ICA @ cs.bham

Ata Kabán

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

A.Kaban@cs.bham.ac.uk

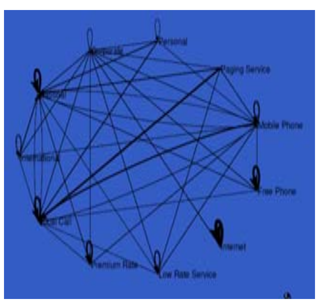
Roadmap

Young components in early-type galaxies

– With: M Harva (@hut),
  L Nolan, S Raychaudhury (@star.bham)

Phantom components in paleo-ecological recordings

– With: E Bingham (@hut), M Fortelius

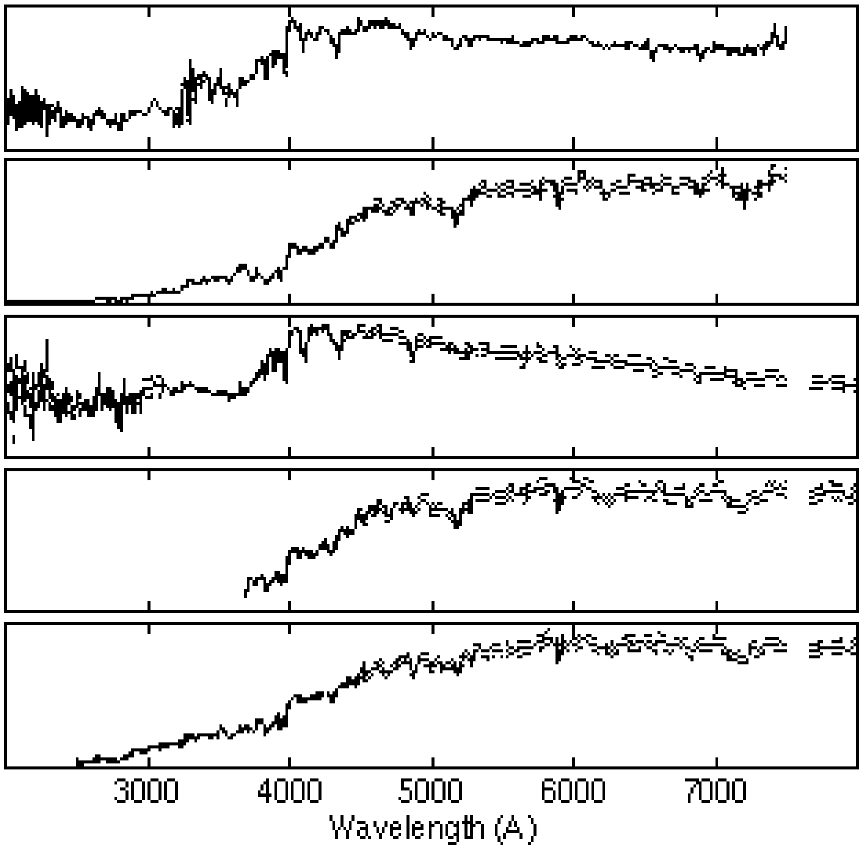


Low-entropy components in heterogeneous network usage

– With: M Girolami (@gla)

Kabán & Nolan and Raychaudhury SDM'05
Harva & Kabán, IJCNN'05; TecRep CSR-05-1, U.Bham.

Finding young components in early-type galaxies



Elliptical galaxies:

- oldest galactic systems

- believed to consist of a single population of old stars

- recent theories indicate the presence of younger populations of stars

- what does the data tell us?

Problem analysis

- Data:
  - One optical spectrum per galaxy
  - This is a linear superposition of the spectra of its billions of constituent stars
  - If both young and old stars exist in a galaxy, their emissions cannot be measured separately
- Inverse problem:
  - Find out spectral components that can explain all observed spectra by an unknown linear superposition
- Prior knowledge from the data characteristics
  - Spectral elements are positive
  - The flux at neighboring wavelength is likely to be similar
  - The concentration of the histogram of flux values does not tend to be skewed toward zero (not sparse)
  - No previously existing such model solution

Data driven model design

The factor model of the flux: $\mathbf{x}_t = \mathbf{A}f(\mathbf{s}_t) + \mathbf{n}$

