Independent Component Analysis in Multimedia Modeling

Jan Larsen

Intelligent Signal Processing Group
Informatics and Mathematical Modelling
Technical University of Denmark

www.imm.dtu.dk
jl@imm.dtu.dk

Motivation

- modeling of multimedia and multimodal data becomes increasingly important with the digitalization of the world
- signal detection/selection and pattern recognition are key components for new content based and context sensitive standards such as XML and MPEG7/MPEG21
- demonstrate the potential of ICA and BSS for modeling and understanding of multimedia data

Special multimedia signal processing session at ICA2003

1. *Independent component analysis in multimedia modeling*
   Jan Larsen, Lars Kai Hansen, Thomas Kolenda, Finn Årup Nielsen
2. *Learning the semantics of multimedia content with application to Web image retrieval and classification*
   Alexei Vinokourov, David Hardoon and John Shawe-Taylor
3. *Review of ICA and HOS methods for retrieval of natural sounds and sound effects*
   Shlomo Dubnov and Adiel Ben-Shalom
4. *Audio/visual independent componenets*
   Paris Smaradgis and Michael Casey
5. *A tentative typology of audio source separation tasks*
   Emmanuel Vincent, Cédric Févotte, Laurent Benaroya, Rémi Gribonval

Multimedia Modeling

from single medium (text/image/audio) to fusion of more media — heterogeneous multimodal fusion

The utility of ICA/BSS

- ICA/BSS are learning-based models, hence relevant for "intelligent multimedia processing", e.g., content based extraction and retrieval
- the independence assumption is amenable for interpretation and well-aligned with human perception of natural sound and images – cognitive component analysis
- ICA can be viewed as data projection onto latent subspace, i.e., feature extraction e.g., useful for data security and multimedia standards

Outline

- ICA/BSS multimedia analysis
- multimedia applications of ICA/BSS
- learning algorithms
- example of content extraction from combined text and images — the ICA assumption for text and images
- example of chat room topic spotting
- conclusions and future challenges

ICA/BSS multimedia analysis

Blind Source Separation based on

- spatial separation (direction of arrival)
- spectral separation (Wiener filter)
- time-frequency separation (masking)
- statistical independence

Linear mixing generative ICA model

$$\underset{\substack{P \text{ dim.} \\ \text{feature}}}{\boldsymbol{x}} = \underset{\substack{P \times K \\ \text{mixing mtx.}}}{\boldsymbol{A}} \cdot \underset{\substack{K \text{ dim.} \\ \text{source}}}{\boldsymbol{s}} + \underset{\substack{P \text{ dim.} \\ \text{noise}}}{\boldsymbol{\varepsilon}}$$

The ICA model in a multimedia perspective

- one medium – combined media
- model assumptions
- learning algorithms
  - maximum likelihood optimization
  - optimization of contrast functions, e.g., higher-order cumulants
  - kernel methods
  - Bayesian learning

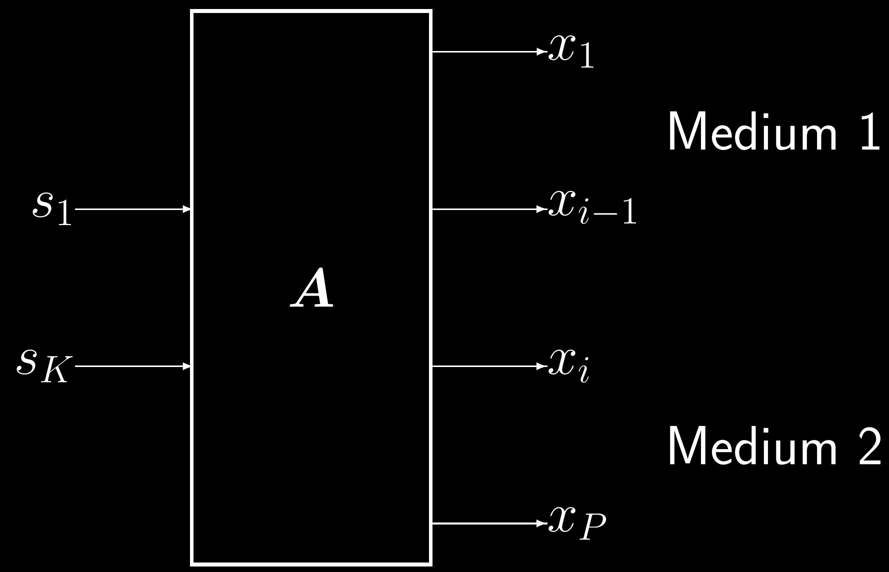
Single medium

- ICA is an extension of PCA and supplement to non-negative matrix factorization, independent and hierarchical factor analysis, ICA mixture model, projection pursuit, generative topographic mapping
- ICA is used for detection, extraction and explanation of hidden causes
- multimedia analysis results in interpretation of mixing matrix and sources

Example: If $\boldsymbol{x}$ grey values of image pixels, $\boldsymbol{a}_k$ is an eigenimage and s_k is the strength of the source.

Combined media

- ICA establishes a common latent space of combined media
- can be viewed as a supervised method for establishing relations between media
- extension of e.g., partial least squares, canonical correlation analysis



Model assumptions

- noise
  - no noise in image analysis
  - complex noise model in audio
- over- (#sensors>#sources) and under-determined cases
  - overdetermined in image and text
  - undertermined in audio
- convolutive
  - acoustical response in audio
  - filtering of image sequences
- constraints
  - incorporation of prior knowledge, e.g., non-negativity, spatial separation constraints
  - common and individual modality latent spaces in multimedia cases

Outline

- ICA/BSS multimedia analysis
- multimedia applications of ICA/BSS
- learning algorithms
- example of content extraction from combined text and images
  — the ICA assumption for text and images
- example of chat room topic spotting
- conclusions and future challenges

Multimedia applications of ICA/BSS

- most reported multimedia applications using ICA/BSS can also be tackled by other models
- main attraction is that ICA provides unsupervised grouping well-aligned with human perception

Multimedia applications of ICA/BSS

Medium	Topics	Refs.
Image/Video	natural scenes, feature extraction, noise reduction	8
	watermark detection	2
	content based retrieval	5
Multimodal brain data	EEG, MEG, fMRI	2
Audio	general	3
	auditory perception	2
	source separation, scene analysis	5
Text	document filtering, retrieval	5
Combined media	document content and inter-connectivity	3
	cross-language document retrieval	1
	combined text/image content extraction	2
	audio-visual segmentation	5

Outline

- ICA/BSS multimedia analysis
- multimedia applications of ICA/BSS
- learning algorithms
- example of content extraction from combined text and images
  — the ICA assumption for text and images
- example of chat room topic spotting
- conclusions and future challenges

Learning Algorithms

Routes to ICA

- non-Gaussianity
- time-correlation
- non-stationarity
- constraints (mixing matrix, source priors)

Two algorithms

- InfoMax
- Dynamic decorrelation

Likelihood and minimum mutual information

$$KL(p^*(\boldsymbol{x})\,;\,p(\boldsymbol{x};\boldsymbol{W})) = KL(p^*(\widehat{\boldsymbol{s}})\,;\,p(\widehat{\boldsymbol{s}};\boldsymbol{W}))$$
$$= KL(\prod_{i=1}^{m} p^*(\widehat{s}_i)\,;\,p(\widehat{\boldsymbol{s}};\boldsymbol{W}))$$

$$\widehat{\boldsymbol{s}} = \boldsymbol{A}^{-1}\boldsymbol{x} = \boldsymbol{W}\boldsymbol{x}$$

minimal MI between sources is optimal in likelihood sense

- approximate MI by kurtosis, skewness (Cardoso)
- expectation of fixed nonlinearities (Bell,Oja,Amari)

Linear ICA model

Noise free, quadratic, linear ICA

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$$

Ambiguity

$\boldsymbol{A}$ and $\boldsymbol{s}$ can be identified up to scaling (sign) and permutations

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} = \left[\boldsymbol{A}(\boldsymbol{P}\boldsymbol{S})^{-1}\right]\left[\boldsymbol{P}\boldsymbol{S}\boldsymbol{s}\right]$$

- $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{P}\boldsymbol{P}^{-1} = \boldsymbol{I}$ is a permutation matrix
- $\boldsymbol{S}$ is a diagonal scaling matrix

