seq_{68127})\}_{\{x_m\} \text{ equally spaced}}$$



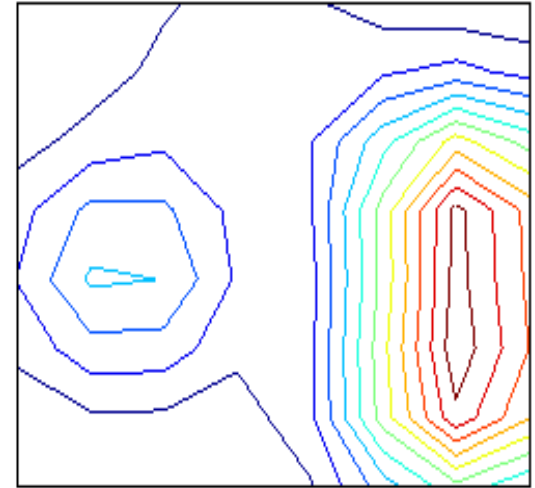
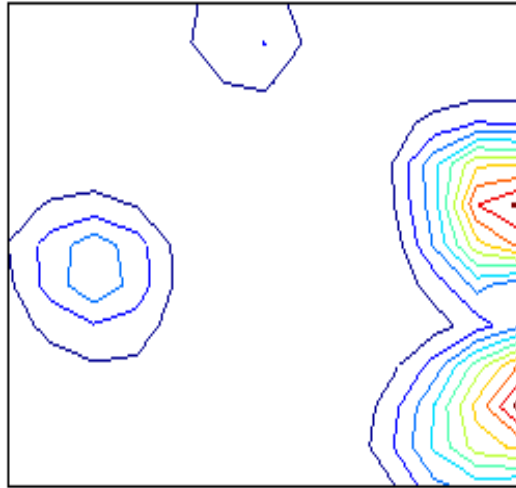
$$\{P(k | Seq_{68127})\}_{\{\mu_k\} \text{ equally spaced}}$$



Seq_{68127} =[frontpg misc misc frontpg misc misc misc misc misc misc misc misc misc

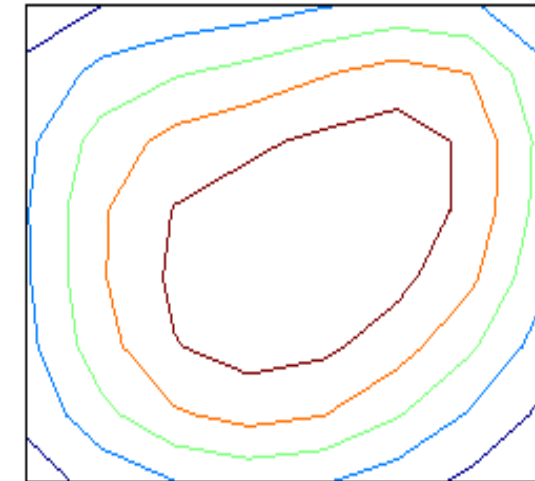
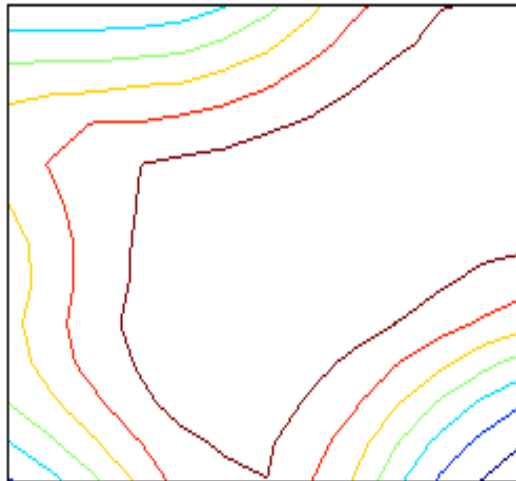
misc misc misc misc misc frontpg frontpg news news news onair misc misc]

User Profile 2



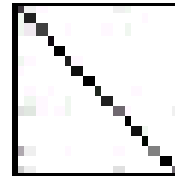
Seq_{83179} = [frontpg tech msnnews msnnews msnnews msnnews news sports msnnews]