Candidate modelling assumptions

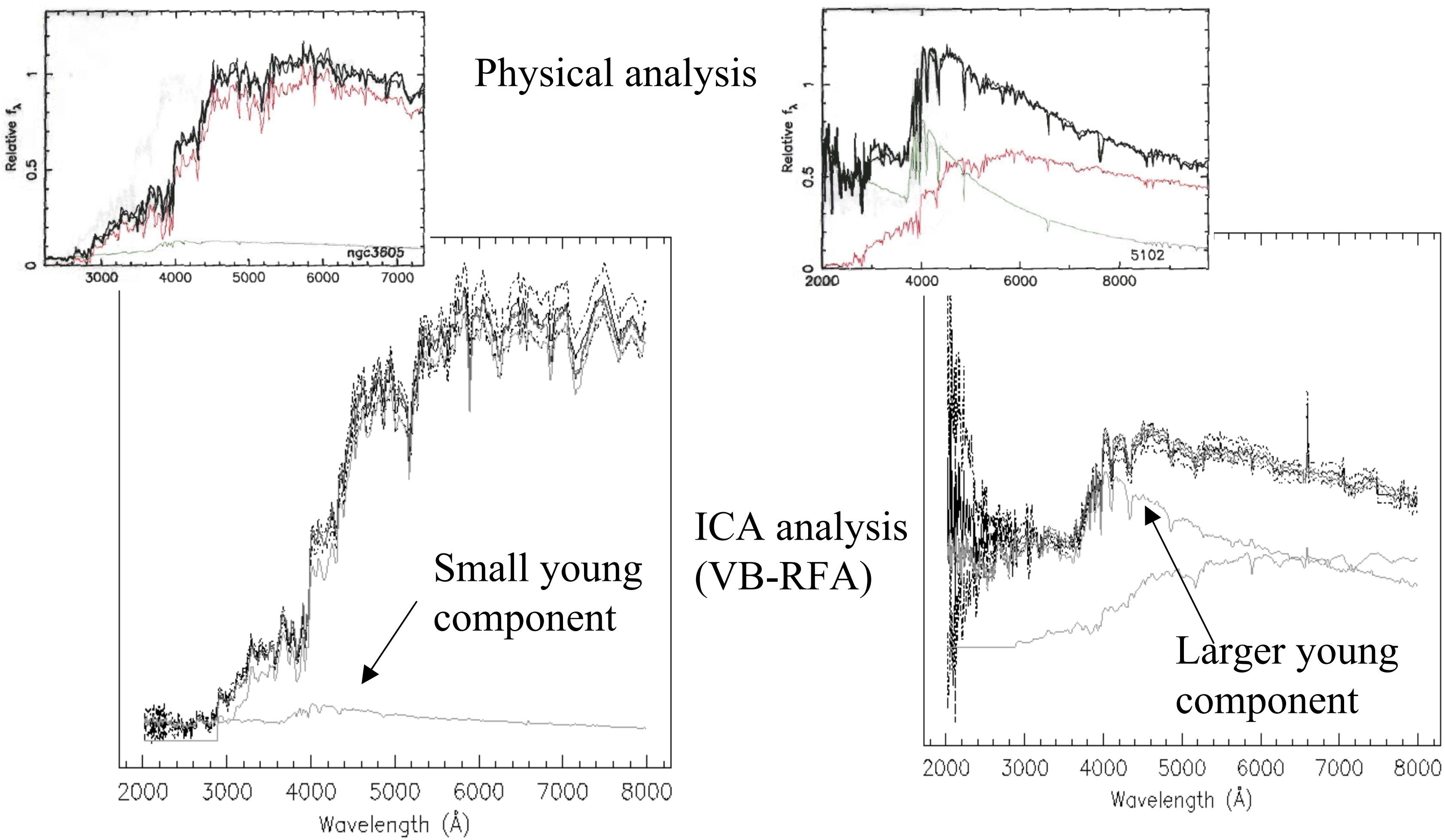
	Model	$f(s)$	$p(s)$	$p(a)$	v_{x_n}
baselines	VB-PCA	s	$\mathcal{N}(s \mid m_s, e^{-v_s})$	$\mathcal{N}(a \mid 0, 1)$	v_x
	VB-FA	s	$\mathcal{N}(s \mid m_s, e^{-v_s})$	$\mathcal{N}(a \mid 0, 1)$	v_{x_n}
Miskin	VB-PFA	s	$\mathcal{N}^R(s \mid 0, e^{-v_s})$	$\mathcal{N}^R(a \mid 0, 1)$	v_{x_n}
new	VB-RFA	$\max(0, s)$	$\mathcal{N}(s \mid m_s, e^{-v_s})$	$\mathcal{N}^R(a \mid 0, 1)$	v_{x_n}

Table 1: Distributional assumptions: VB-PCA: Variational Bayesian PCA; VB-FA: Variational Bayesian Factor Analysis; VB-PFA: Variational Bayesian Positive Factor Analysis; VB-RFA: Variational Bayesian Rectified Factor Analysis.

