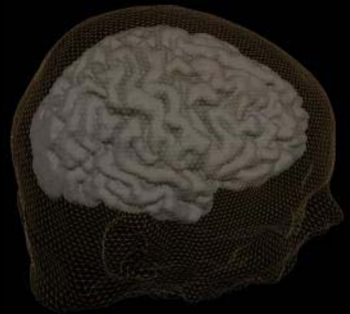
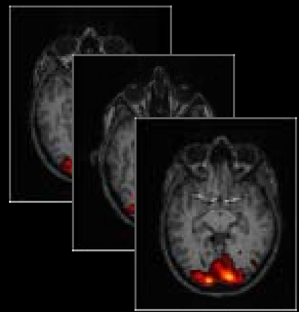
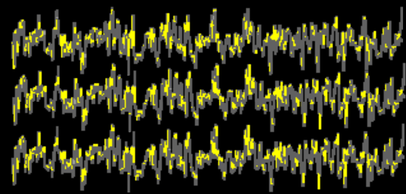
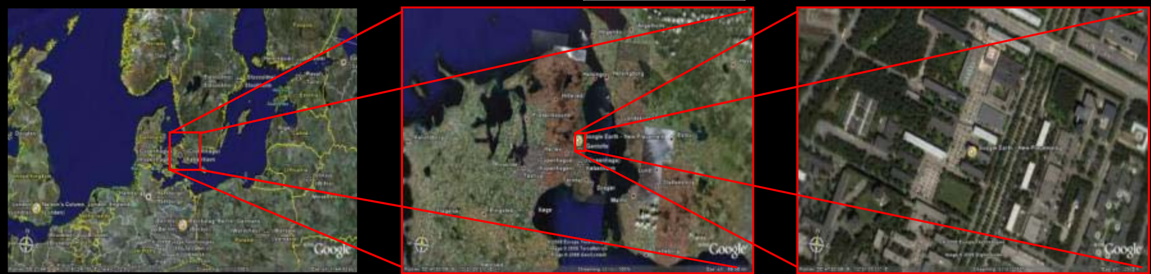
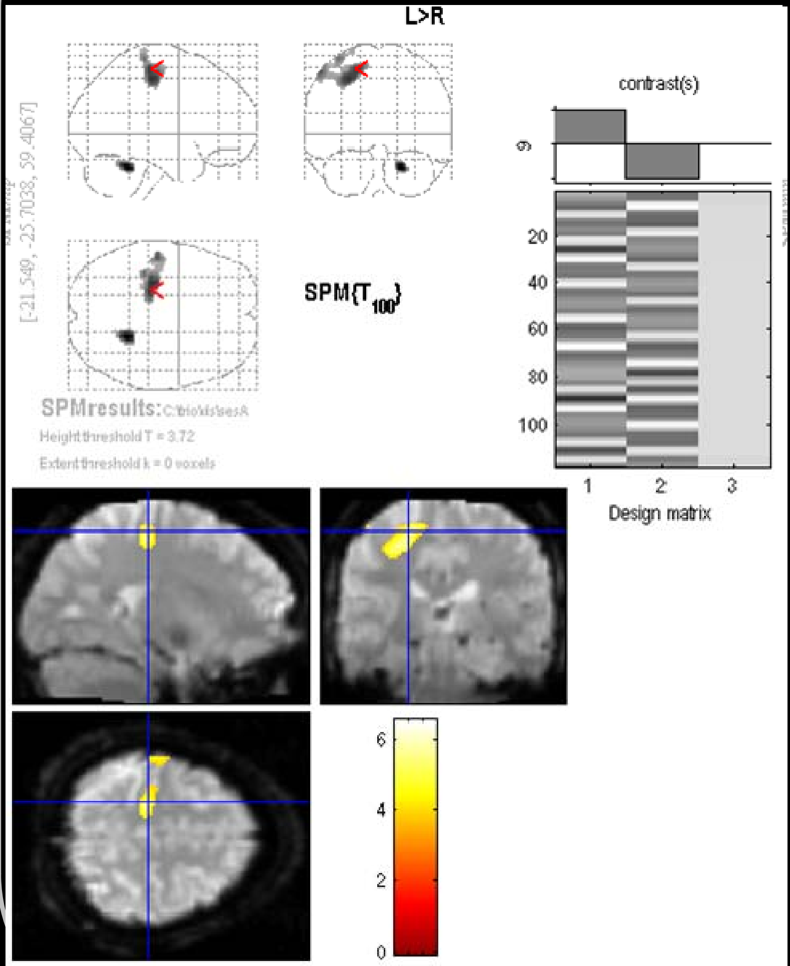
Tensor decomposition for mining the consistent reproducible patterns in neuroimaging data


Morten Mørup
Informatics and Mathematical Modeling
Cognitive Systems
Technical University of Denmark
joint work with Kristoffer Hougaard Madsen, Sidse Marie Arnfred and Lars Kai Hansen

Univariate statistical analysis in NeuroImaging



Problems:

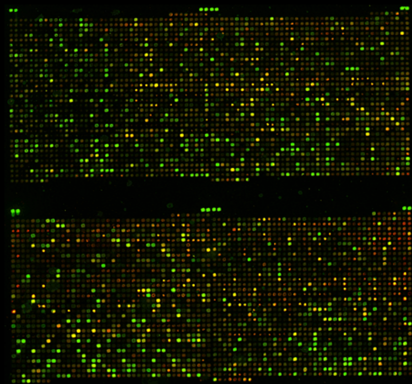
1) Multiple comparisons, i.e. many voxels tested.
2) What is the true number of independent tests, i.e. voxels are highly correlated
3) Data extremely noisy, i.e. low SNR rendering tests insignificant.

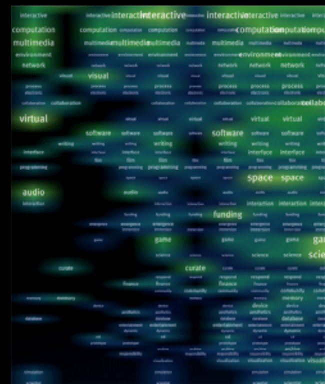
⇓

Need for advanced multivariate methods that can efficiently extract the underlying sources in the data

This problem is no different than the problems encountered in general in Modern Massive Datasets (MMDS)

$\mathbf{X}^{Space \times Time}$ $\mathbf{X}^{Gene\ \ seq. \times Samples}$ $\mathbf{X}^{Webpages \times Webpages}$ $\mathbf{X}^{Term \times Document}$





NeuroInformatics **BioInformatics** **ComplexNetworks** **WebDataMining**

Unsupervised Learning attempts to find the hidden causes and underlying structure in the data.

(Multivariate exploratory analysis – driving hypotheses)

Goal of unsupervised Learning

(Ghahramani & Roweis, 1999)

- Perform dimensionality reduction
- Build topographic maps
- Find the hidden causes or sources of the data
- Model the data density
- Cluster data

Purpose of unsupervised learning

(Hinton and Sejnowski, 1999)

- Extract an efficient internal representation of the statistical structure implicit in the inputs

2008

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the

"**All models are wrong,** but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't

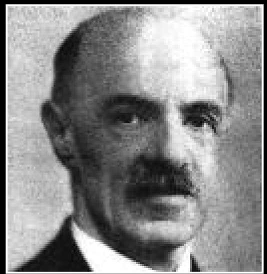
Analysis of massive amounts of data will be the main driving force of all sciences in the future!!

Outline of the talk

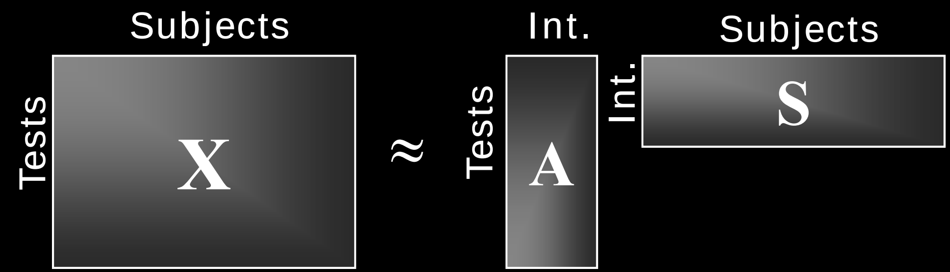
- NeuroImaging data modeled as tensors (CandeComp/PARAFAC(CP), ShiftCP and ConvCP)
- Bayesian methods for estimating the number of components in tensor decomposition (Automatic Relevance Determination)
- Tensor decomposition of complex functional networks (Infinite Relational Modeling)

NeuroImaging data modeled as tensors

Factor Analysis



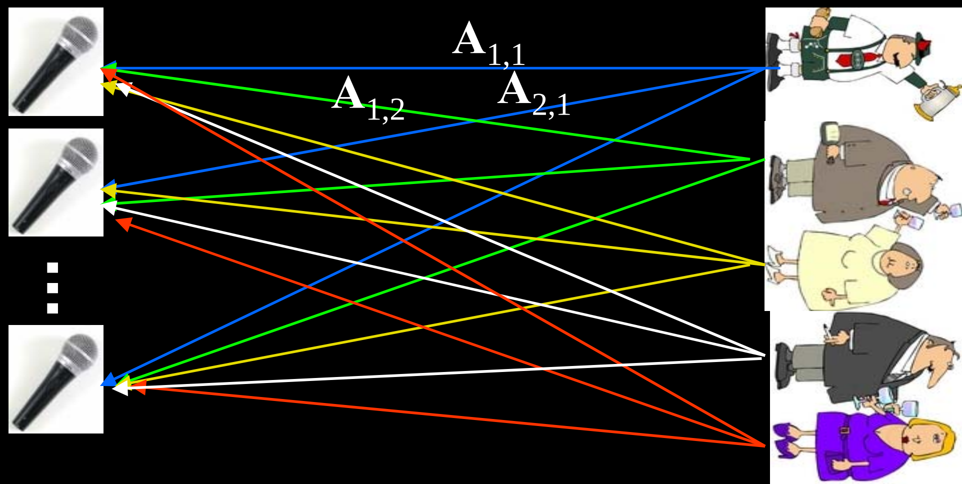
Spearman ~1900



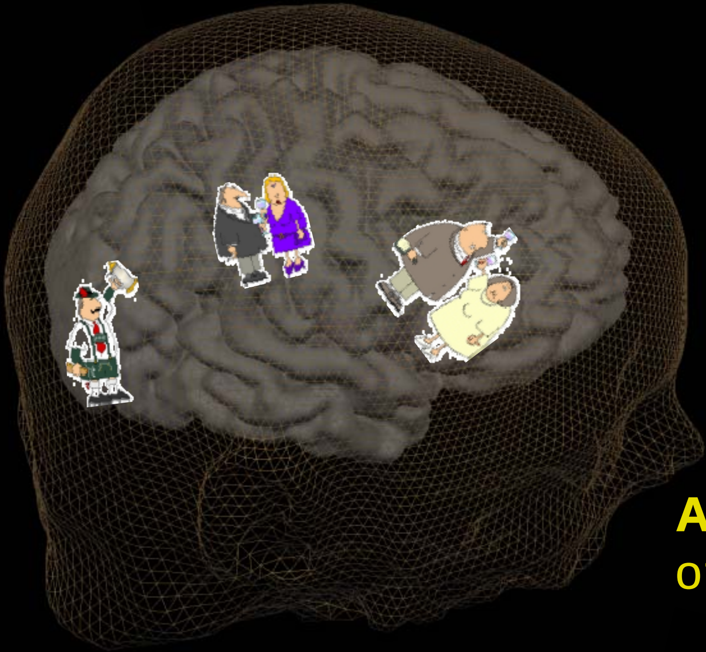
$\mathbf{X}^{\text{tests x subjects}} \approx \mathbf{A}^{\text{tests x int.}}\mathbf{S}^{\text{int. x subjects}}$