Likelihood

Generative model

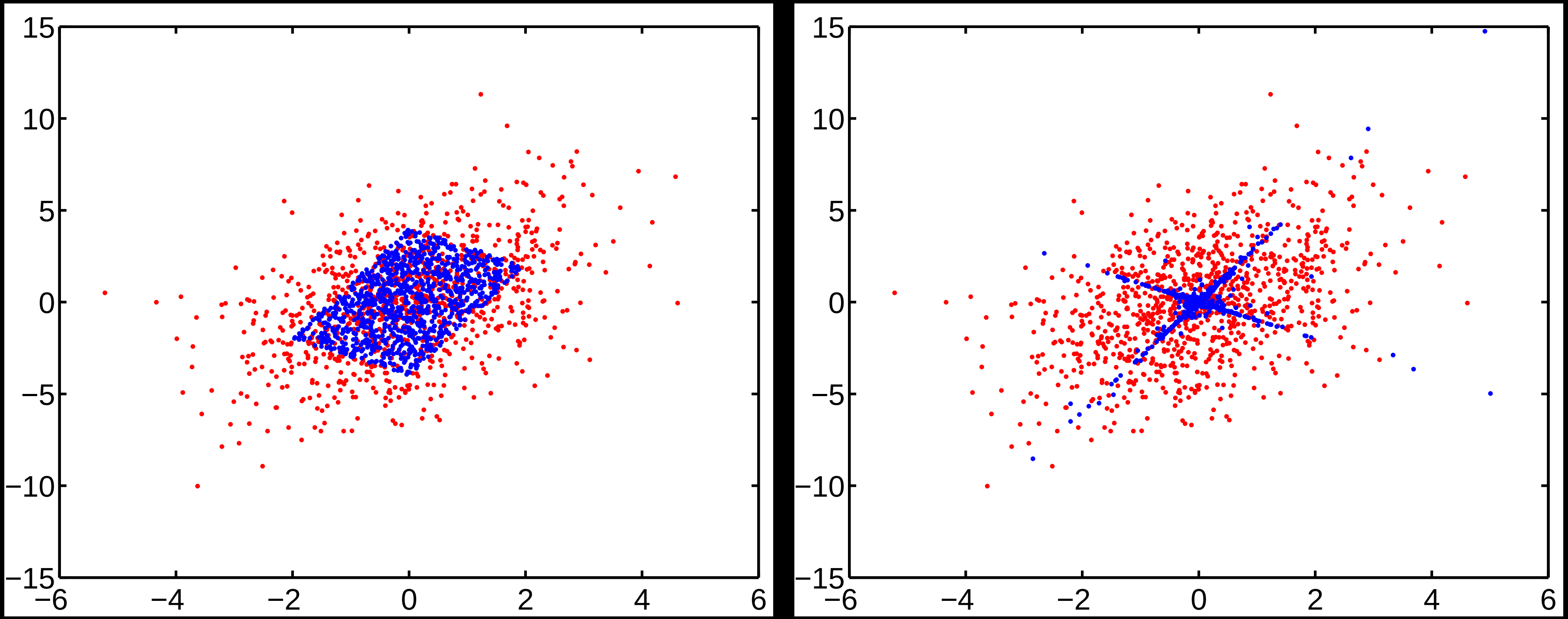
$$p(\boldsymbol{x}|\boldsymbol{A}) = \int \delta\left(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}\right) p_s(\boldsymbol{s})\, d\boldsymbol{s}$$

$p_{\boldsymbol{s}}(\boldsymbol{s})$ is prior sources distribution which is assumed non-Gaussian

Data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$

$$p(\boldsymbol{X}|\boldsymbol{A}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{A}) = \prod_{n=1}^{N} \int \delta\left(\boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{s}_n\right) p_s(\boldsymbol{s}_n)\, d\boldsymbol{s}_n$$

Why is non-Gaussianity required?



Why is non-Gaussianity required?

Independent sources and scale ambiguity implies

$$E\{\boldsymbol{s}\boldsymbol{s}^\top\} = \boldsymbol{I}$$

A multi-variate Gaussian is fully specified by its covariance

$$\begin{aligned} E\{\boldsymbol{x}\boldsymbol{x}^\top\} &= E\{\boldsymbol{A}\boldsymbol{s}\boldsymbol{s}^\top\boldsymbol{A}^\top\} \\ &= \boldsymbol{A}\boldsymbol{A}^\top = (\boldsymbol{A}\boldsymbol{Q})(\boldsymbol{A}\boldsymbol{Q})^\top \end{aligned}$$

where $\boldsymbol{Q}$ is a rotation matrix $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$

All mixing matrices $\widetilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{Q}$ give same measurement covariance matrix

Source estimation

Maximum aposteriori principle

$$\widehat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{A}) \propto$$

$$\begin{aligned} p(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{A}) &\propto p(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{s}) p_{\boldsymbol{s}}(\boldsymbol{s}) \\ &= \delta\left(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}\right) p_{\boldsymbol{s}}(\boldsymbol{s}) \end{aligned}$$

$\widehat{\boldsymbol{s}} = \boldsymbol{A}^{-1}\boldsymbol{x}$ independent on $p_{\boldsymbol{s}}(\boldsymbol{s})$

Choice of source distribution

Pham & Garat, 1997 studied asymptotic fluctuations due to mis-specified source distribution.

Under mild conditions asymptotic consistence is achieved

$$\boldsymbol{\delta} = \boldsymbol{I} - \widehat{\boldsymbol{A}}^{-1}\boldsymbol{A}, \quad vec(\boldsymbol{\delta}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}/N)$$

Contamination

$$\boldsymbol{s} = \boldsymbol{A}^{-1}\boldsymbol{x}, \quad \widehat{\boldsymbol{s}} = \widehat{\boldsymbol{A}}^{-1}\boldsymbol{x}$$

$$\boldsymbol{A}^{-1} - \widehat{\boldsymbol{A}}^{-1} = \boldsymbol{\delta}\boldsymbol{A}^{-1}$$

$$\boldsymbol{s} - \widehat{\boldsymbol{s}} = \left(\boldsymbol{A}^{-1} - \widehat{\boldsymbol{A}}^{-1}\right)\boldsymbol{x} = \boldsymbol{\delta}\boldsymbol{s}$$

$$\widehat{\boldsymbol{s}} = \boldsymbol{s}(\boldsymbol{I} - \boldsymbol{\delta})$$

Likelihood - continued

$$p(\boldsymbol{X}|\boldsymbol{A}) = \prod_{n=1}^{N} \int \delta\left(\boldsymbol{x}_n - \boldsymbol{A}\boldsymbol{s}_n\right) p_{\boldsymbol{s}}(\boldsymbol{s}_n)\, d\boldsymbol{s}_n$$

integration by substitution, $\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{s}_n$, $d\boldsymbol{v}_n = |\det \boldsymbol{A}| d\boldsymbol{s}_n$

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{A}) &= \prod_{n=1}^{N} \int \delta\left(\boldsymbol{x}_n - \boldsymbol{v}_n\right) p_{\boldsymbol{s}}(\boldsymbol{A}^{-1}\boldsymbol{v}_n) \frac{1}{|\det \boldsymbol{A}|}\, d\boldsymbol{v}_n \\ &= \frac{1}{|\det \boldsymbol{A}|^N} \prod_{n=1}^{N} p_{\boldsymbol{s}}(\boldsymbol{A}^{-1}\boldsymbol{x}_n) \end{aligned}$$

Maximum likelihood

$$\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} p(\boldsymbol{X}|\boldsymbol{A}) = \arg\max_{\boldsymbol{A}} \log p(\boldsymbol{X}|\boldsymbol{A})$$

Gradient ascent

$$\boldsymbol{A}^{(i+1)} = \boldsymbol{A}^{(i)} + \eta \cdot \frac{\partial p\left(\boldsymbol{X}|\boldsymbol{A}^{(i)}\right)}{\partial \boldsymbol{A}}$$

where η is a step-size

Gradient of log-likelihood

$$\log p(\boldsymbol{X}|\boldsymbol{A}) = -N \log|\det \boldsymbol{A}| + \sum_{n=1}^{N} \log p_{\boldsymbol{s}}(\boldsymbol{A}^{-1}\boldsymbol{x}_n)$$

$$\frac{\partial p(\boldsymbol{X}|\boldsymbol{A})}{\partial \boldsymbol{A}} = -N\boldsymbol{A}^{-\top} - \boldsymbol{A}^{-\top}\boldsymbol{A}^{-1}\sum_{n=1}^{N} \boldsymbol{x}_n \left.\frac{\partial \log p_{\boldsymbol{s}}(\boldsymbol{v})}{\partial \boldsymbol{v}^{\top}}\right|_{\boldsymbol{v}=\boldsymbol{A}^{-1}\boldsymbol{x}_n}$$

Natural gradient ascent

Gradient ascent is slow due to different length scales in gradient directions

- Natural gradient provides optimization in a non-orthonormal Riemannian space: $\|d\boldsymbol{\theta}\|^2 = d\boldsymbol{\theta}^\top \boldsymbol{G}(\boldsymbol{\theta}) d\boldsymbol{\theta}$, where $\boldsymbol{G}(\boldsymbol{\theta})$ is the (local) metric tensor.
- Euclidian orthonormal space: $\boldsymbol{G} = \boldsymbol{I}$