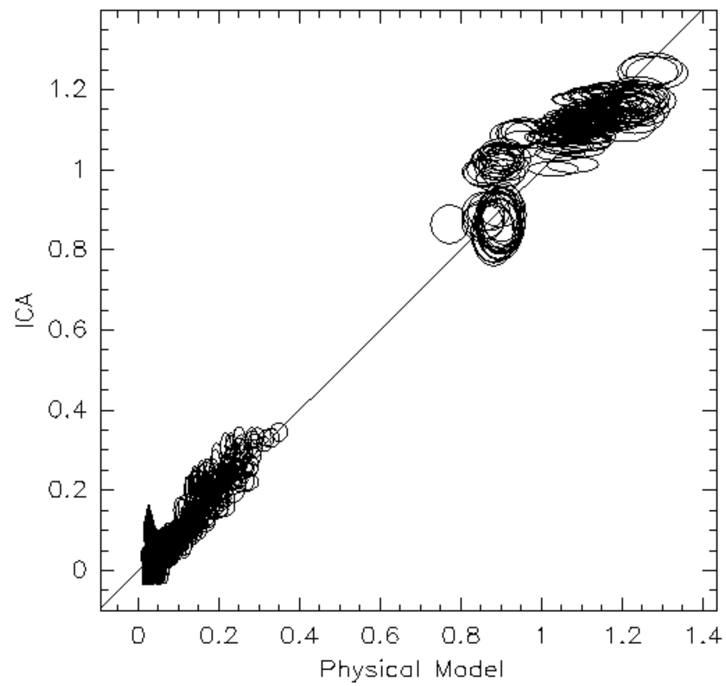
Known measurement errors →included in the model:

$y_{nt} = x_{nt} + e_{nt}$; x is now hidden var.; e's variances are known

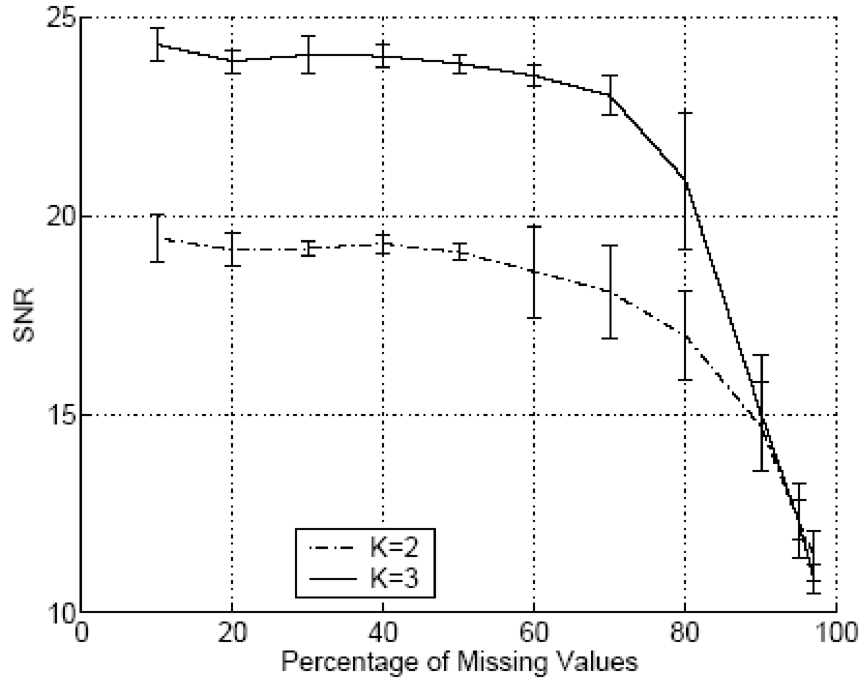
Examples of the reconstruction and analysis of spectra



Data model and physical model correlate in predicting missing entries in the real spectra



Missing value prediction on synthetic spectra



Summing up

- Very few astrophysical processes can be observed directly
- The use of component models may help verifying to what extent a hypothesis is supported by the data
- The assumptions needed for this may be derived from both domain knowledge and from data characteristics
- The probabilistic formulation and Bayesian estimation are tools to achieve good generalisation of component models

Details in: Kaban & Bingham & Hirsimaki, SDM'04
Bingham & Kaban & Fortelius, submitted

Finding phantoms in paleo-ecological recordings

- Results of paleo-ecological excavations are recorded as presences / absences of remains of genera on various sites
- Co-occurrences of genera is typically an indication of the relationship between them
- If a genera is found on a site then we can infer that it lived there
- If it is not found, what can we infer?
- Related scenarios:
  - Binary coded text messages
  - The two-choice questionnaire story
  - Corrupted binary images.
    - → How do we know whether the data is corrupted?

The Aspect Bernoulli Model

$$p(\mathbf{x}_n \mid \mathbf{s}_n, \mathbf{A}) = \prod_{t=1}^{T} \sum_{k=1}^{K} s_{kn} a_{tk}^{x_{tn}} (1 - a_{tk})^{1 - x_{tn}}$$

$$= \prod_{t=1}^{T} (\sum_{k=1}^{K} a_{tk} s_{kn})^{x_{tn}} (1 - \sum_{k=1}^{K} a_{tk} s_{kn})^{1 - x_{tn}}$$