The Cocktail Party problem (Blind source separation)

$\mathbf{X}^{\text{microphones x time}} \approx \mathbf{A}^{\text{microphones x people}}\mathbf{S}^{\text{people x time}}$

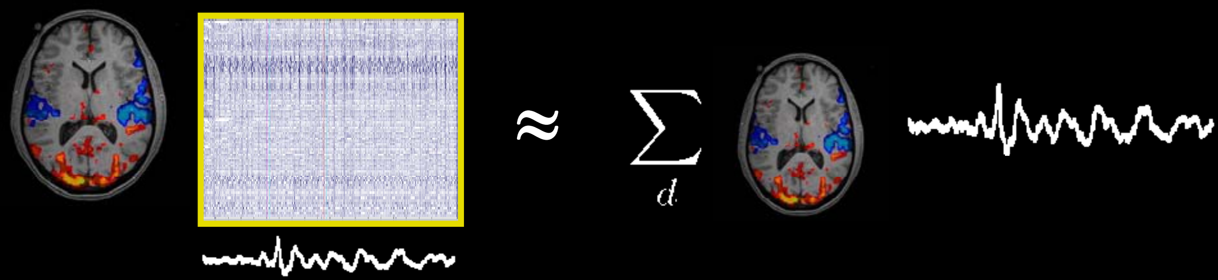


The EEG/MEG/fMRI Party problem



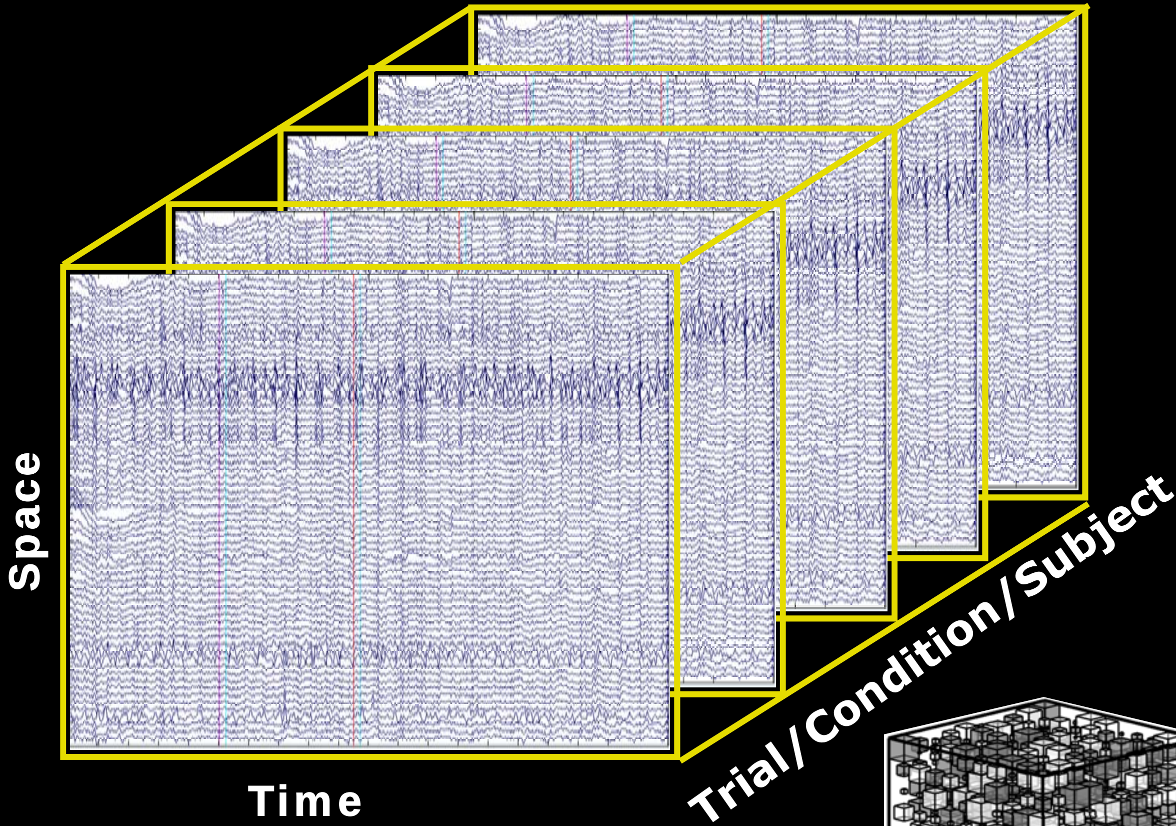
$$\mathbf{X}^{\mathrm{Voxel}\times\mathrm{Time}} \approx \sum_d \mathbf{a}_d^{Voxel}\mathbf{b}_d^{\mathrm{Time}}$$

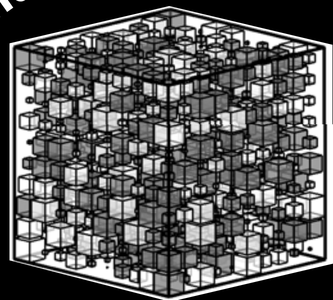
Assumption: Data **instantaneous** mixture of temporal signatures. (PCA/ICA/NMF)



Flaw: $\boxed{X \approx AS = (AQ^{-1})(QS) = \hat{A}\hat{S}}$ $\Longrightarrow$ **Representation not unique!**

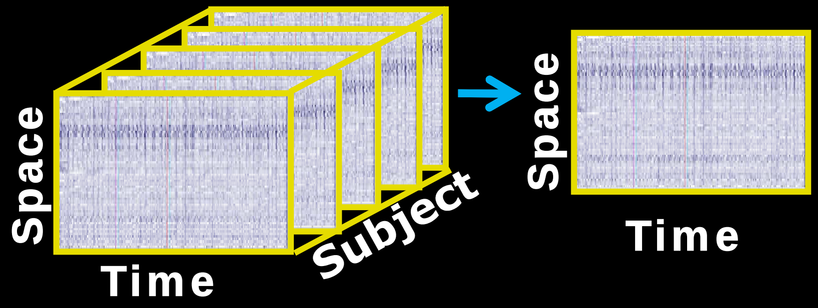
From 2-way to multi-way analysis



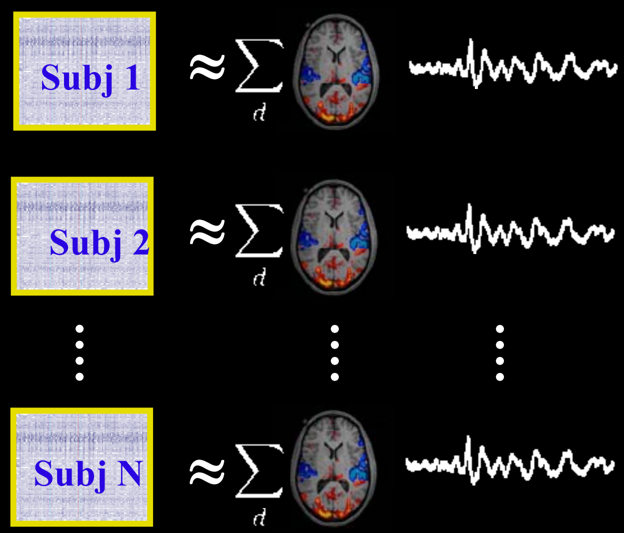


3 common ways of avoiding tensors

Preaverage

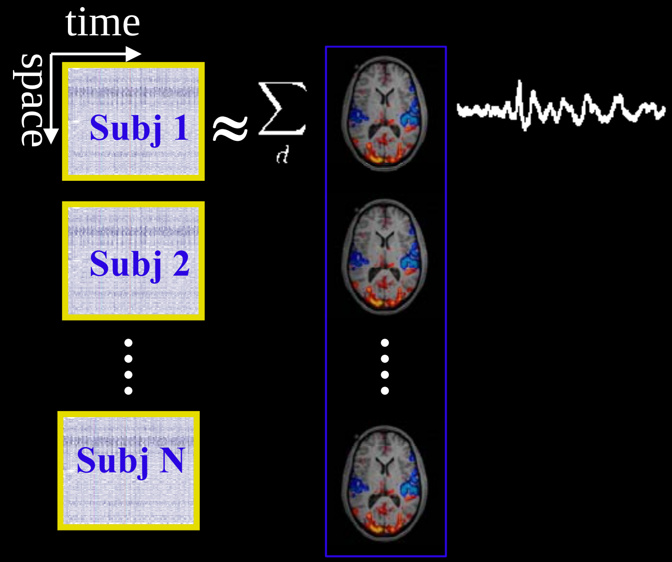


Separate Analysis

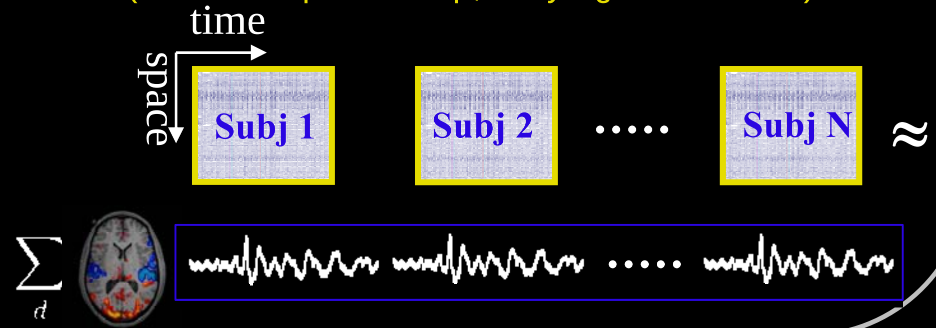


Concatenation

(identical time series varying spatial maps)