Steepest ascent direction

$\boldsymbol{\nabla} L = \arg\max_{d\boldsymbol{\theta}} L(\boldsymbol{\theta} + d\boldsymbol{\theta})$ subject to $\|d\boldsymbol{\theta}\|^2 = \epsilon^2$ sufficiently small

$$\boldsymbol{\nabla} L(\boldsymbol{\theta}) = \boldsymbol{G}^{-1}(\boldsymbol{\theta}) \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

If metric is the Fisher information matrix (Hessian matrix) $\boldsymbol{G}(\boldsymbol{\theta}) = \partial L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$ then natural gradient is the Newton direction

Natural gradient for InfoMax

$\boldsymbol{A}$ is in the matrix space of nonsingular $K \times K$ matrices, Amari, 1996

$$\nabla L(\boldsymbol{A}) = \boldsymbol{A}\boldsymbol{A}^\top \frac{\partial L(\boldsymbol{A})}{\partial \boldsymbol{A}}$$

Natural gradient ascent

$$\boldsymbol{A}^{(i+1)} = \boldsymbol{A}^{(i)} - \eta \left[N\boldsymbol{A} + \sum_{n=1}^{N} \boldsymbol{x}_n \left. \frac{\partial \log p_{\boldsymbol{s}}(\boldsymbol{v})}{\partial \boldsymbol{v}^\top} \right|_{\boldsymbol{v}=\boldsymbol{A}^{-1}\boldsymbol{x}_n} \right]$$

Choice of source distribution

$$p_{\boldsymbol{s}}(\boldsymbol{s}) = \prod_{i=1}^{K} p_{s_i}(s_i) = \prod_{i=1}^{K} \frac{1}{\pi \cosh(s_i)}$$

corresponds to

$$\frac{\partial \log p_{\boldsymbol{s}}(\boldsymbol{v})}{\partial \boldsymbol{v}} = -\tanh(\boldsymbol{v})$$

Overdetermined ICA

$$\boxed{\boldsymbol{x}} = \boxed{\boldsymbol{U}} \cdot \boxed{\boldsymbol{\Phi}} \cdot \boxed{\boldsymbol{s}} = \boxed{\boldsymbol{A}} \cdot \boxed{\boldsymbol{s}}$$

To stages: PCA succeeded by quadratic ICA

- $\boldsymbol{U}$ is the $P \times K$ matrix of K largest eigenvectors of the covariance of $E\{\boldsymbol{x}\boldsymbol{x}^\top\} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$
- $\boldsymbol{\Phi}$ is the $K \times K$ mixing matrix
- Quadratic ICA is thus performed in the subspace $\boldsymbol{y} = \boldsymbol{U}^\top \boldsymbol{x}$
- No noise case and K sources PCA is optimal. In the noisy case PCA gives an approximate solution.

Probabilistic Model Selection

Model Hypotheses

Consider a finite set of models enumerated by $m = 0, 1, \cdots, M$. $m = 0$ signifies the null-hypothesis corresponding to no non-trivial independent components

Bayes optimal model selection

Bayes optimal decision rule (under 0/1 loss function) leads to the optimal model

$$m_{opt} = \arg\max_{m} p(m|\boldsymbol{X})$$

$$p(m|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|m)P(m)}{\sum_{m=0}^{M} p(\boldsymbol{X}|m)P(m)}$$

is the probability of the model given data

Probabilistic Model Selection

Uniform Model Belief

In the case of equal model priors, i.e., $P(m) = 1/(M+1)$, the model selection concerns computing the **evidence** $p(\boldsymbol{X}|m)$

Evidence

$$p(\boldsymbol{X}|m) = \int P(\boldsymbol{X}, \boldsymbol{\theta}|m)\, d\boldsymbol{\theta} = \int p(\boldsymbol{X}|\boldsymbol{\theta}, m) p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

- $\boldsymbol{\theta}$ are model parameters
- $p(\boldsymbol{X}|\boldsymbol{\theta}, m)$ is the likelihood
- $p(\boldsymbol{\theta}|m)$ is the prior which is normally assumed vague and normalizable

Evidence approximation

Simpler than more involved methods like Laplace, variational Bayes and MCMC

Normalized log-posterior

$$C(\boldsymbol{\theta}) = \frac{1}{N} \left(\log p(\boldsymbol{X}|\boldsymbol{\theta}, m) + \log p(\boldsymbol{\theta}|m) \right)$$

and the maximum a posteriori (MAP) solution

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} C(\boldsymbol{\theta})$$

Approximations

Gaussian MAP approximation

$$C(\boldsymbol{\theta}) = C(\widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\boldsymbol{J}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top$$

$$\boldsymbol{J} = -\left.\frac{\partial^2 C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = O(1)$$

Laplace Approximation

$$\begin{aligned}
p(\boldsymbol{X}|m) &= \int \exp\left(NC(\boldsymbol{\theta})\right)\, d\boldsymbol{\theta} \\
&\approx \int \exp\left(NC(\widehat{\boldsymbol{\theta}}) - \frac{N}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\boldsymbol{J}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top\right) d\boldsymbol{\theta} \\
&\approx p(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}, m) \cdot p(\widehat{\boldsymbol{\theta}}|m) \cdot \left(\frac{2\pi}{N}\right)^{\frac{\dim(\boldsymbol{\theta})}{2}} \cdot |\boldsymbol{J}|^{-\frac{1}{2}}
\end{aligned}$$

Bayesian Information Criterion

Since $\boldsymbol{J} = O(1)$, the leading term for large N does not involve the often complicated Hessian, hence, the evidence is approximated as

$$p(\boldsymbol{X}|m) \approx p(\boldsymbol{X}|\widehat{\boldsymbol{\theta}}, m) \cdot p(\widehat{\boldsymbol{\theta}}|m) \cdot \left(\frac{2\pi}{N}\right)^{\frac{\dim(\boldsymbol{\theta})}{2}}$$

$$\dim(\boldsymbol{\theta}) = 1 + KN + K^2 + \frac{K(2P - K + 1)}{2} + P$$

On the likelihood for overdetermined case

- signal space spanned by the first K eigenvectors has full covariance structure
- noise space $\boldsymbol{e}$ spanned by the remaining $P-K$ eigenvectors assumed to be isotropic with noise variance

$$\sigma_\varepsilon^2 = (P-K)^{-1} \sum_{i=K+1}^{N} D_{ii}^2$$

- Assuming independence of signal and noise space we model

$$p(\boldsymbol{x}|\boldsymbol{\theta}, K) = p(\boldsymbol{y}|\boldsymbol{\Phi}, K)p(\boldsymbol{e}|\sigma_\varepsilon^2)$$

$\boldsymbol{y} = \boldsymbol{U}^\top \boldsymbol{x}$ is the signal space in which ICA is performed

Likelihood terms

Noise space likelihood

$$p(\boldsymbol{E}|\sigma_\varepsilon^2) = (2\pi\sigma_\varepsilon^2)^{-N(P-K)/2} \cdot \exp\left(\frac{-N(P-K)}{2}\right)$$

Signal space likelihood

$$p(\boldsymbol{Y}|\boldsymbol{A}, m) = p_{\boldsymbol{s}}\left(\left(\boldsymbol{U}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{Y}\right) \frac{1}{\left|\det\left(\boldsymbol{U}^\top \boldsymbol{A}\right)\right|^N}$$

Outline

- ICA/BSS multimedia analysis
- multimedia applications of ICA/BSS
- learning algorithms
- example of content extraction from combined text and images
  — the ICA assumption for text and images
- example of chat room topic spotting
- conclusions and future challenges

Example of content extraction from combined text and images

- content based retrieval of combined text and image data from webpages
- mutual support between media are ensured by adjacency
- we want to demonstrate a synergistic effect
- components can be interpreted by providing key features (words,color,texture)