$$where \quad a_{tk} = P(1 \mid k, t)$$

$$s_{kn} = P(k \mid \mathbf{s}_n)$$

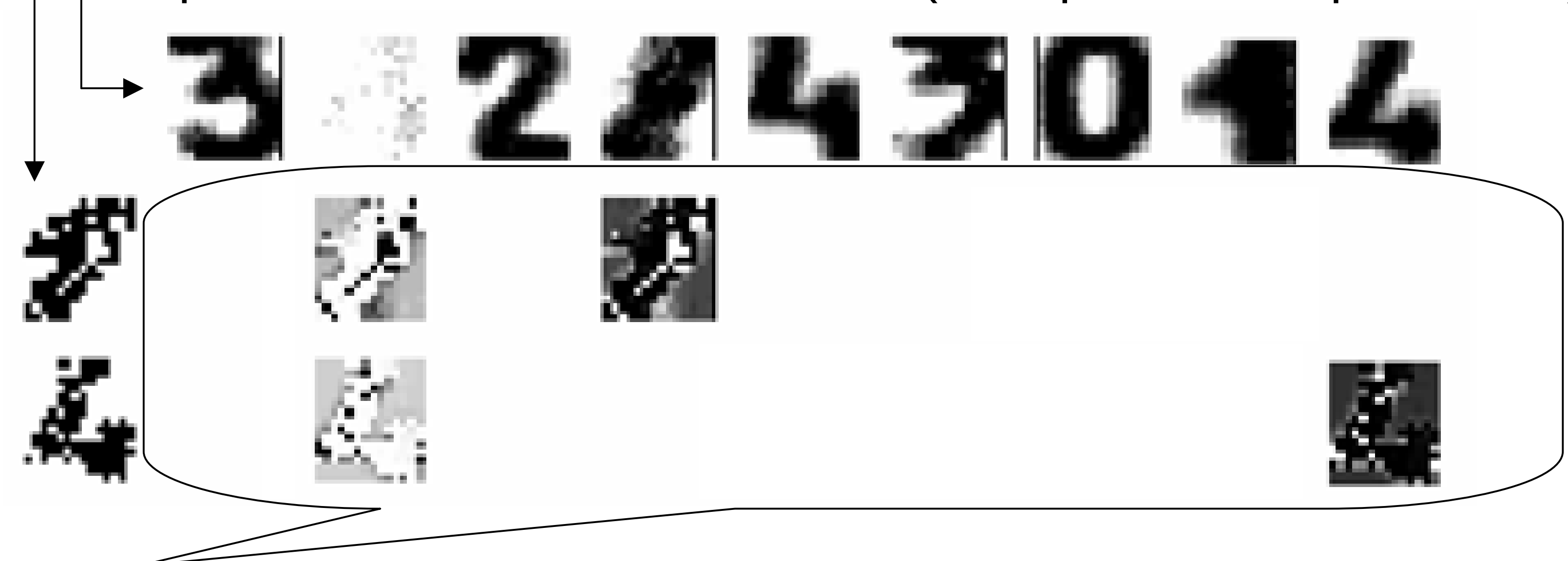
$a_k \sim Beta(a_k \mid \alpha_k, \beta_k)$ independent Beta priors can be assumed

Note. The convexity constraint $\sum_{\mathrm{k}} s_{\mathrm{k}} = 1$ of the mixing does not imply necessarily a single - cause model (it can be shown that it only implies a low - entropy mixing distribution)

Illustrative example: Explaining each pixel

Example input data instances

Components identified from the data (Beta posterior expectations)



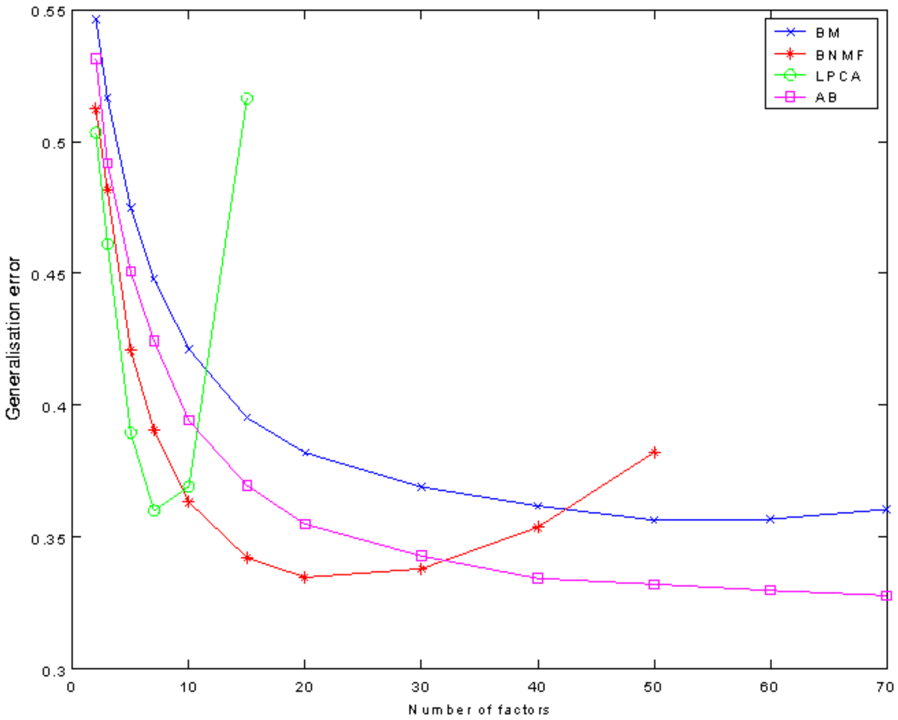
‘Explanation’ of each pixel value of each data instance in terms of how likely is it explained by any of the components. Darker means higher probability. Note that the white pixels in the corners of a raster are explained by ‘content-bearing’ components whereas the occluded pixels come from a ‘phantom’-component.