User Profile 3



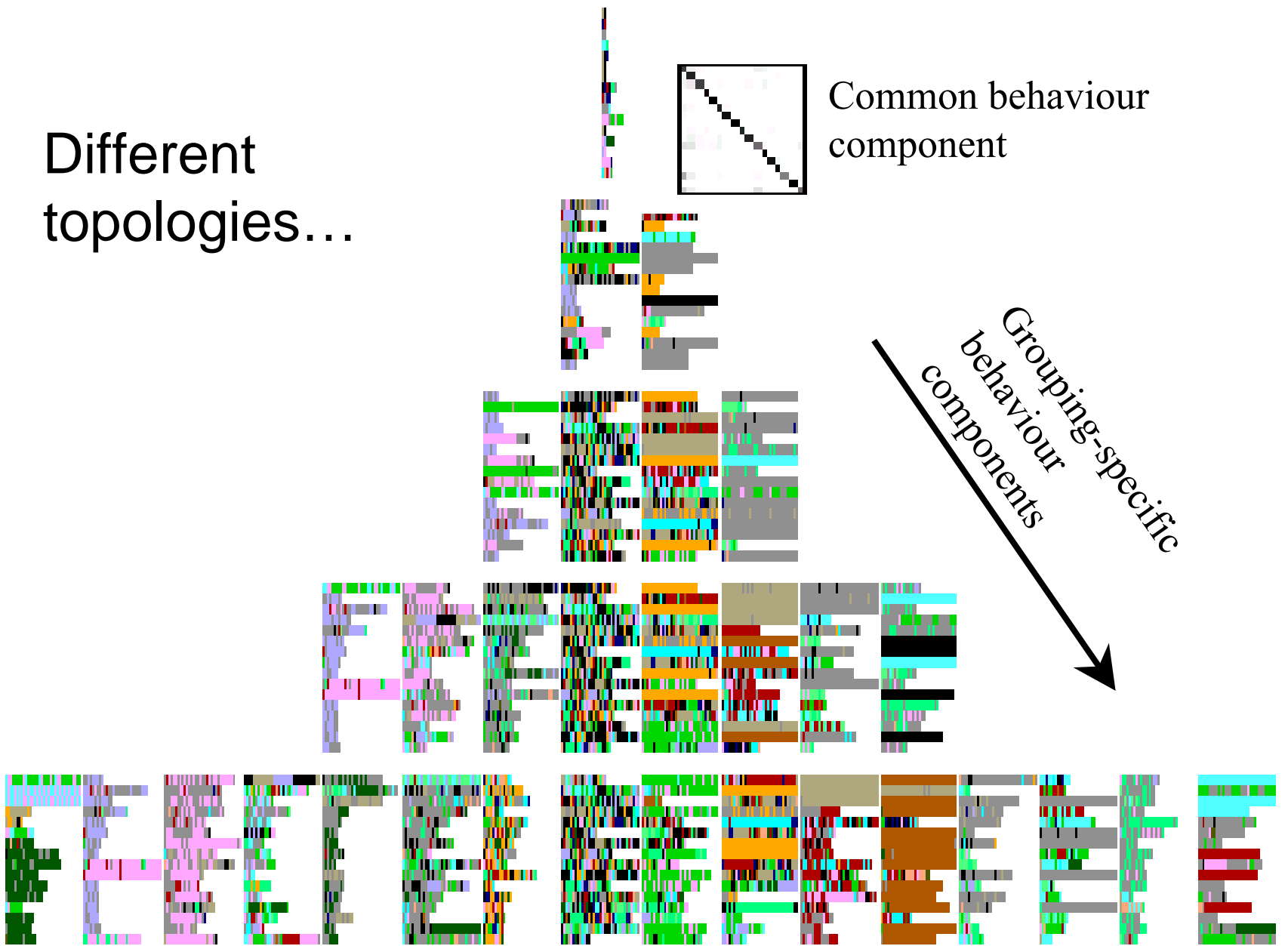
Seq_{1963} = [weather weather weather weather]

Different topologies...



Common behaviour component

Grouping-specific behaviour components



Conclusions

- Generative model
- Topographic mapping
- Single consistent probabilistic framework
- Simple efficient algorithm derived
- Representation characteristics studied
- Two application examples demonstrated
- Large heterogeneous sequence collections analysed
- Prediction & generalisation performance evaluated against the state of the art
- Predictive model
- Exploratory tool

Future Challenges

- “The most important goal for theoretical computer science in 1950-2000 was to understand the von Neumann computer. The most important goal for theoretical computer science from 2000 onwards is to understand the Internet”

Christos H. Papadimitriou