Modeling Framework

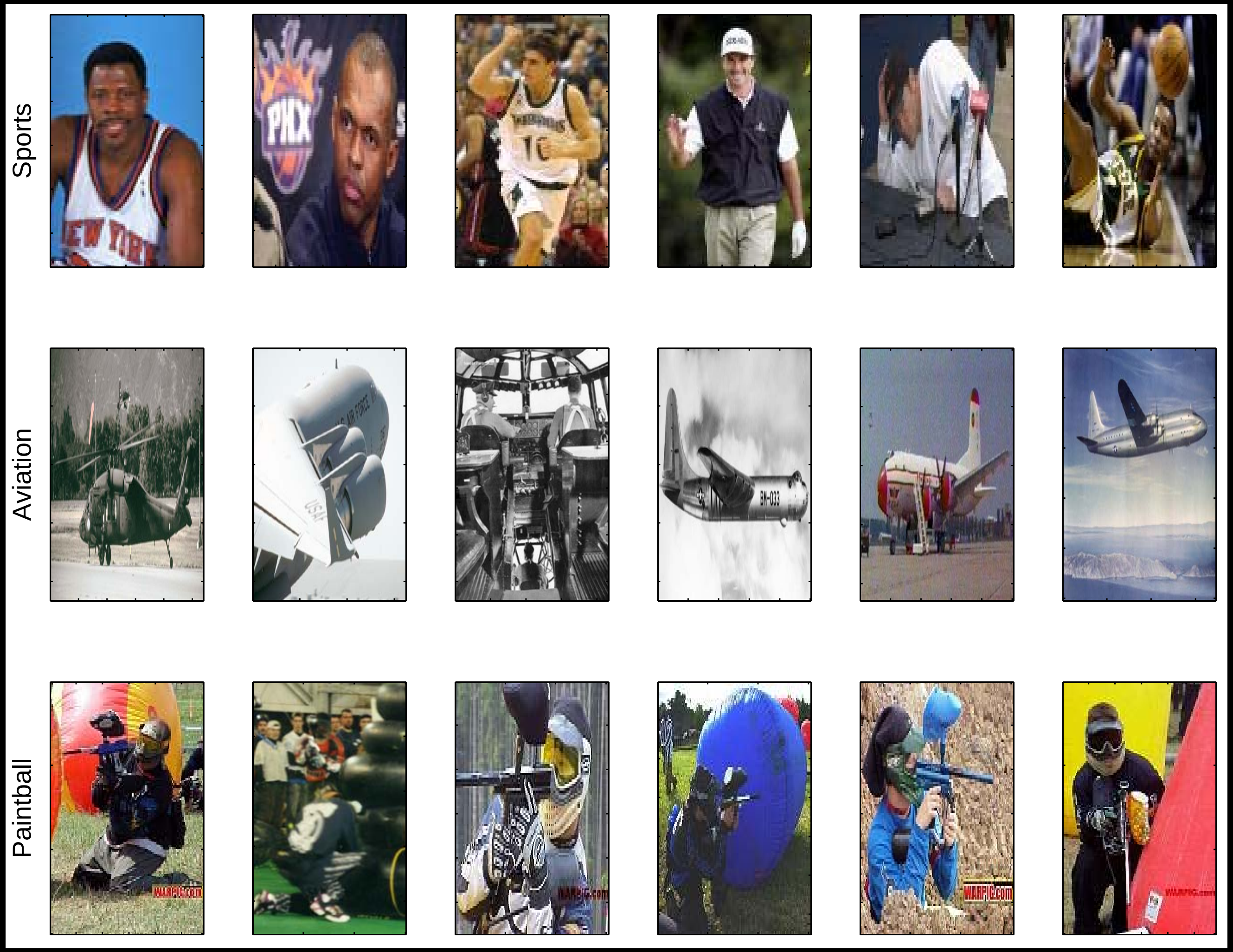
$$\boldsymbol{z} = [\boldsymbol{z}_I; \boldsymbol{z}_T]$$

is the column vector of image (I) and text (T) features.

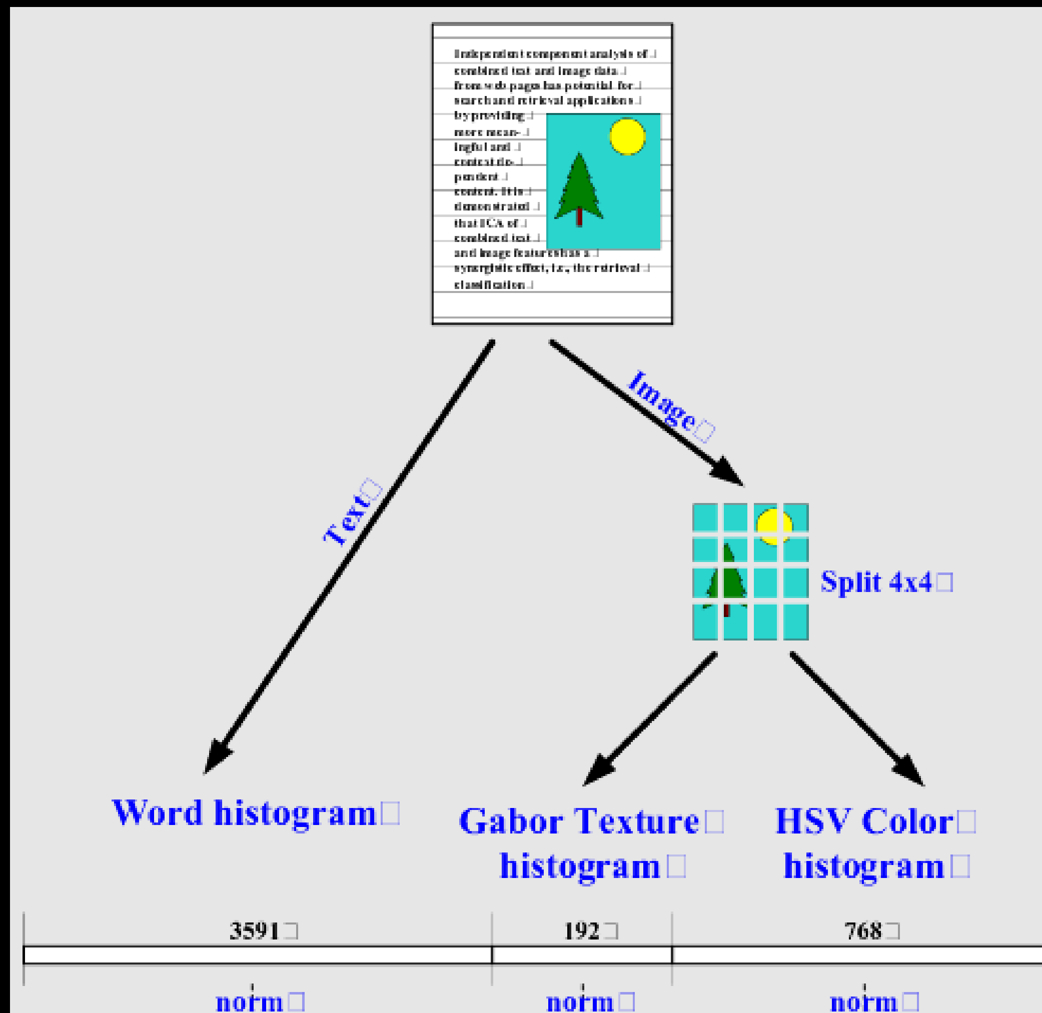
Unsupervised ICA provides a probability density $p(\boldsymbol{z})$ model from which meaningful clusters are identified

Supervised modeling is the conditional class-feature probability, $p(y|\boldsymbol{z})$, where $y = \{1, 2, \cdots, C\}$ is the class label.
A simple classifier can be obtained from the unsupervised ICA

Image examples from three classes



Extracting features



Text features

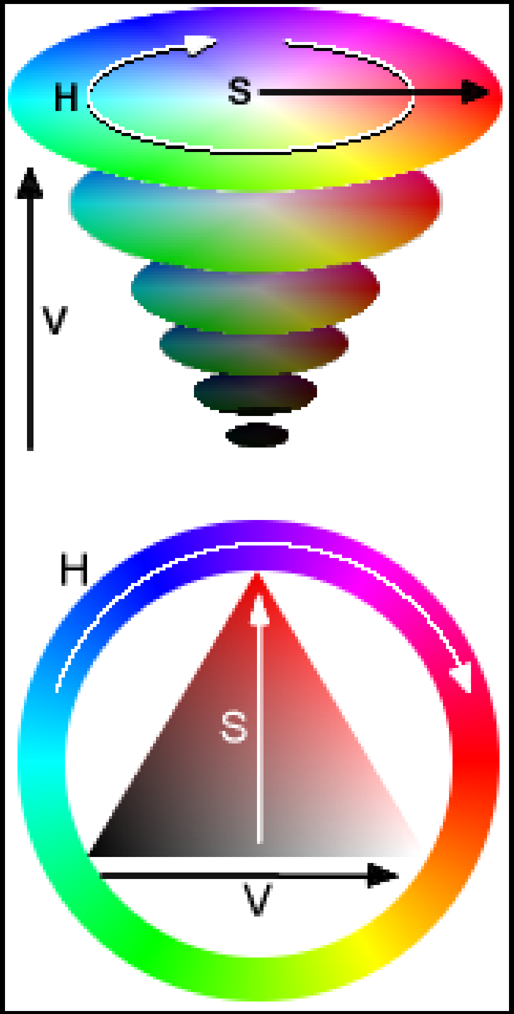
- collection of terms adjacent to an image is represented by a vector: the histogram of term occurrence
- filtered by removing low frequency words
- filtered by removing high frequency words - stop words. manually constructed to form a list of 585 words
- stemming by merging words with different endings, e.g., ing or s