Text example: expanding short text messages

'govern' 'secur' 'access' 'scheme' 'system' 'devic' kei 0.99 encrypt 0.99 public 0.98 clipper 0.92 chip 0.91 peopl 0.89 comput 0.84 escrow 0.83 algorithm 0.76
'encrypt' 'decrypt' 'tap' system 1.00 kei 1.00 public 1.00 govern 0.98 secur 0.98 clipper 0.97 chip 0.97 peopl 0.96 comput 0.94
'algorithm' 'encrypt' 'secur' 'access' 'peopl' 'scheme' 'system' 'comput' kei 0.98 public 0.97 govern 0.92 clipper 0.87 chip 0.85 escrow 0.75 secret 0.63 nsa 0.63 devic 0.62
'peopl' 'effect' 'diseas' 'medicin' 'diagnos' medic 0.98 doctor 0.77 patient 0.75 treatment 0.71 physician 0.66 food 0.66 symptom 0.65 med 0.65 diet 0.65
'system' 'medicin' effect 0.97 medic 0.96 peopl 0.96 doctor 0.92 patient 0.92 diseas 0.91 treatment 0.91 physician 0.89 food 0.89

Generalisation performance

Negative out of sample log likelihood



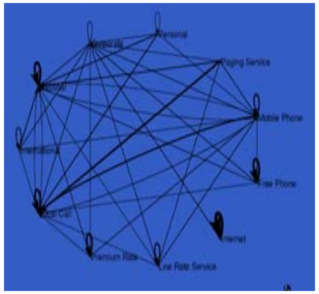
Generalisation performance

Results are averaged over all test examples.

Magenta line is the Aspect Bernoulli model

Summing up

- Multiple cause model for binary data that infers hidden causes of observing both ones and zeros
- Model with convex mixing allows us to compute posteriors of each feature of each observation being explained by each of the components
- Model with convex mixing implicitly implies a sparse mixing matrix
- Similarly it can be shown that the estimated components also tend to have low entropy
- These properties together result in ‘phantom’ components being created when there is noise in the binary data – the noise is separated out without requiring us to know a priori about it
- Good generalisation in the case of high-dimensional binary data because of being a factor model

