



# ICA and related models

Kaare Brandt Petersen  
Technical University of Denmark, ISP Group

Niels Bohr Institute  
February 2005



## Outline

- Introduction to Learning and Probabilistic Modelling
- Introduction to ICA
- Example: Mean Field ICA
- Example: A Network Estimation Model
- Summary



## Introduction: Learning from data

Typical **challenges** considered:

- Classification
- Regression
- Separation
- Monitoring
- Issues such as:
  - Bayesian or not
  - Generalization
  - Supervised / Unsupervised
  - Robustness
  - Outliers



## Introduction: Examples of Applications

**Projects** active at the moment (excl. ICA):

- Lipreading / picture coupling
- Monitoring of large diesel engines
- Music Classification
- Article database search
- The Demining Project
- Medical data mining
- fMRI Brainscannings



## Introduction: Applications of ICA

Examples of **applications** of ICA

- EEG measurements of the brain
- fMRI scannings
- CNN Chatroom
- Hearing Aids
- Telecommunication
- EKG measurements of the heart
- Speech separation – the cocktail party problem



## Introduction: Methods at ISP

### **Methods** of use and interest:

- Independent Component Analysis
- Hidden Markov Models (HMM)
- Graphical Models
- Feature Extraction
- Kalman Filters
- State Space Models
- Support Vector Machines
- Relevance Vector Machines
- Neural Networks?

### Important **concepts**

- Likelihood
- Prior
- Posterior
- Bayesian approach
  - priors on everything
  - integrate out the unknown



## Introduction: The ICA problem

- Tricky: Guess two numbers

$$1.7 = \frac{2}{3}s_1 + \frac{4}{3}s_2$$

(Statistical knowledge is useful)

- More tricky: Guess six numbers

$$\begin{bmatrix} 1.7 \\ 0.2 \end{bmatrix} = \mathbf{A}\mathbf{s}$$

- Less tricky: Guess many numbers using statistics

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t \quad t = 1, \dots, N$$



## Introduction: Some ICA approaches

### Approaches to ICA:

- Maximum of non-gaussianity
  - Kurtosis
  - Negentropy
- Information Theory
  - Minimum Mutual Information
  - Infomax
- Maximum Likelihood
  - Parameters in a generative model

### Model differences

- Square, over- or underdetermined
- Noise
- Convolutional
- Source prior possibilities



## Mean Field ICA: Probabilistic framework

### ■ Observation model

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t$$

**X** is the observed data

**A** is unknown

**S** is unknown, but  $p(\mathbf{S})$  is known

**n** is unknown, but  $p(\mathbf{n}) = \mathcal{N}(\mathbf{0}, \mathbf{W})$  is assumed gaussian

### ■ Maximum Likelihood

$$\frac{\partial \ln p(\mathbf{X}|\mathbf{A}, \mathbf{W})}{\partial \mathbf{A}} = 0$$

$$\frac{\partial \ln p(\mathbf{X}|\mathbf{A}, \mathbf{W})}{\partial \mathbf{W}} = 0$$

using Bayes we get

$$\mathbf{A} = \mathbf{X} \langle \mathbf{S} \rangle^T \langle \mathbf{S}\mathbf{S}^T \rangle^{-1}$$

$$\mathbf{W} = \frac{1}{N} \langle (\mathbf{X} - \mathbf{A}\mathbf{S})(\mathbf{X} - \mathbf{A}\mathbf{S})^T \rangle$$

Average wrt  $p(\mathbf{S}|\mathbf{X})$



## Mean Field ICA: The MF part of it

- **Bounding** the log likelihood

$$B(\Theta, \Phi) = \int q(\mathbf{S}|\Phi) \ln \frac{p(\mathbf{X}, \mathbf{S}|\Theta)}{q(\mathbf{S}|\Phi)} = \ln p(\mathbf{X}|\Theta) - KL[q(\mathbf{S}|\Phi)||p(\mathbf{S}|\mathbf{X}, \Theta)] \leq \ln p(\mathbf{X}|\Theta)$$