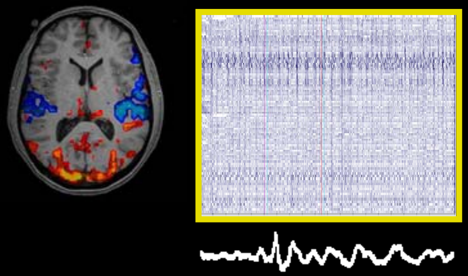


(identical spatial map, varying time series)



Multilinear modelling

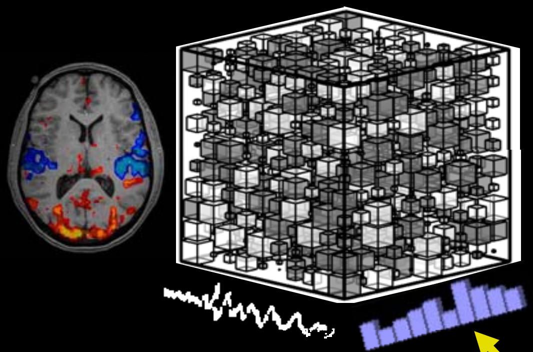
Bilinear Model: $\mathbf{X}^{\mathrm{Voxel}\times\mathrm{Time}} \approx \sum_d \mathbf{a}_d^{Voxel}\mathbf{b}_d^{\mathrm{Time}}$



$\sum_d$

Assumption: Data **instantaneous** mixture of temporal signatures.
(PCA/ICA/NMF)

Trilinear Model: $\mathbf{X}^{\mathrm{Voxel}\times\mathrm{Time}\times Trial} \approx \sum_d \mathbf{a}_d^{Voxel}\mathbf{b}_d^{\mathrm{Time}}\mathbf{c}_d^{Trial}$



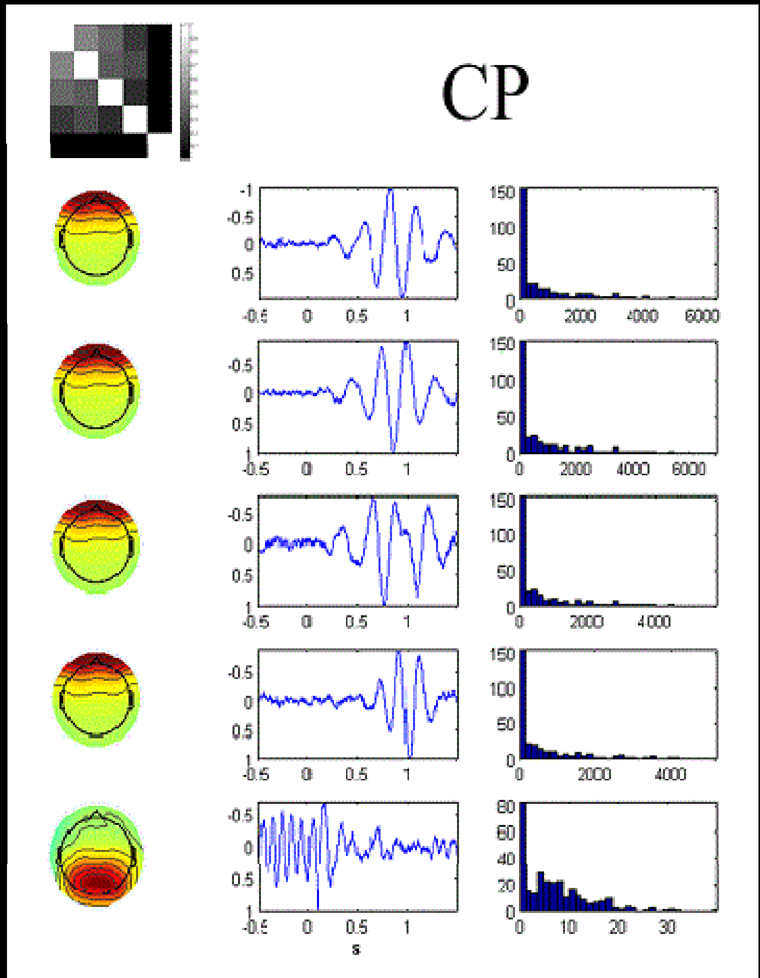
$\sum_d$

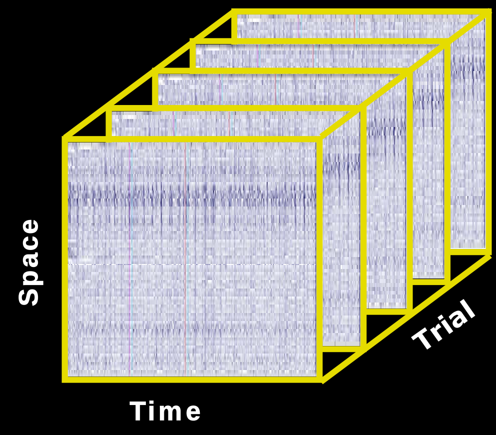
Assumption: Data **instantaneous** mixture of temporal signatures that are expressed to **various degree** over the Subjects/trials
(Canonical Decomposition, Parallel Factor (CP))

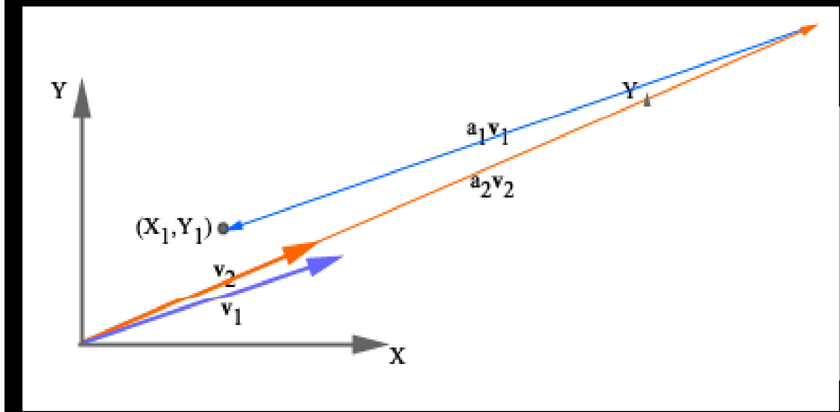
(weighted averages over the trials)

Mult. Mod. admits non-ambiguous extraction of concistent patterns of activation
(see also k-rank criterion by Kruskal 1976,1977)