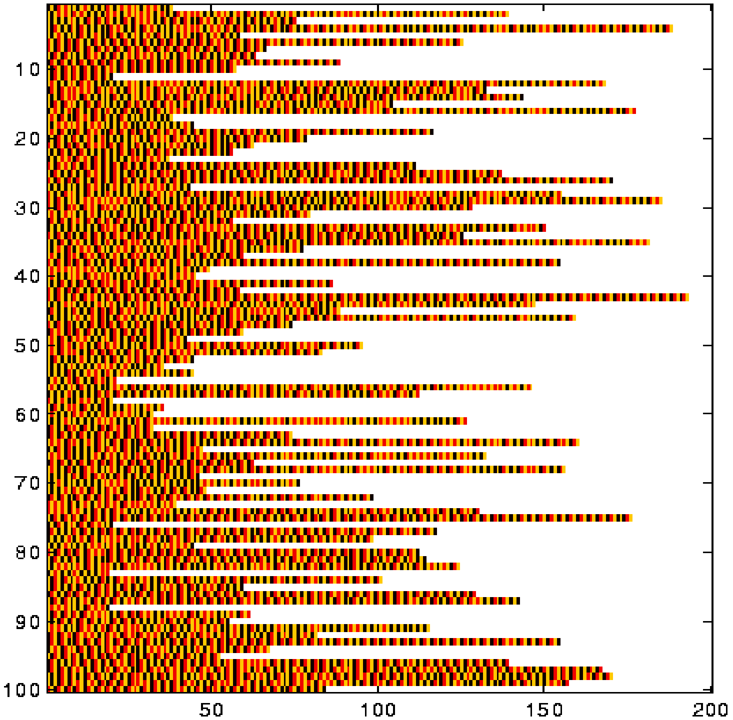


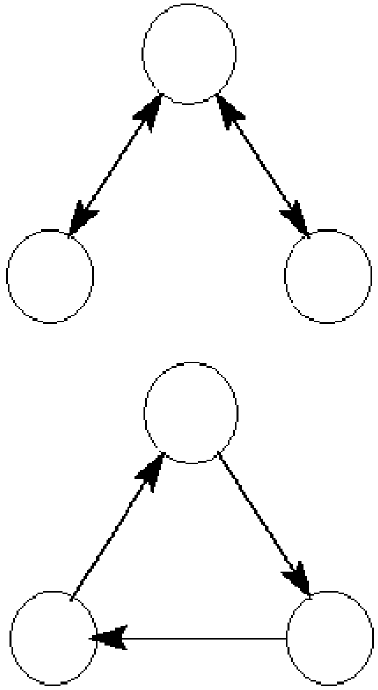
Finding low-entropy components in heterogeneous network usage

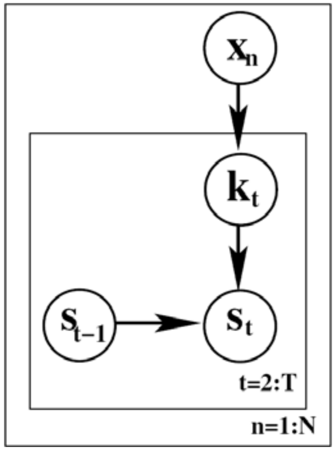
- Heterogeneous symbolic sequences over time
  - Finding shared behavioural patterns (analogous to procedures of computer programs)
  - These are the basis of multiple relationships between users and groups of users
  - Existing models are either global or assume homogeneous prototypical behaviour within groups

Details in: Girolami & Kaban, NIPS'03

Simplicial Mixtures of Markov Chains





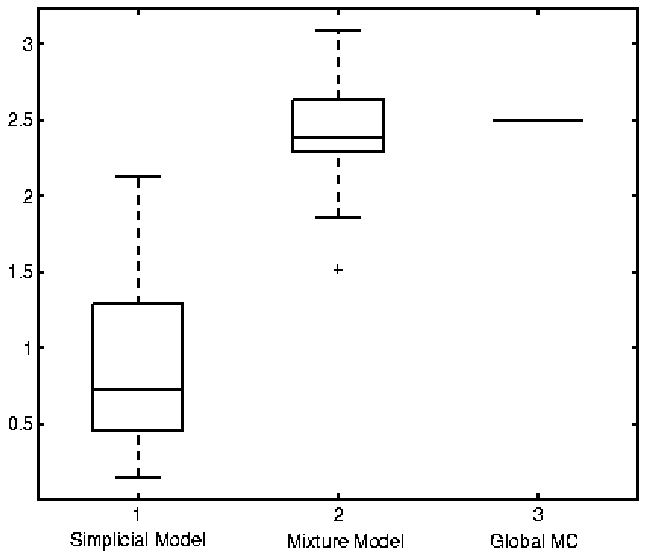


$$P(Seq^{(n)} \mid \mathbf{x}) = \prod_{t=1}^{T_n} \sum_{k=1}^{K} P(s_t^{(n)} \mid s_{t-1}^{(n)}, k) x_k$$

where $p(\mathbf{x}) = Dirichlet(\mathbf{x} \mid \boldsymbol{\alpha})$, therefore $x_k = P(k \mid \mathbf{x})$

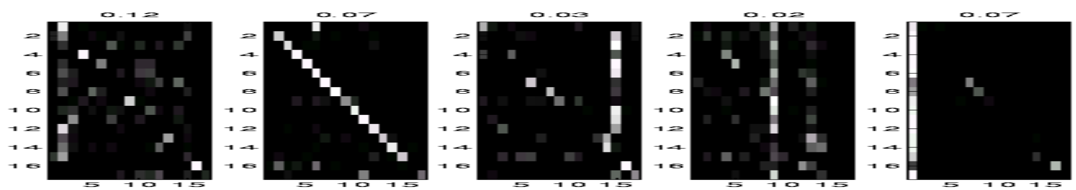
Low-entropy representations compared to mixtures and global models

Distribution of entropy rates of he dynamic components

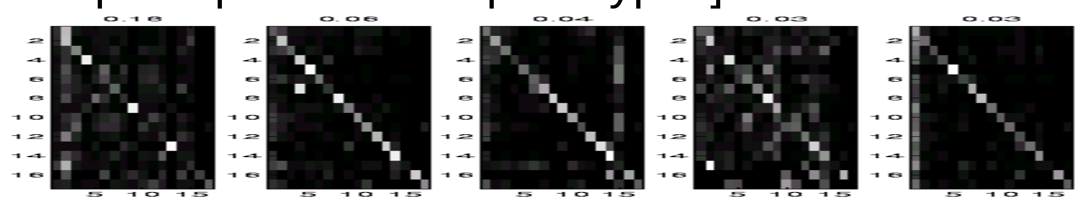


Selected basis-transitions obtained from web navigation sequences

SMMC component MCs [separates common behaviour in one component]