- Approximating the posterior mean using **Mean Field** Theory

$$\begin{aligned} \langle \mathbf{S} \rangle &= \int \mathbf{S} p(\mathbf{S}|\mathbf{X}, \Theta) d\mathbf{S} \cong \int \mathbf{S} q(\mathbf{S}|\Phi) d\mathbf{S} = f_1(\Phi, \langle \mathbf{S} \rangle, \langle \mathbf{S}\mathbf{S}^T \rangle) \\ \langle \mathbf{S}\mathbf{S}^T \rangle &= \int \mathbf{S}\mathbf{S}^T p(\mathbf{S}|\mathbf{X}, \Theta) d\mathbf{S} \cong \int \mathbf{S}\mathbf{S}^T q(\mathbf{S}|\Phi) d\mathbf{S} = f_2(\Phi, \langle \mathbf{S} \rangle, \langle \mathbf{S}\mathbf{S}^T \rangle) \end{aligned}$$

where  $q(\mathbf{S})$  is a suitably factorized distribution

- Variants: Linear Response, TAP, Bethe, Kikuchi, ...



## Mean Field ICA: Optimization

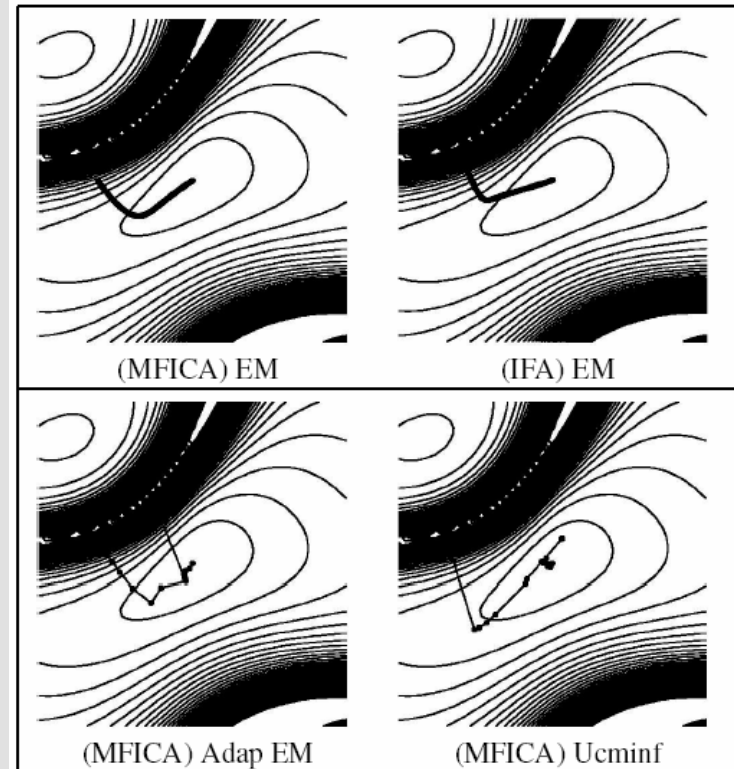
- Different optimization schemes
  - an example:

**EM:** 729 iterations

**Adap EM:** 16 iterations

**UCMINF:** 25 iterations

- Some challenges of Maximum likelihood:
  - Choosing model
  - The optimization
 (Don't use the EM algorithm!)





## Mean Field ICA: Discussion

### ■ **Pros**

- Can estimate both  $A, W$  and prior parameters
- Flexible wrt constraints on  $A$  and  $W$
- Flexible wrt source priors