Unfortunately, violation of multi-linearity causes degeneracy

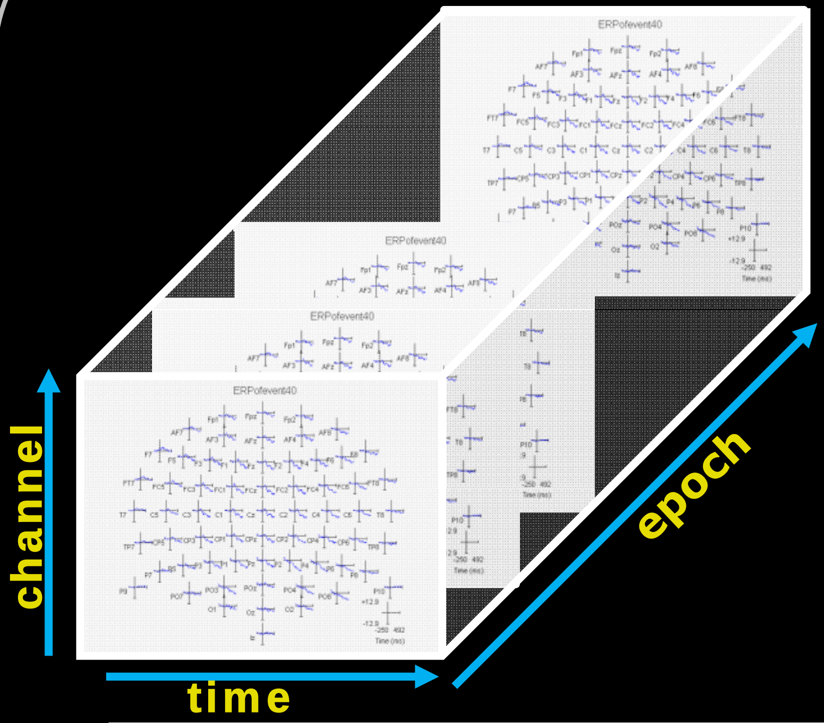


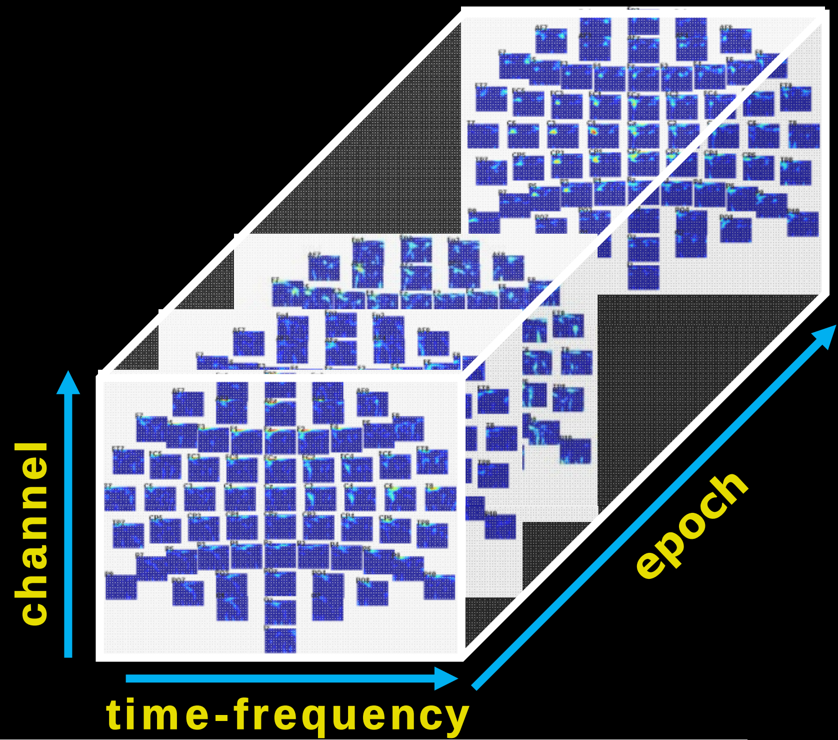


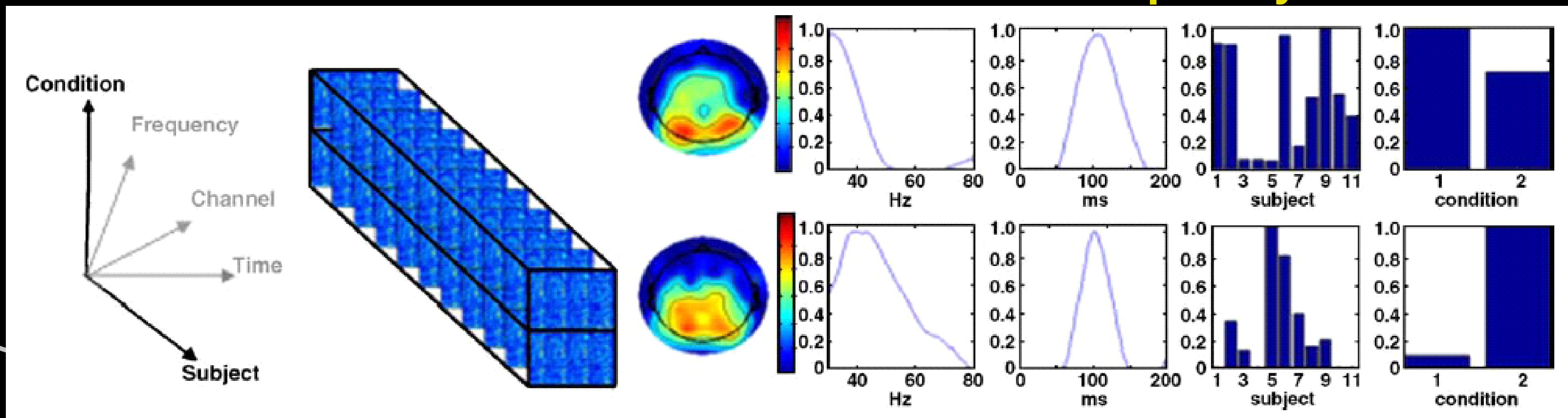


Common fixes: Impose orthogonality, regularization or non-negativity constraints by analyzing data transformed to a time-frequency domain representation

Wavelet transformed data





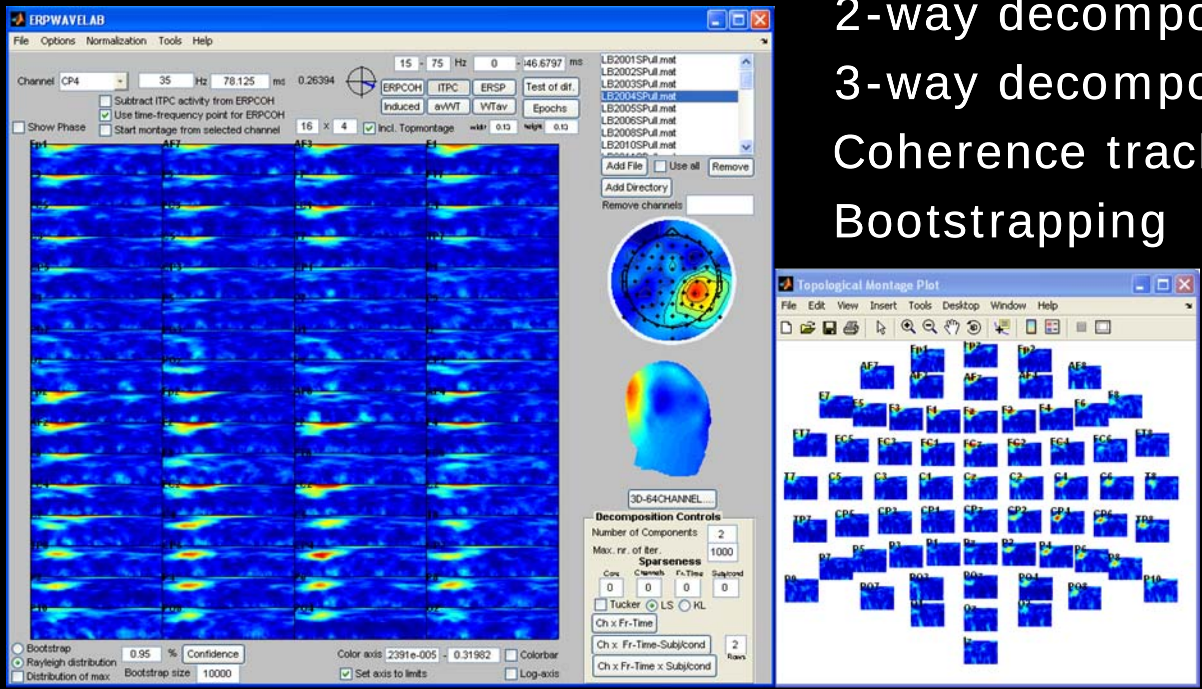


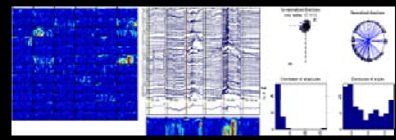
(Mørup et al., NeuroImage 2006)

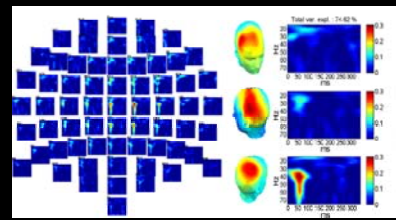
www.ERPWAVELAB.org

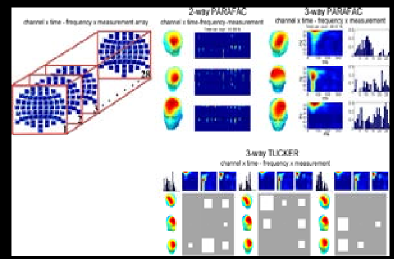
Features:

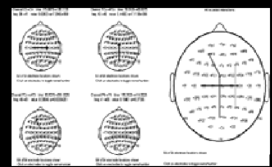
- Wavelet analysis
- Data visualization
- Artifact Rejection
- 2-way decomposition
- 3-way decomposition
- Coherence tracking
- Bootstrapping

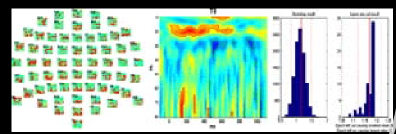












(Mørup et al, Journ. of Neurosc. Meth. 2007)
(Algorithms described in Mørup et al, Neural Computation 2008)

Degeneracy often a result of multi-linear models being too restrictive

Trilinear model can encompass:

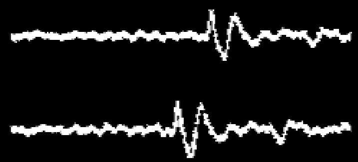
- Variability in strength over repeats

However, other common causes of variation are:

- Delay Variability

Trial 1

Trial 2

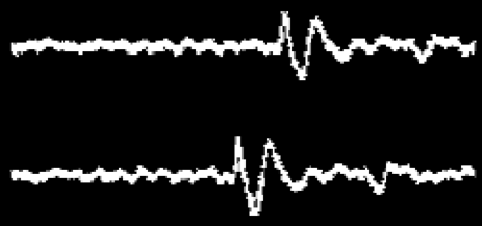


- Shape Variability

Trial 1

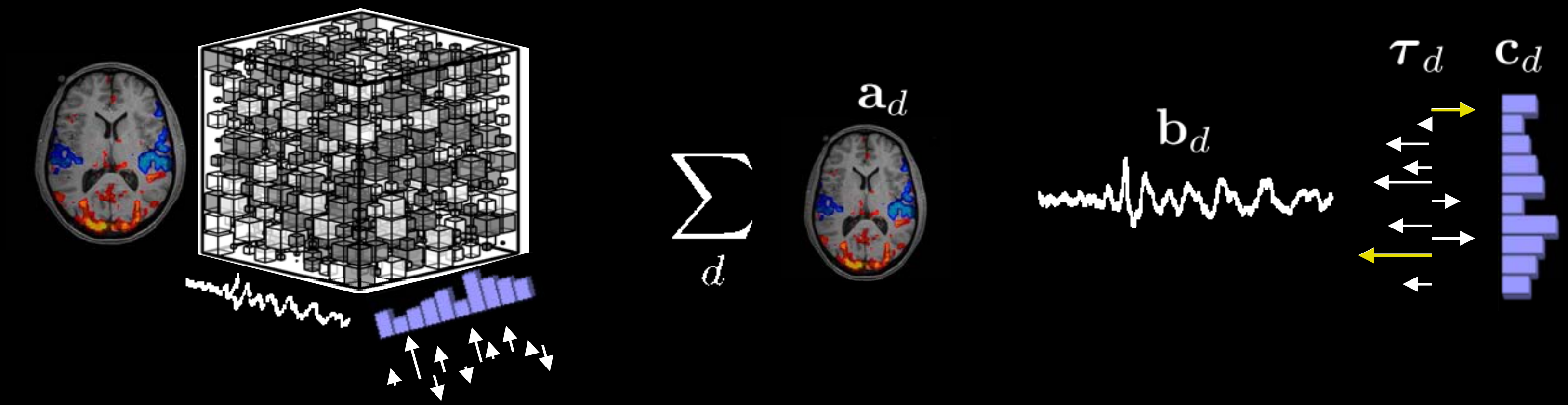
Trial 2

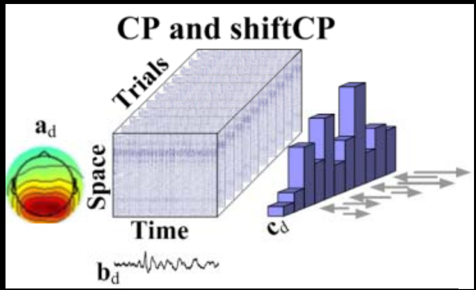
Modelling Delay Variability



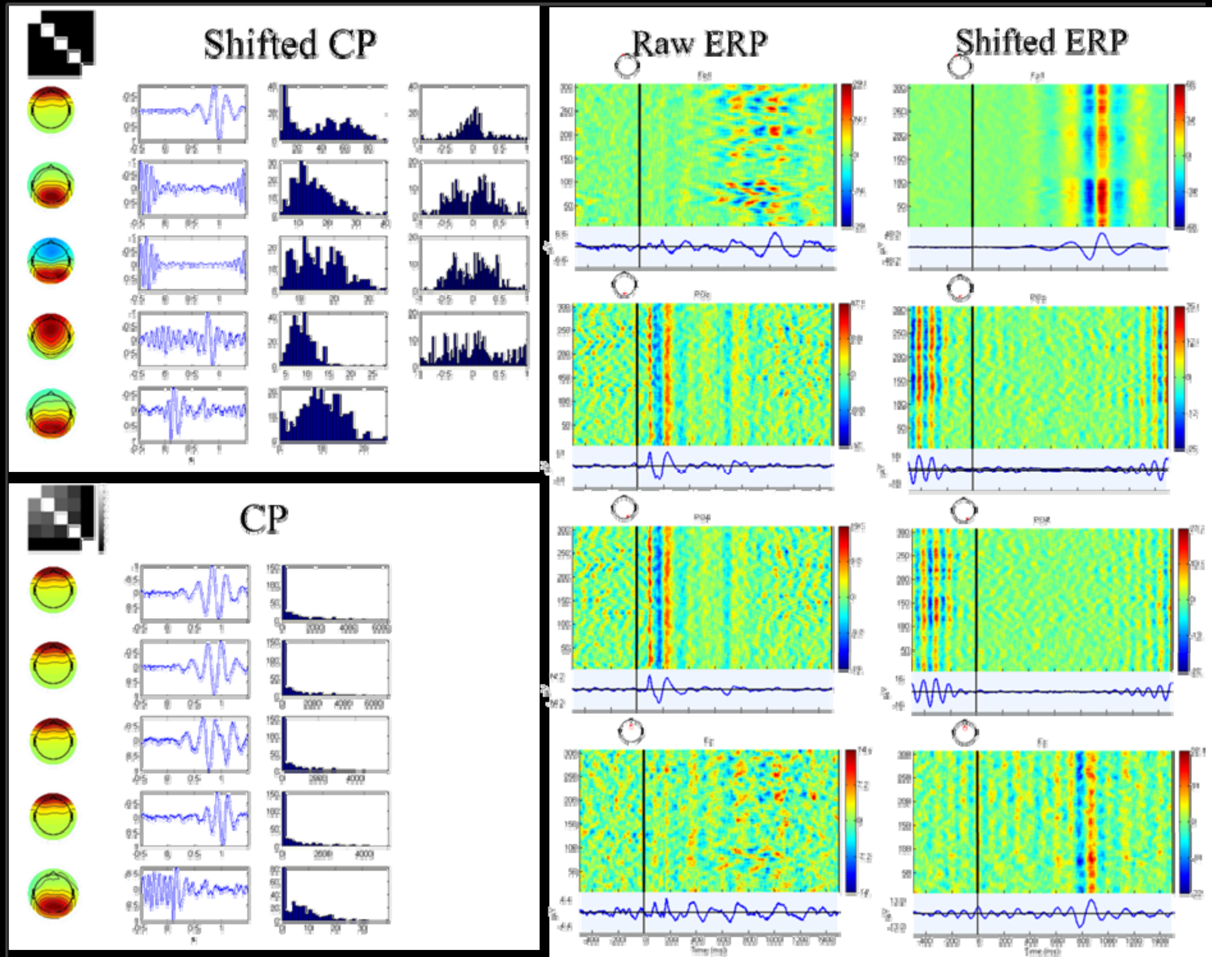
Shifted CP:

$$x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t - \boldsymbol{\tau}_{k,d}) c_{k,d}$$





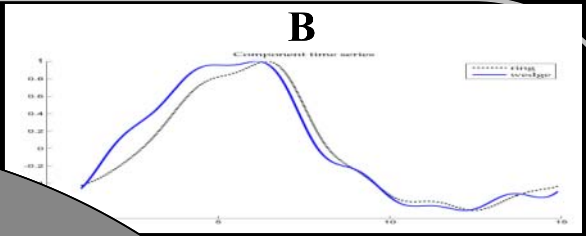
$$x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t - \tau_{k,d}) c_{k,d}$$

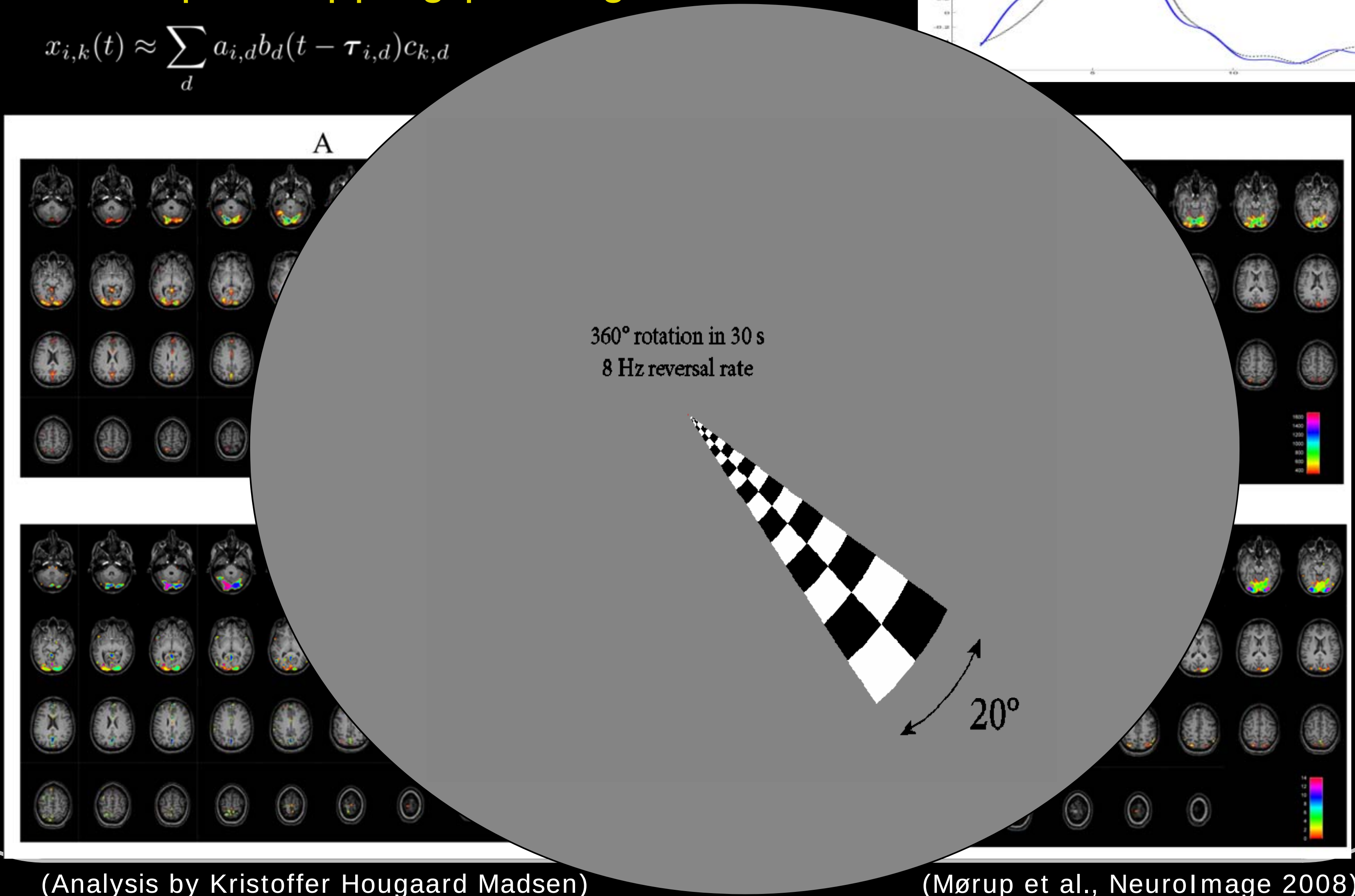


(Mørup et al., NeuroImage 2008)

Delay modelling of fMRI data from retinotopic mapping paradigm

$$x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t - \tau_{i,d}) c_{k,d}$$

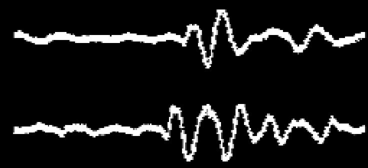




(Analysis by Kristoffer Hougaard Madsen)

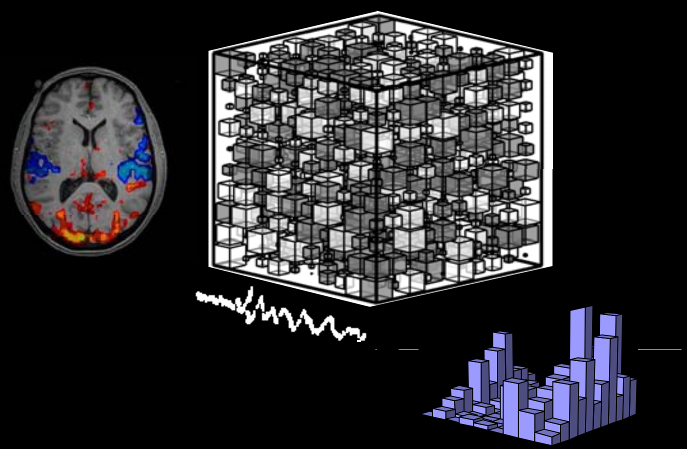
(Mørup et al., NeuroImage 2008)

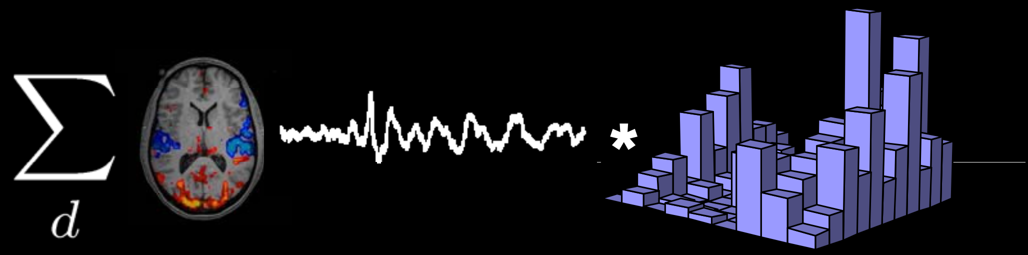
Modeling Shape (and delay) Variability



convolutive CP:

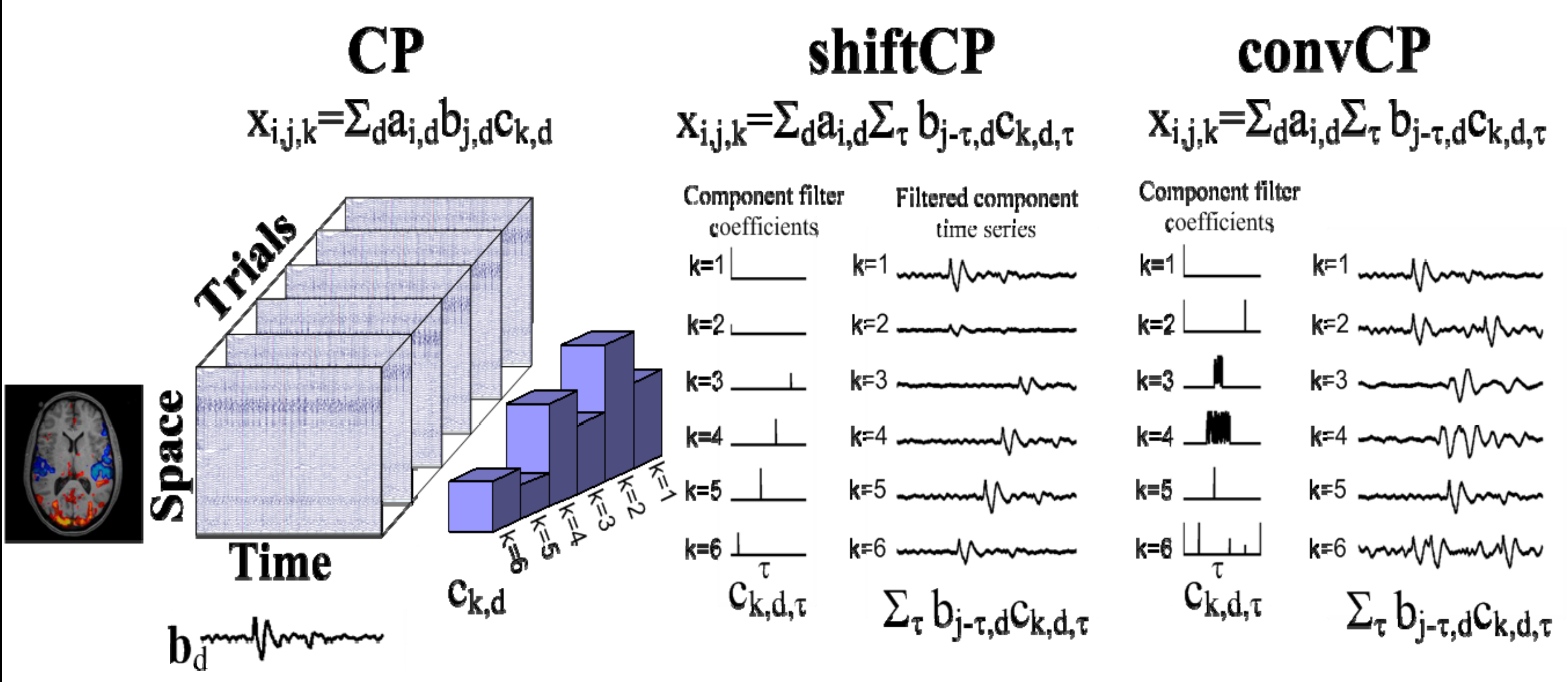
$$x_{i,k}(t) \approx \sum_{d,\tau} a_{i,d} b_d(t-\tau) c_{k,d}(\tau)$$





(Mørup et al., Nips workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis 2008)

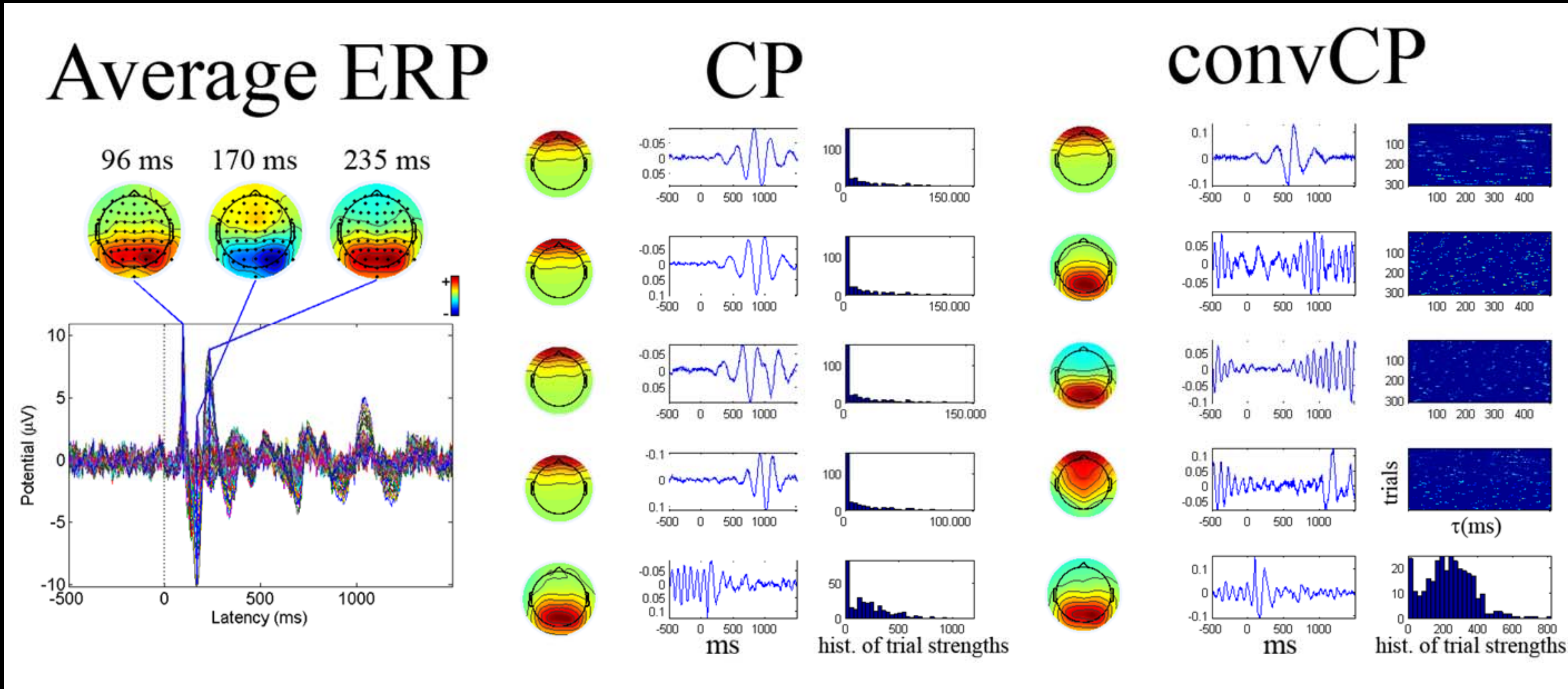
CP, ShiftCP and ConvCP



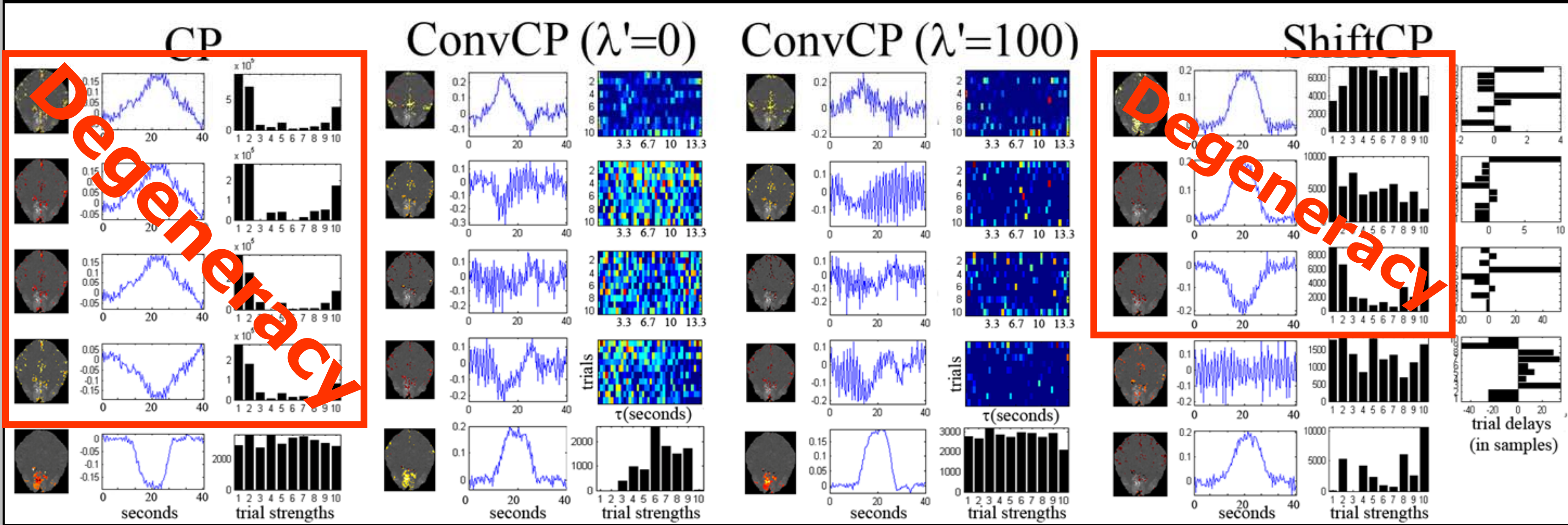
ConvCP: Can model arbitrary number of component delays within the trials and account for shape variation within the convolutional model representation. Redundancy between what is coded in C and B resolved by imposing sparsity on C.