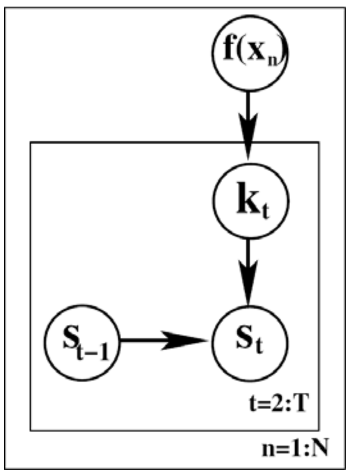


MMC cluster-prototypes [common behaviour superimposed on all prototypes]



black=0, white=1

Topographic Mixture of Markov Chains



$$P(Seq^{(n)} \mid \mathbf{x}) = \prod_{t=1}^{T_n} \sum_{k=1}^{K} P(s_t \mid s_{t-1}, k)\, \varphi_k(\mathbf{x})$$

$$where\ \ p(\mathbf{x}) = Uniform_{[-1,1]^2}(\mathbf{x})$$

and

$$\varphi_k(\mathbf{x}) = P(k \mid \mathbf{x}) \equiv \frac{\exp(-\frac{1}{2\sigma^2} \mid \boldsymbol{\mu}_k - \mathbf{x} \mid)}{\sum_{k'} \exp(-\frac{1}{2\sigma^2} \mid \boldsymbol{\mu}_{k'} - \mathbf{x} \mid)}$$

smooth function

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

- Using the big mnbc.com web log sequence collection previously used in Cadez et al.

- Training on randomly chosen 100,000 user traces, totalling 801,745 page requests

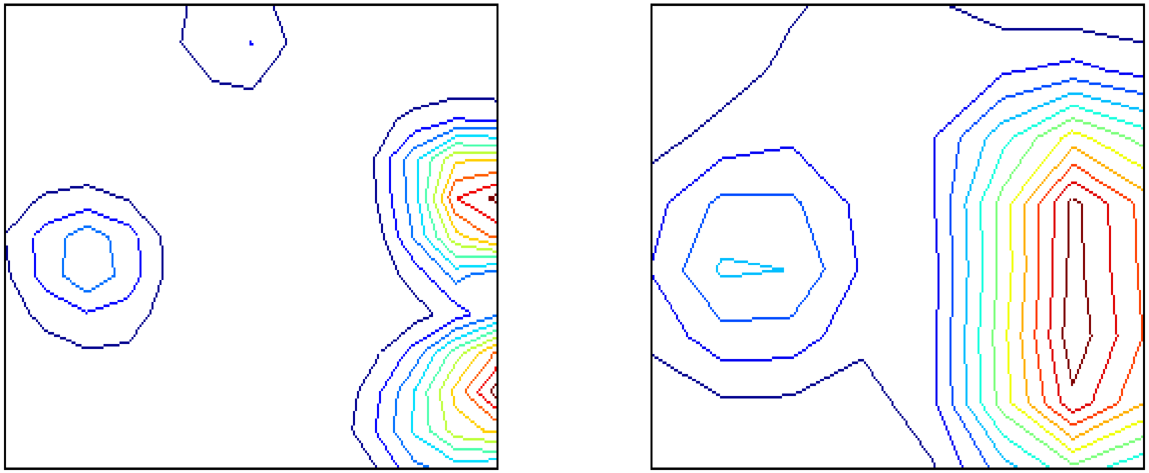
- Testing on further, previously unseen 88,181 user traces, totalling 714,280 page requests – prediction results equal or better than both MMC and SMMC.

- As a by-product, nice topographic data organisation emerges

Summary overview the web browsing activity estimated from 100,000 navigation sequences – the top probable sequences at equal locations of the latent space

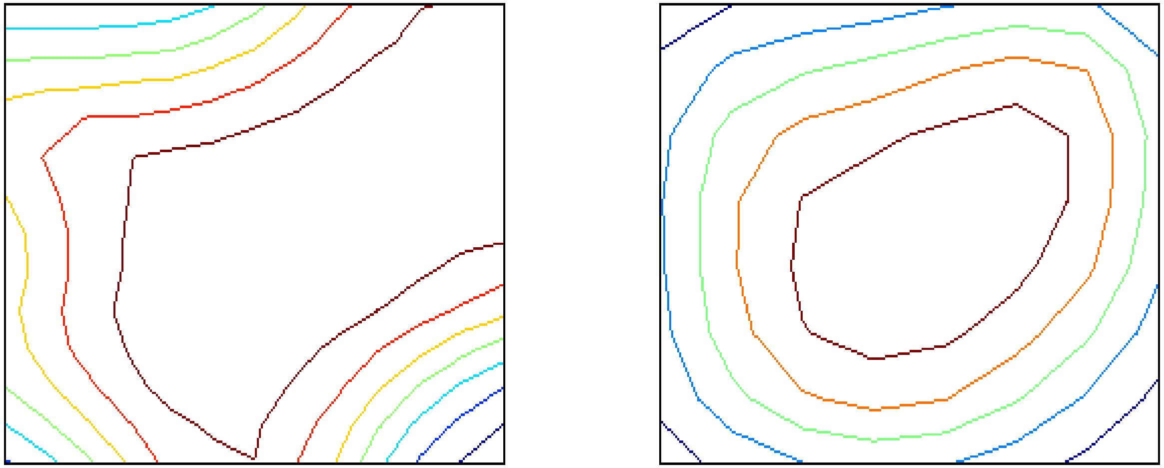
Individual user profiles extracted from the topographic mixture