Image features

- intention is to employ Vector Space Model on image features
- suggest low level image features of the ISO/IEC MPEG-7 standard
- image is divided into 4×4 patches to increase sensitivity

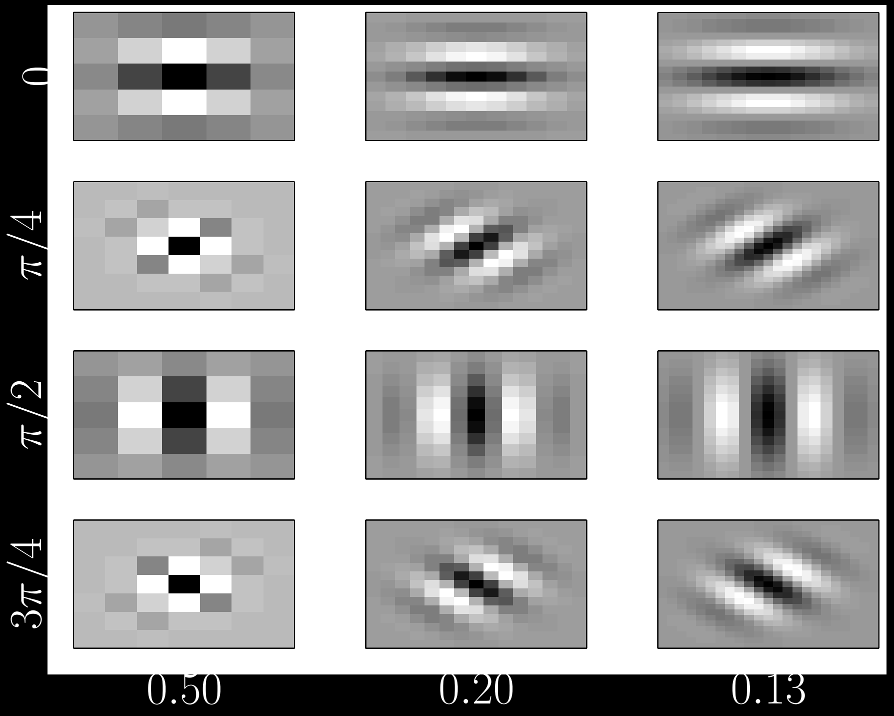
Color

- Hue-Saturation-Value color space (better than RGB)
- Hue is dominant wavelength,
- color components quantized into 16 levels
- $48 = 16{\cdot}3$ features for each of the 16 patches, i.e., $48 \cdot 16 = 768$ features

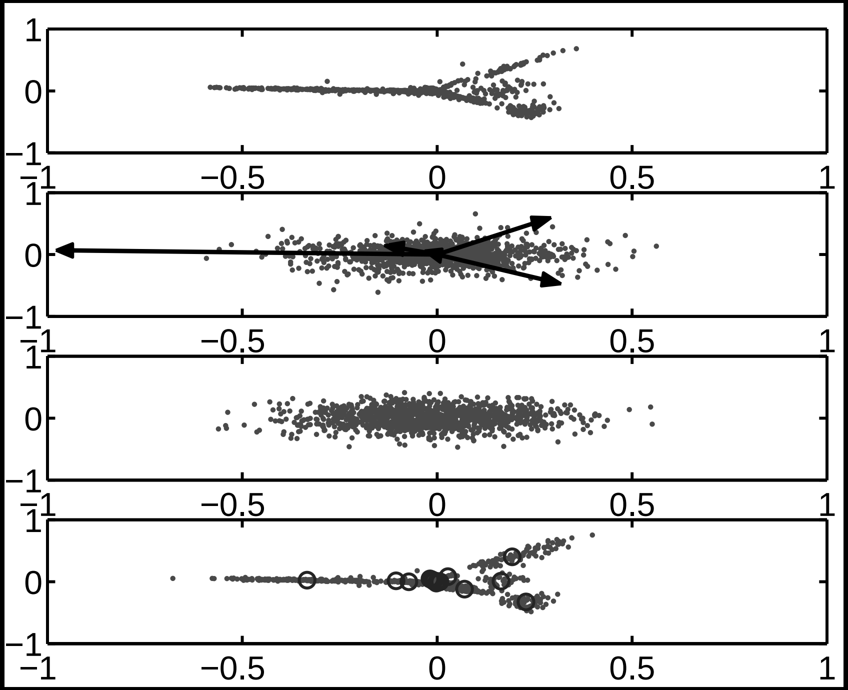


Texture

- Gabor filter bank
- filter output captures a specific texture frequency and direction
- energy of filtered outputs are the texture feature
- total of 16 patches times 12 filters, 192 texture features



The ICA Assumption



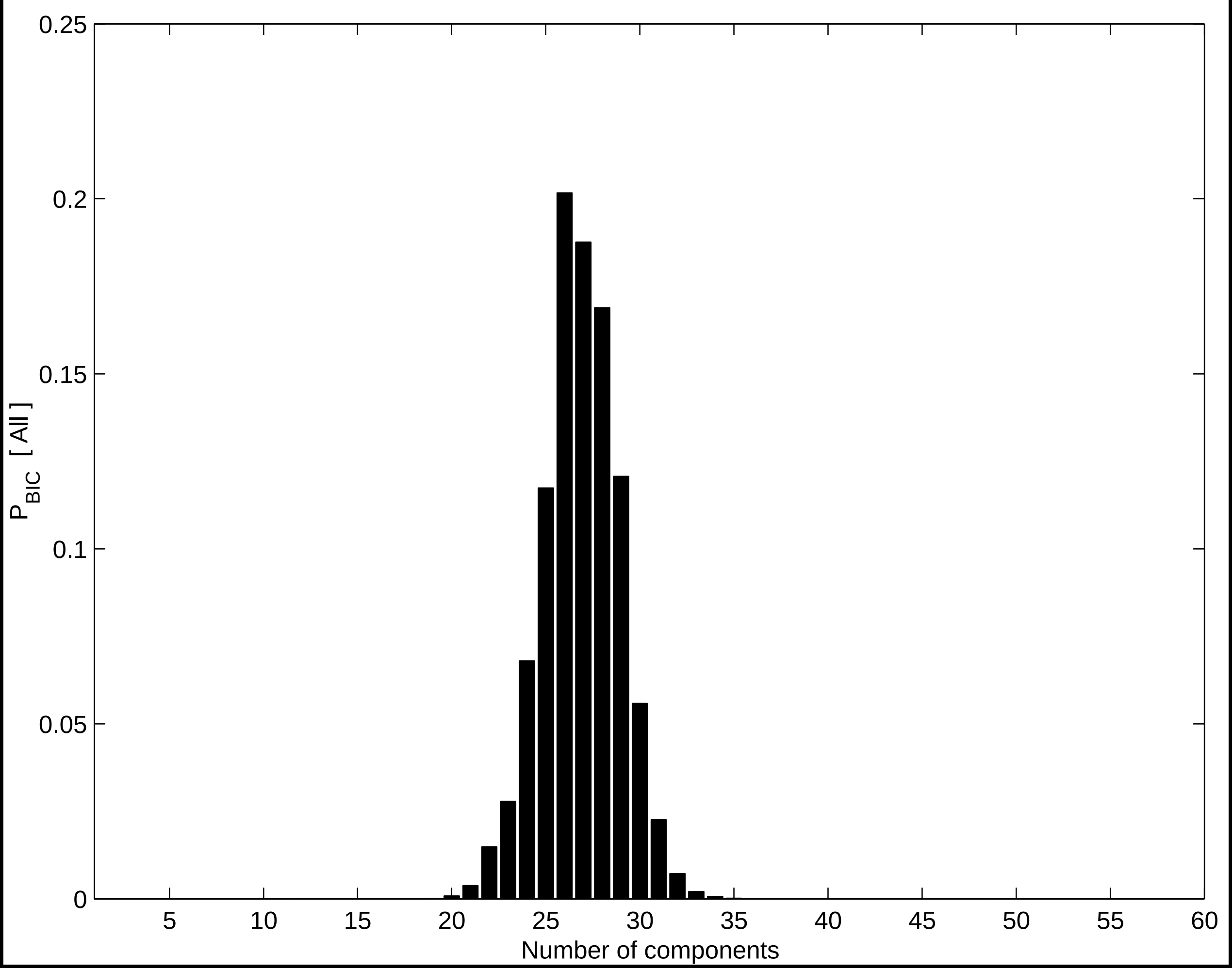
Panel 1: scatterplot of text data along two PC's show ray structure