### ■ **Cons**

- Somewhat complicated to implement
- not scaling well in number of sources

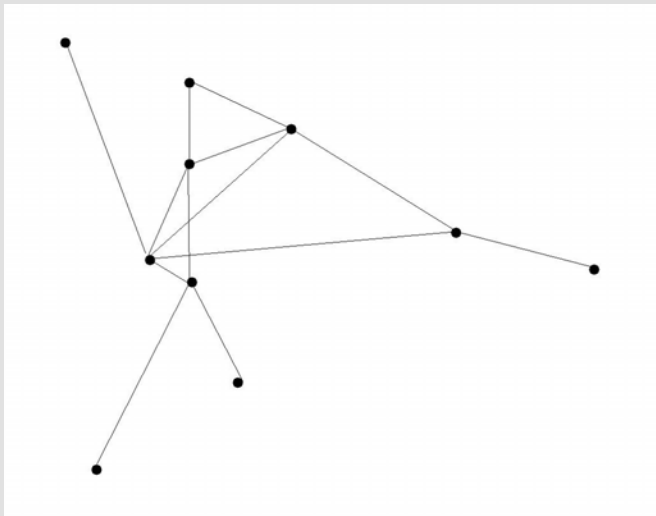
### ■ **Outlook**

- ICA is moving towards convolutive and non-linear mixtures
- Non-linear probably ok, what about convolutive?



## Network Model: Matrix representation

- A sparse network



Represented by  
a sparse matrix,  $M$

- **Matrices** can represent networks.  $i$  and  $j$  are connected using  $M_{ij}$ .
- **Internet** analogy:  
Vertices = Computers  
Edges = Wiring
- **Question:** If we can control and measure some part of the signals send, can we then infer knowledge about  $M$ ?



## Network Model: Idea 1

- **Observation** model

Known:  $x, M$  and  $v$

- Reformulating into (underdetermined) **ICA**

$$x = Mv + A_2 u$$

Estimate  $A_2$  and  $u$

- **Challenge:** Size and underdetermined-ness  
 $A_2$  is app.  $20 \times 1e12$

$$\begin{bmatrix} x \\ \vdots \\ z \\ \vdots \end{bmatrix} = \begin{bmatrix} M & \dots & A_2 & \dots \\ \vdots & & & \\ A_3 & & & \\ \vdots & & & \end{bmatrix} \begin{bmatrix} v \\ \vdots \\ u \\ \vdots \end{bmatrix}$$



## Network Model: Idea 2

- **Representation** of the internet as binary (symmetric) matrices

- **Number of paths** as potents

$$M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad M^6 = \begin{bmatrix} 5 & 7 & 9 \\ 9 & 12 & 16 \\ 7 & 9 & 12 \end{bmatrix}$$

- **Challenges:** size and non-linearity. Countermove: utilize sparseness

- Part of a **larger network**

$$\begin{aligned} \mathbf{J} &= \left[ \begin{array}{c|c} \mathbf{M} & \mathbf{A}_2 \\ \hline \mathbf{A}_3 & \mathbf{A}_4 \end{array} \right] \\ \mathbf{J}^2 &= \left[ \begin{array}{c|c} \mathbf{M}\mathbf{M} + \mathbf{A}_2\mathbf{A}_3 & \mathbf{M}\mathbf{A}_2 + \mathbf{A}_2\mathbf{A}_4 \\ \hline \mathbf{A}_3\mathbf{M} + \mathbf{A}_4\mathbf{A}_3 & \mathbf{A}_3\mathbf{A}_2 + \mathbf{A}_4\mathbf{A}_4 \end{array} \right] \\ \mathbf{J}^3 &= \dots \end{aligned}$$

- Observations from packages

$$(\hat{\mathbf{J}}^p)_{obs} = \sum_{n=0}^p f_n(\mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4, \mathbf{M})$$



## Summary

- ISP Group: Learning from data
  - Information theory
  - Probabilistic Models
  
- Independent Component Analysis, ICA
  - Separation of linear mixtures
  - Assuming the sources are independent
  
- Brainstorm on network models
  - Idea 1: Hugely underdetermined ICA problem
  - Idea 2: Sample the path/potens matrix



## Acknowledgements and References

- Details on ISP Group: <http://isp.imm.dtu.dk>
- A **tutorial on ICA**: A. Hyvarinen and E. Oja "Independent Component Analysis", Neural Networks, 13 (4-5), 411-430, 2000.
- Details on **Mean Field ICA**: P. Højen-Sørensen, L. K. Hansen and O. Winther "Mean Field Approaches to Independent Component Analysis", Neural Computation, 14, 889-918, 2002.