User $Profile_{83179}$

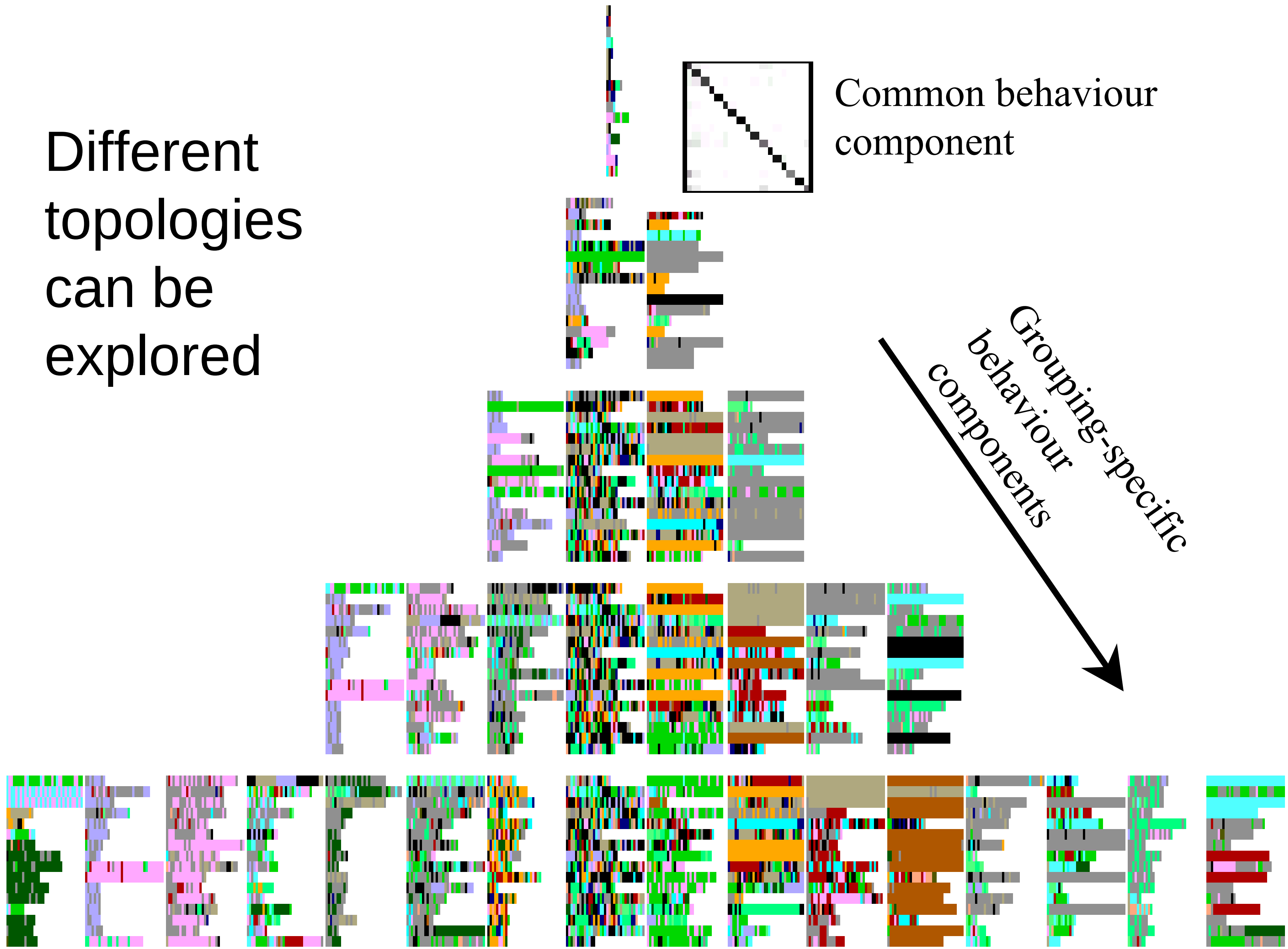


Seq_{83179}=[frontpg tech msnnews msnnews msnnews msnnews news sports msnnews]

User $Profile_{1963}$



Seq_{1963} =[weather weather weather weather]


Different topologies can be explored
Common behaviour component
Grouping-specific behaviour components