Panel 2: 5 component InfoMax and **prior** source scatterplot $\propto 1/\cosh(s_k)$. **posterior** source scatterplot resembles first panel

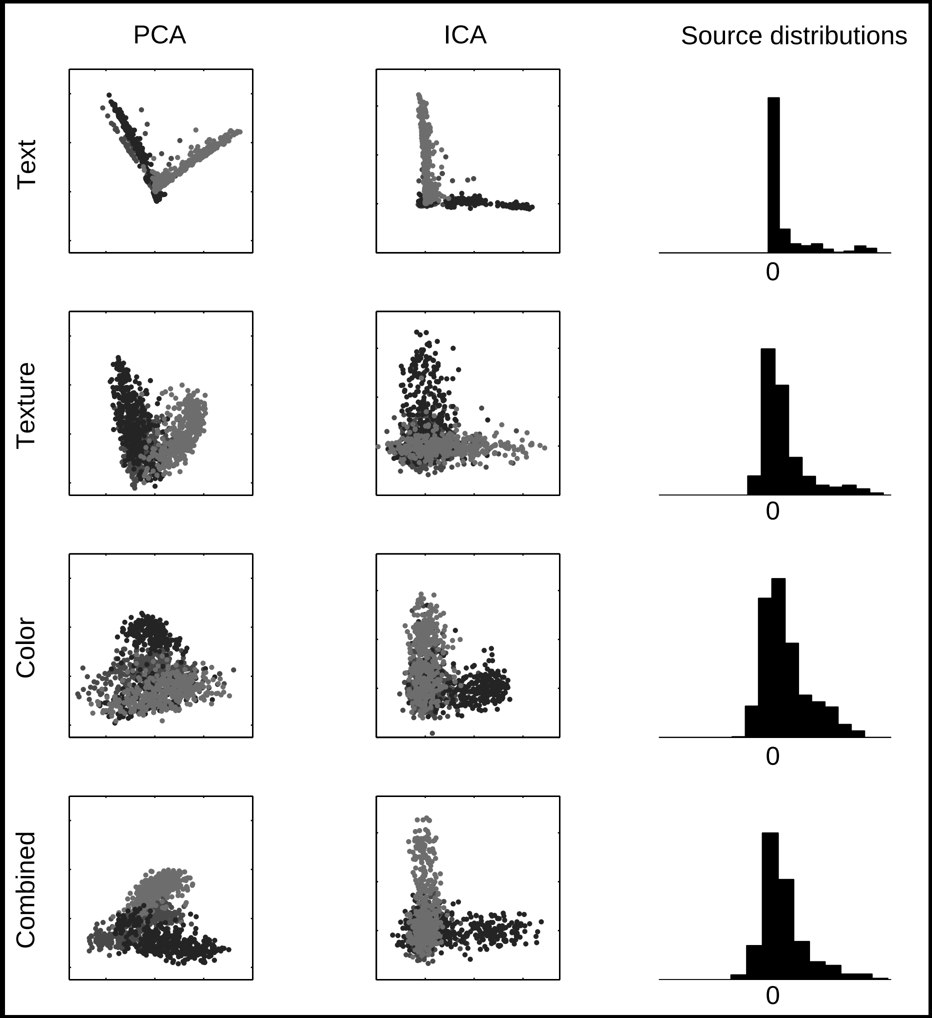
Panel 3: Gaussian with same mean and variance as data in first panel. Major axes parallel with PC axes - does not reveal the structure

Panel 4: 15 component Gaussian mixture model. Good density estimator but does not reveal independent components

Selection of components using BIC



ICA modeling



ICA classification

Interpretation as conditional component probabilities

$$\widehat{p}(k|\boldsymbol{x}) = \frac{\exp(\widehat{s}_k)}{\sum_{k=1}^{K} \exp(\widehat{s}_k)}, \qquad \widehat{\boldsymbol{s}} = [\widehat{s}_1, \cdots \widehat{s}_K]^\top = \widehat{\boldsymbol{\Phi}}^{-1} \widehat{\boldsymbol{U}}^\top \boldsymbol{x}$$

Posterior class probabilities

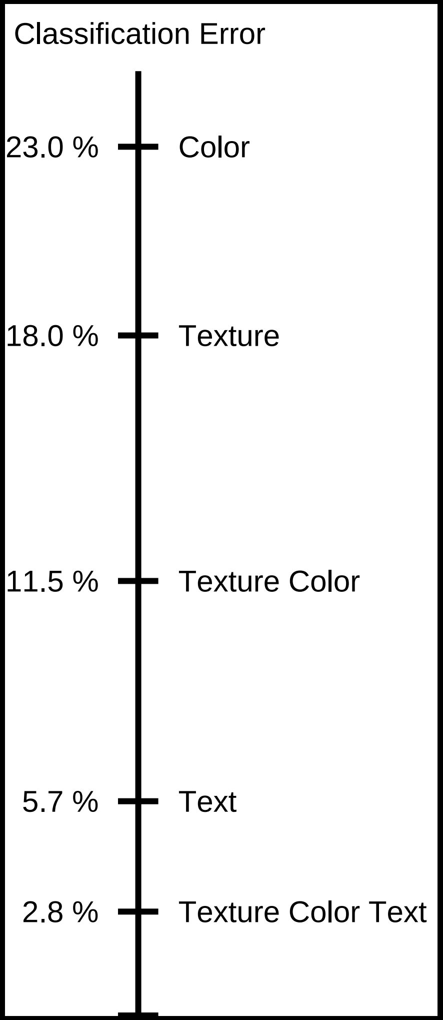
$$p(y|\boldsymbol{x}) = \sum_{k=1}^{K} p(y|k) p(k|\boldsymbol{x})$$

$p(y|k)$ are estimated from data as the frequency of occurrence for specific component-class combination $k \in [1;K]$, $y \in [1;C]$

$$\widehat{p}(y,k) = \frac{1}{N} \sum_{n=1}^{N} \delta(\, y - y(n)\,) \cdot \delta(\, k - \arg\max_{\ell} \widehat{p}(\ell|\boldsymbol{x}(n))\,)$$

$$\widehat{p}(y|k) = \frac{\widehat{p}(y,k)}{\sum_y \widehat{p}(y,k)}$$

Performance



Performance

Texture (K=13)

69.75	7.75	6.5
11.5	88.5	5.75
18.75	3.75	87.75

Color (K=16)

70.75	3.75	10
12	81.5	11.25
17.25	14.75	78.75

Text (K=45)

93	2	2.25
0.5	94.75	2.5
6.5	3.25	95.25

Combined errorrate: 2.8%

Single best errorrate: 5.7%

Texture Color (K=26)

82	1.75	4.5
9	93.75	5.75
9	4.5	89.75

Texture Color Text (K=26)

98.25	0.25	0.75
0.75	98	3.75
1	1.75	95.5

Component explanation

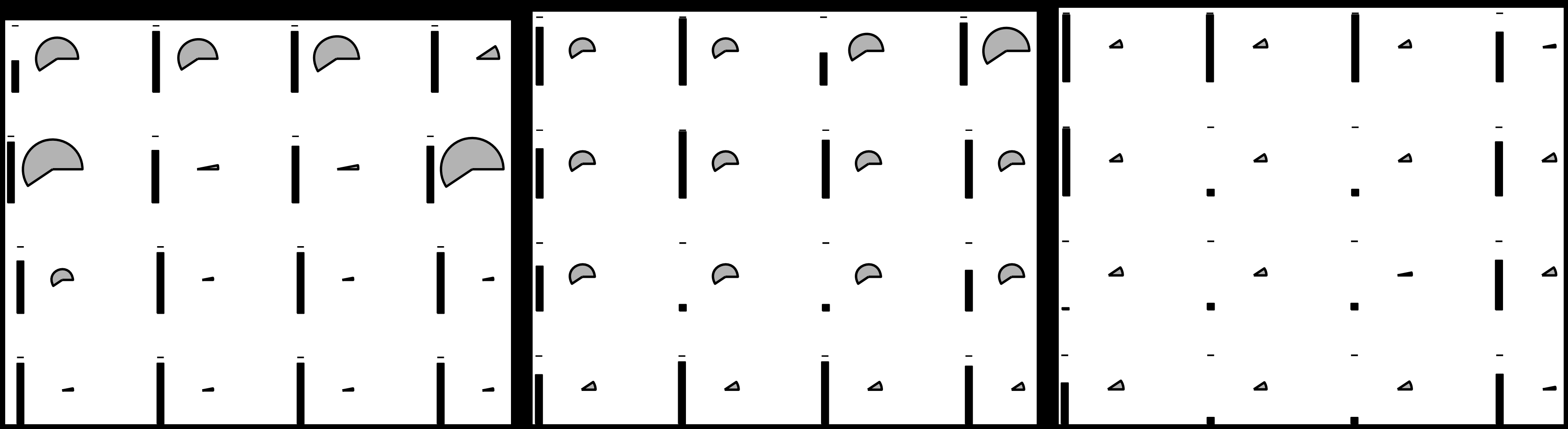
ICA model

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} = \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{s},$$