(Mørup et al., Nips workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis 2008)

Convolutive Multi-linear decomposition



Analysis of fMRI data



Each trial consists of a visual stimulus delivered as an annular full-field checkerboard reversing at 8 Hz.

λ' is L_1 sparsity regularization imposed on third mode

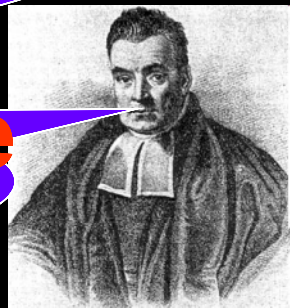
(Mørup et al., Nips workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis 2008)

Bayesian Learning and the Principle of Parsimony

William of Ockham

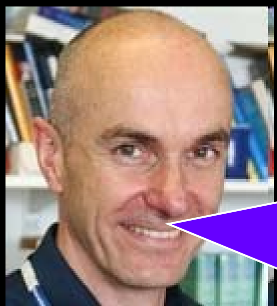
The explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory.

Open problem: What is an adeqaute degree of sparsity and the "correct" number of components?



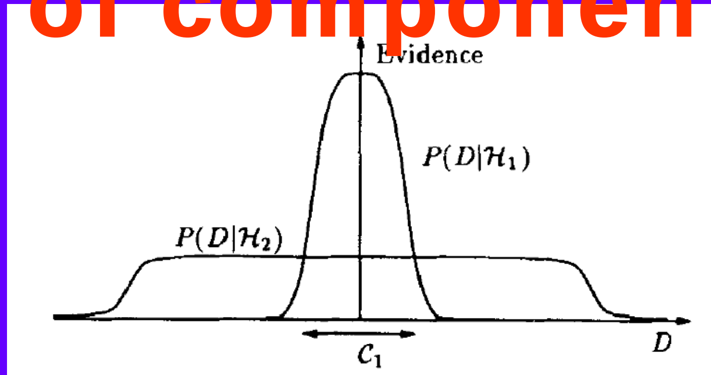
Thomas Bayes

To get the posterior probability distribution, multiply the prior probability distribution by the likelihood function and then normalize.



David J.C. MacKay

Bayesian learning embodies Occam's razor, i.e. Complex models are penalized.



Many inference paradigms in Bayesian Learning

- **Maximum a posteriori estimation (MAP)**
  seeks optimal solution (admit standard optimization) however, the approach does not take parameter uncertainty into account
- **Sampling methods**
  Marcov Chain Monte Carlo (MCMC)
- **Variational methods (VB) and Belief Propagation (BP)**
  Approximate likelihood $P(\theta)$ by factorized form $Q(\theta)$ that is tractable
  VB: minimize the Kulback Leibler divergence $KL(P(\theta)|Q(\theta))$
  BP: minimize the Kulback Leibler divergence $KL(Q(\theta)|P(\theta))$

(Notice: MAP estimation admits direct use of standard optimization tools)

Automatic Relevance Determination (ARD)

- Automatic Relevance Determination (ARD) is a hierarchical Bayesian approach widely used for model selection
- In ARD hyper-parameters explicitly represents the relevance of different features by defining their range of variation.
  (i.e., Range of variation$\rightarrow$0 $\Rightarrow$ Feature removed)

A motivating example: A Bayesian formulation of the Lasso /Basis Pursuit Denoising (BPD) problem

$$\text{LASSO/BPD:} \quad \arg\min_{\mathbf{s}} \frac{1}{2\sigma^2}\|\mathbf{x}^I - \mathbf{A}^{I\times J}\mathbf{s}^J\|_F^2 + \lambda|\mathbf{s}|_1$$

$$P(\mathbf{x}|\mathbf{A},\mathbf{s},,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}^I} e^{-\frac{\|\mathbf{x}-\mathbf{A}\mathbf{s}\|_F^2}{2\sigma^2}}$$ **Likelihood**

$$P(\mathbf{s}|\lambda) = (\frac{\lambda}{2})^J e^{-\lambda|\mathbf{s}|_1}$$ **Prior**

$$P(\mathbf{s}|\mathbf{A},\mathbf{x},\sigma^2,\lambda) = \frac{P(\mathbf{x}|\mathbf{A},\mathbf{s},\sigma^2)P(\mathbf{s}|\lambda)}{P(\mathbf{x})}$$ **Bayes**

$$-\log P(\mathbf{s}|\mathbf{A},\mathbf{x},\sigma^2,\lambda) = -\log P(\mathbf{x}|\mathbf{A},\mathbf{s},\sigma^2) - \log P(\mathbf{s}|\lambda) + \log P(\mathbf{x})$$

$$= \underbrace{\frac{\|\mathbf{x}-\mathbf{A}\mathbf{s}\|_F^2}{2\sigma^2} + \lambda|\mathbf{s}|_1}_{\text{BPD / LASSO term}} + \underbrace{\frac{I}{2}\log 2\pi\sigma^2 - J\log\frac{\lambda}{2}}_{\text{Normalization terms}} + Const.$$

$$\boxed{\frac{\partial \log P(\mathbf{s}|\mathbf{A},\mathbf{x},\sigma^2,\lambda)}{\partial\lambda} = 0 \Rightarrow \lambda = \frac{J}{|\mathbf{s}|_1}}$$

ARD in reality a ℓ_0-norm optimization scheme. As such ARD based on Laplace prior corresponds to ℓ_0-norm optimization by re-weighted ℓ_1-norm

In particular if we define λ for each entry in $\mathbf{s}$, i.e.

$$\frac{1}{2\sigma^2}\|\mathbf{x}^I - \mathbf{A}^{I\times J}\mathbf{s}^J\|_F^2 + \sum_j \lambda_j |\mathbf{s}_j|$$

Corresponding to the Laplace prior $P(\mathbf{s}|\boldsymbol{\lambda}) = \prod_j \frac{\lambda_j}{2} e^{-\lambda_j |s_j|}$ optimizing for λ_j gives $\lambda_j = \frac{1}{|s_j|}$ such that

$$\frac{1}{2\sigma^2}\|\mathbf{x}^I - \mathbf{A}^{I\times J}\mathbf{s}^J\|_F^2 + \sum_j \frac{|\mathbf{s}_j|}{|\widetilde{\mathbf{s}}_j|}$$

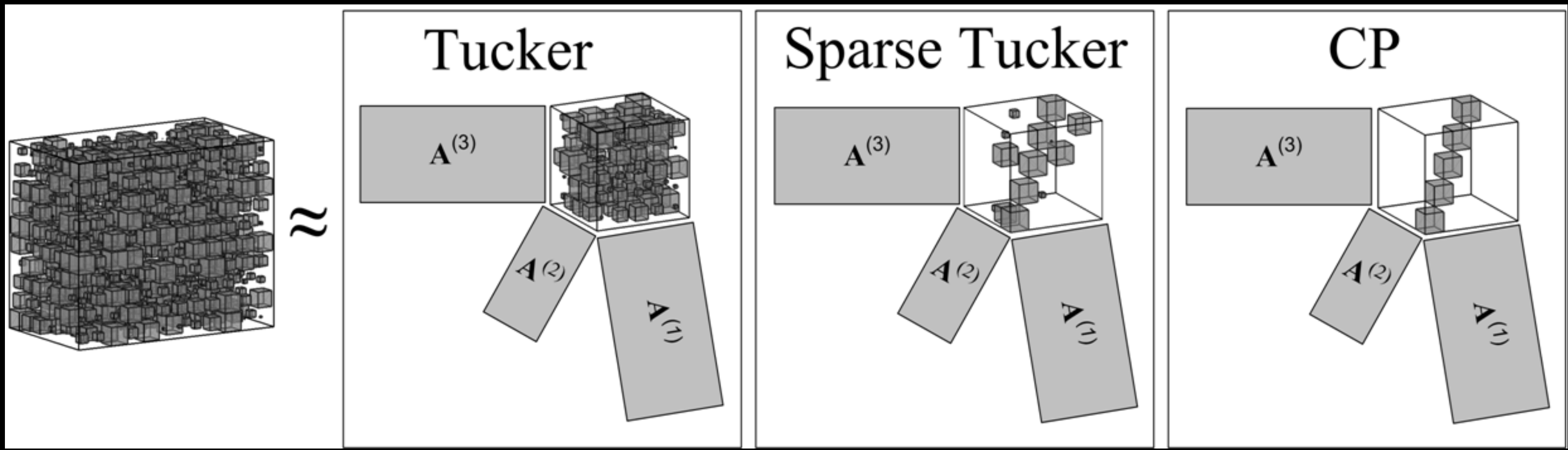
ℓ_0 norm by re-weighted ℓ_2 follows by imposing Gaussian prior instead of Laplace

Notice that we are all the time monotonically decreasing

$$-\log P(\mathbf{s}|\mathbf{A}, \mathbf{x}, \sigma^2, \lambda)$$

Agenda for model order selection

- To use regularization to simplify the Tucker core forming a unique representation as well as enable interpolation between the Tucker (full core) and CP (diagonal core) model.
- To use regularization to turn off excess components in the CP and Tucker model and thereby select the model order.
- To tune the regularization strength from data by Automatic Relevance Determination (ARD) based on Bayesian learning.