- $\boldsymbol{U}$ is the $P \times K$ matrix of K largest eigenvectors of the covariance of $\boldsymbol{x}$
- $\boldsymbol{\Phi}$ is the $K \times K$ mixing matrix. Quadratic ICA is thus performed in the subspace $\widetilde{\boldsymbol{x}} = \boldsymbol{U}^\top \boldsymbol{x}$

K'th column of $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Phi}}$ are features of K'th source. High values of the column provide a compact explanation

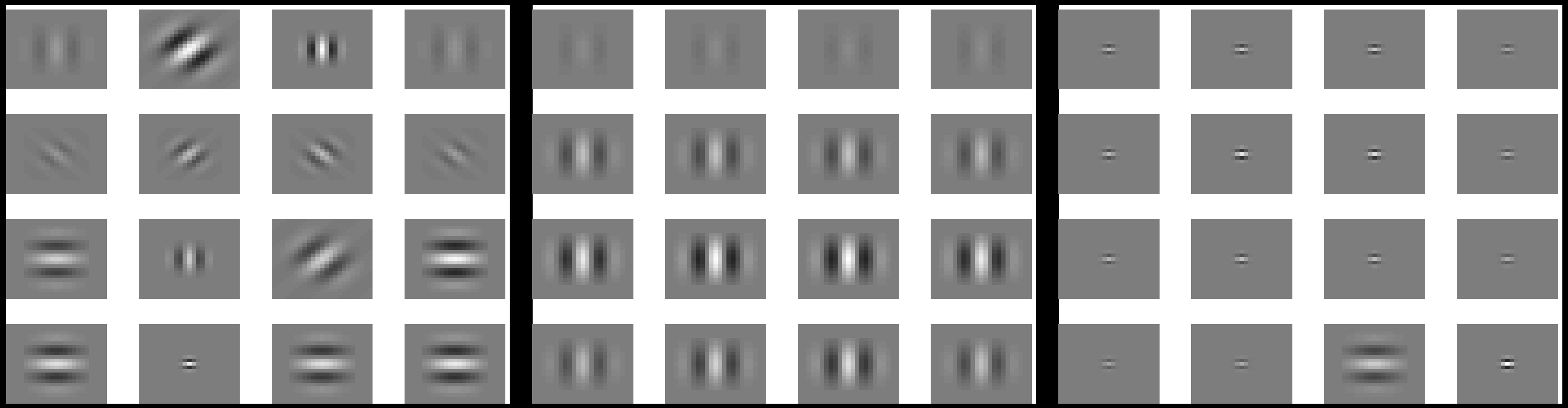
Interpreting color features



Sports Aviation Paintball

Bar is Value, Pie diameter is Saturation, Pie angle is Hue

Interpreting texture features



Sports Aviation Paintball

Image Annotation

Image	Label	Keywords
I_1	Sports	position college weight born lbs height guard
I_2	Aviation	na air convair wing
I_3	Paintball	check darkside force gog strike odt

Outline

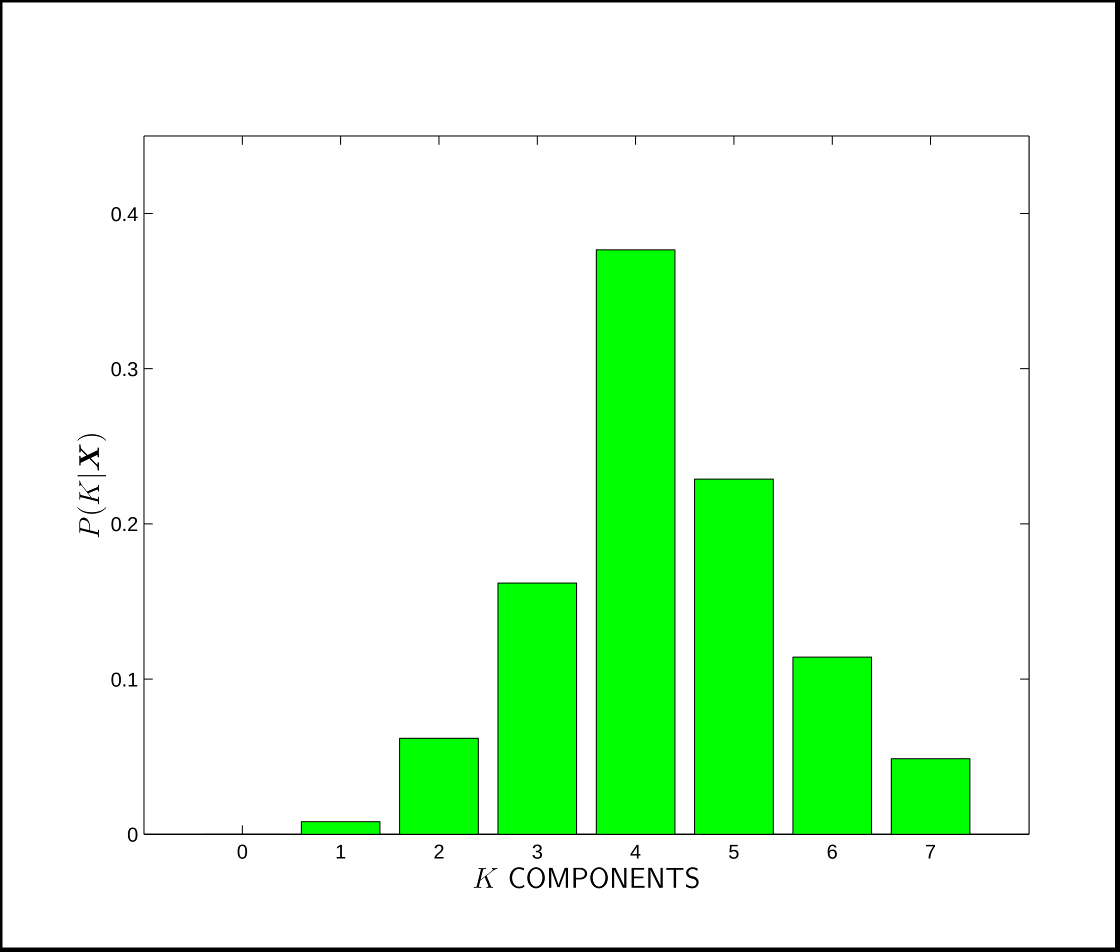
- ICA/BSS multimedia analysis
- multimedia applications of ICA/BSS
- learning algorithms
- example of content extraction from combined text and images
  — the ICA assumption for text and images
- example of chat room topic spotting
- conclusions and future challenges

Chat analysis

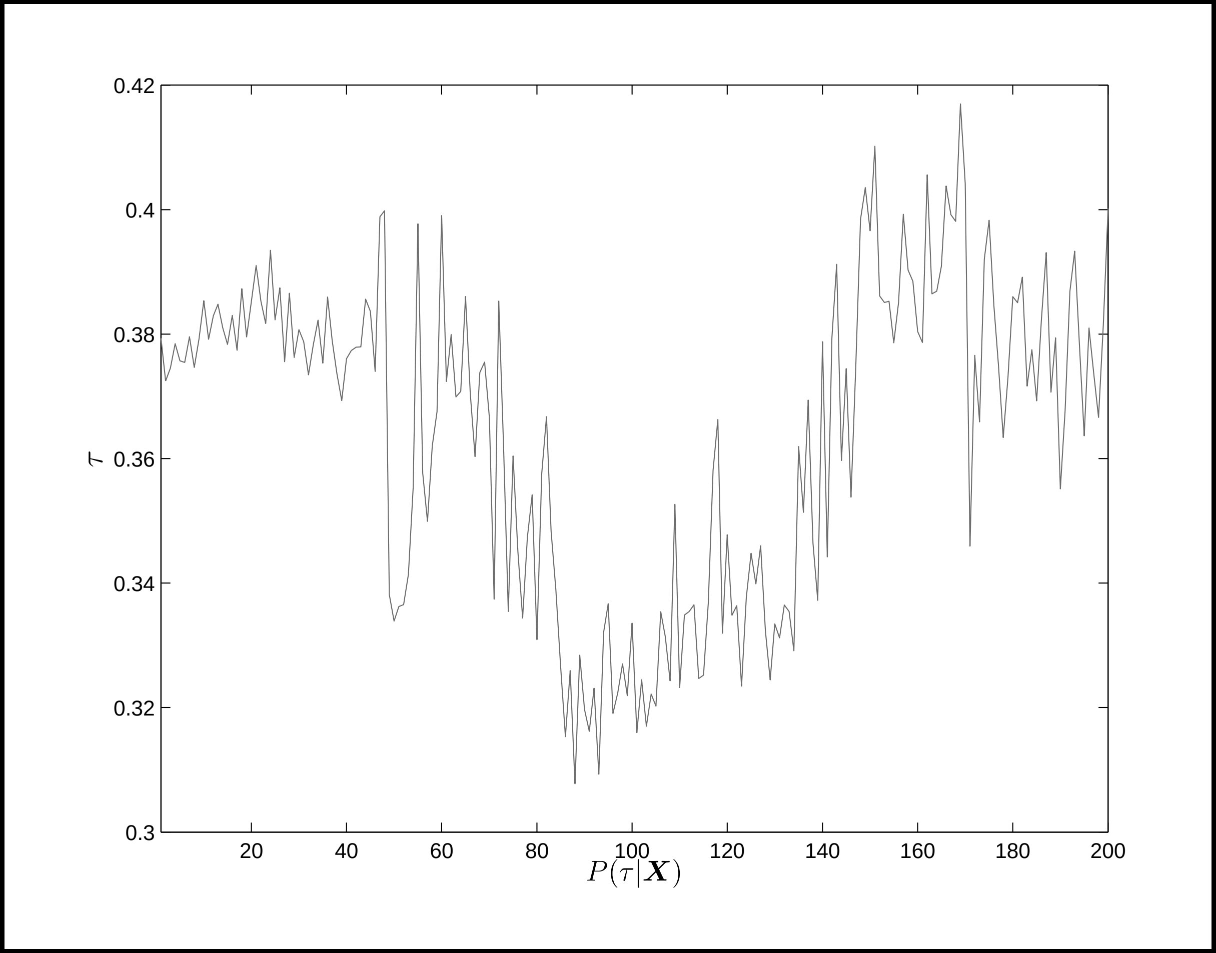
Kolenda, Hansen & Larsen, 2001

- retrospective analysis of CNN.com #CNN moderated chat channel
- topic spotting in unstructured chat text stream
- 8.5 hours on April 5th, 2000 was logged
- 4900 lines with 128 unique chatters participating
- preprocessing leaves $T = 2498$ terms
- the stream was segmented into 300 characters long pseudo-documents windows
- windows were shifted 150 characters without breaking words
- $N = 1114$ pseudo-documents was generated

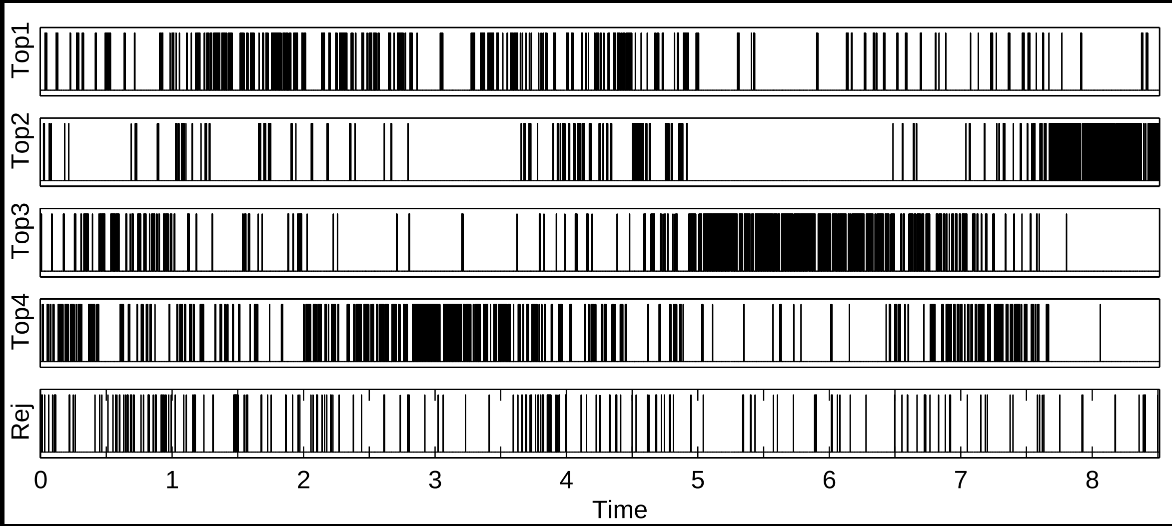
Detecting the number of topics



Selecting algorithm lag parameter



Retrospective event detection



	keywords
topic 1	chat join pm cnn board message allpolitics visit check america
topic 2	gun show
topic 3	susan smith mother children kid life
topic 4	people census elian state clinton government thing year good father time

Chat analysis demo

http://mole.imm.dtu.dk/webchat http://mole.imm.dtu.dk/webdemo/

Outline

- ICA/BSS multimedia analysis
- multimedia applications of ICA/BSS
- learning algorithms
- example of content extraction from combined text and images
  — the ICA assumption for text and images
- example of chat room topic spotting
- conclusions and future challenges

Conclusions

- independent component analysis and blind sources separation methods have potential for modeling, understanding and intelligent processing of multimedia data
- the unique feature of ICA is an unsupervised grouping of data which are amenable for interpretation and well-aligned with human perception
- combined text/image data for the purpose of cross-media retrieval and web search was presented
- synergy among text and image features leads to significant better classification performance
- retrospective dynamic topic spotting in chat rooms

Probing the future

- relevant and specific multimedia features for which linear ICA is the appropriate model. Nonlinear ICA?
- representation issues in image/video, e.g., facial animation, motion parameters and active appearance models
- incorporation of natural language and semantic features in text processing
  - R.E. Madsen, J. Larsen & L.K. Hansen: "Part-of-Speech Enhanced Context Recognition," MLSP2004
- Cognitive component analysis
  - Hansen, L. K., Ahrendt, P., Larsen, J.: "Towards Cognitive Component Analysis, AKRR'05.

Probing the future

- processing from mono/binaural audio signals: underdetermined convolutive mixture models, e.g., by invoking more specific audio priors

1. – Olsson, R. K., Hansen, L. K.: "A harmonic excitation state-space approach to blind separation of speech, Advances in Neural Information Processing Systems," NIPS 2005.
   – R.K. Olsson, L.K. Hansen: "A harmonic excitation state-space approach to blind separation of speech," NIPS 2004.
   – R.K. Olsson, L.K. Hansen: "Probabilistic blind deconvolution of non-stationary sources," 12th ESPC, 2004.
   – R.K. Olsson, L.K. Hansen: " Estimating the number of sources in a noisy convolutive mixture using BIC," ICA 2004.

   – non-stationary source AR(p)+sinusoidal
   – state space model and EM estimation

2. – Pedersen, M. S., Wang, D., Larsen, J., Kjems, U.: "Overcomplete Blind Source Separation by Combining ICA and Binary Time-Frequency Masking," MLSP'2005, 2005.
   – M.S. Pedersen, C.M. Nielsen: "Gradient Flow Convolutive Blind Source Separation," MLSP 2004.

   – gradient flow for four close microphones
   – extension to convolutive

3. – M.S. Pedersen, L.H. Hansen, U. Kjems, K.B. Ramussen: " Semi-blind Source Separation Using Head-Related Transfer Functions, ICASSP 2004.
   - acoustical propagation contraint - HRTF
   - two microphones
4. – Dyrholm, M., Makeig, S., Hansen, L. K.: "Model structure selection in convolutive mixtures, Neural Information Processing Systems," 2005.
   – M. Dyrholm, L.K. Hansen: "CICAAR: Convolutive Independent Component Analysis with an Auto-Regressive Inverse Model," IEEE Signal Processing Letters, 2003.
   - can handle more sensors than sources
   - IIR estimation of the sources
   - example: speaker in office with loud background music, two microphone, distance 60cm.
     mic1.wav, mic2.wav, sep-music.wav, sep-speech.wav
5. – T. Beierholm, B.D. Pedersen, O. Winther: "Low Complexity Bayesian Single Channel Source Separation," ICASSP'2004
   - single channel separation using source priors

Probing the future

- computational issue in training and recall in large multimedia databases
- estimation and optimization beyond natural gradient: advanced active data subset selection methods, online learning algorithms, and adaptation to changing environment
- intelligent fusion of media types, the ability to handle missing data and learning from combined labeled/unlabeled data

Acknowledgements

- Thomas Kolenda
- Lars Kai Hansen
- Finn Årup Nielsen
- Ole Winther