(Mørup and Hansen, Journal of Chemometrics 2009)

Sparse Tucker decomposition by ARD

$$
\begin{aligned}
P(\mathcal{X}|\mathcal{R},\sigma^2) &= (2\pi\sigma^2)^{-\frac{I_1 I_2 \cdots I_N}{2}} e^{-\frac{\|\mathcal{X}-\mathcal{R}\|_F^2}{2\sigma^2}} \\
P(\mathcal{G}|\alpha^{\mathcal{G}}) &= \left(\frac{\alpha^{\mathcal{G}}}{2}\right)^{J_1 J_2 \cdots J_N} e^{-\alpha^{\mathcal{G}}|\mathcal{G}|_1} \\
P(\mathbf{A}^{(n)}|\boldsymbol{\alpha}^{(n)}) &= \prod_{j_n} \left(\frac{\alpha^{(n)}}{2}\right)^{I_n} e^{-\alpha_j^{(n)}|\mathbf{a}_j|_1}
\end{aligned}
$$

CP follows setting $\mathcal{G} = \mathcal{I}$

$$
\begin{aligned}
L &= P(\mathcal{G}, \boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(N)}|\mathcal{X}, \sigma^2, \boldsymbol{\alpha}^{\mathcal{G}}, \boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(n)}) \\
&\propto P(\mathcal{X}|\mathcal{R},\sigma^2) P(\mathcal{G}|\alpha^{\mathcal{G}}) P(\boldsymbol{A}^{(1)}|\boldsymbol{\alpha}^{(1)}) \cdots P(\boldsymbol{A}^{N}|\boldsymbol{\alpha}^{(N)}).
\end{aligned}
$$

Thus the negative log likelihood based on Laplace priors is proportional to

$$
-\log L \propto \mathrm{c} + \frac{1}{2\sigma^2}\|\mathcal{X}-\mathcal{R}\|_F^2 + \sum_n \sum_{j_n} \alpha_{j_n}^{(n)} |\boldsymbol{a}_{j_n}^{(n)}|_1 + \alpha^{\mathcal{G}}|\mathcal{G}|_1
$$

$$
+\frac{1}{2} I_1 I_2 \cdots I_N \log \sigma^2 - \sum_n \sum_{j_n} I_n \log \alpha_{j_n}^{(n)} - J_1 J_2 \cdots J_n \log \alpha^{\mathcal{G}}.
$$

Maximum a posteriori (MAP) estimation

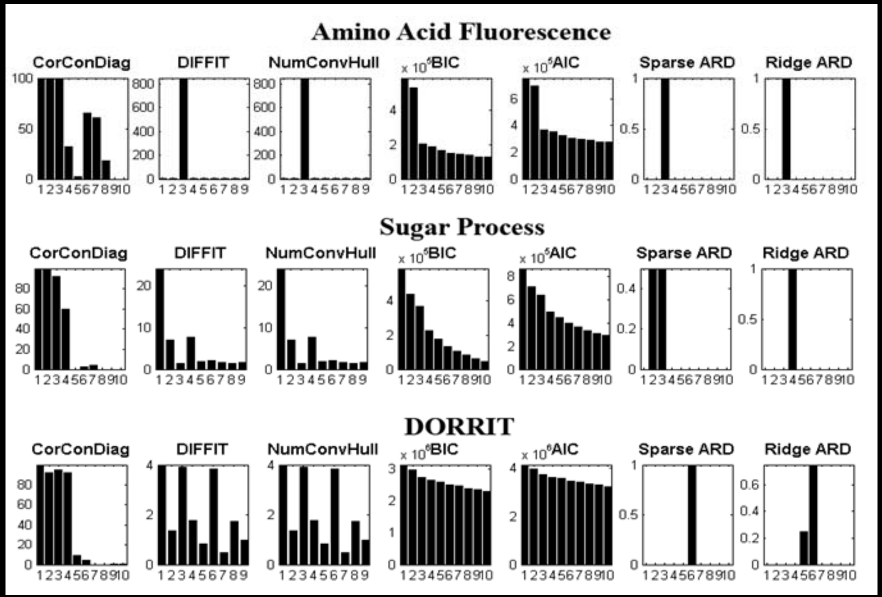
Brakes into standard Lasso/BPD sub-problems of the form

$$
\arg\min_{\boldsymbol{A}^{(n)}} \frac{1}{2\sigma^2} \|\boldsymbol{X}_{(n)} - \boldsymbol{A}^{(n)} \boldsymbol{Z}_{(n)}\|_F^2 + \sum_j \lambda_j |\boldsymbol{a}_j^{(n)}|_1
$$

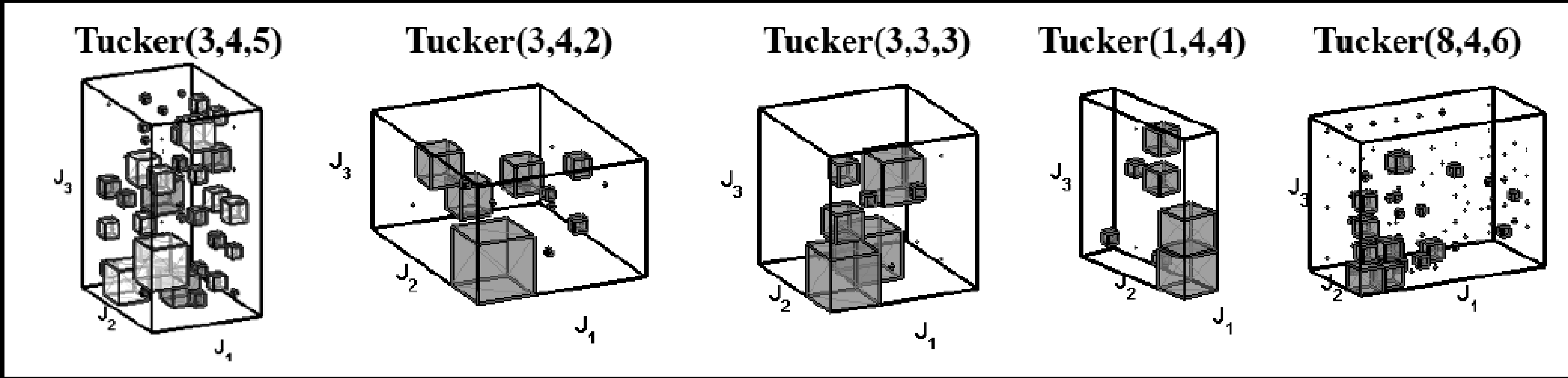
Update of regularization parameters by ARD

$$
\alpha_{\mathcal{G}} = \frac{J_1 J_2 \cdots J_N}{|\mathcal{G}|_1}, \quad \boldsymbol{\alpha}_d^{(n)} = \frac{J_n}{|\boldsymbol{A}_d^{(n)}|_1}
$$

CP models

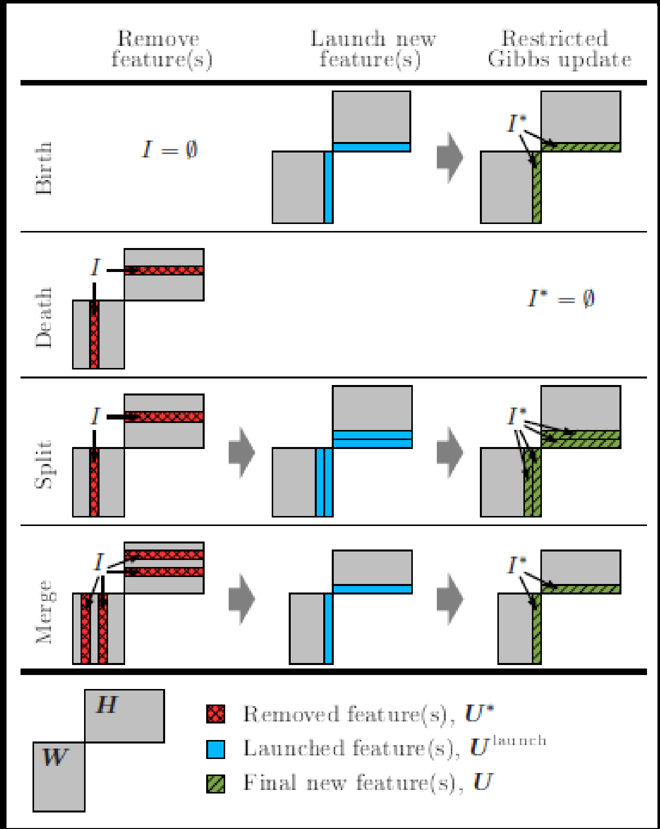


Tucker(10,10,10) models were fitted to the data, given are below the extracted cores



Synthetic Tucker(3,4,5) | Tucker(3,6,4) | Tucker(3,3,3) | Tucker(?,4,4) | Tucker(4,4,4)

Reversible jump Markov Chain Monte Carlo - a fully Bayesian approach to estimate parameter uncertainty and model order.



(For details see: Schmidt and Mørup, Infinite Non-negative Matrix Factorization, 2010)

Tensor models for complex networks

The Infinite Relational Model
(A Bayesian generative model for graphs)

Learning Systems of Concepts with an Infinite Relational Model (AAAI2006)


Charles Kemp


Josh Tenenbaum


Thomas Griffith


Takeshi Yamada


Naonori Ueda

See also: Infinite Hidden Relational Model (UAI 2006)


Zhao Xu

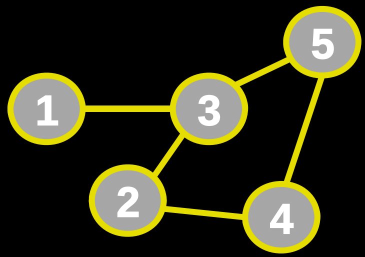

Kai Yu


Volker Tresp


Hans-Peter Kriegel

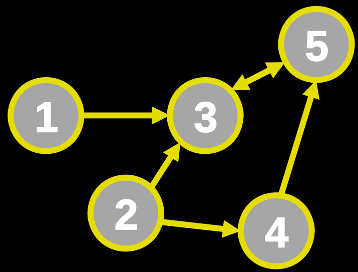
The relational model for various types of graphs:

UnDirected



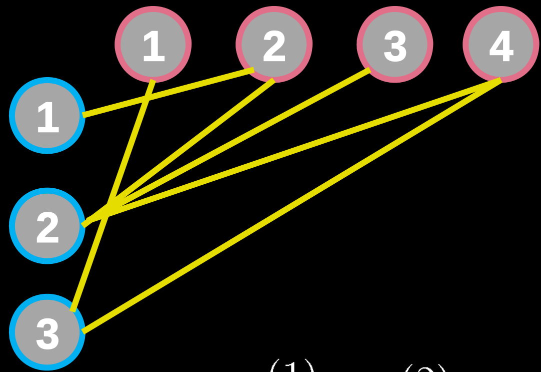
$$\pi_{ij} = \mathbf{z}_i \boldsymbol{\eta} \mathbf{z_j}$$
$$\boldsymbol{\eta} = \boldsymbol{\eta}^\top$$

Directed



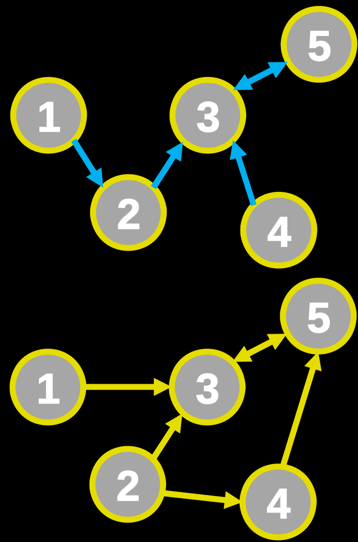
$$\pi_{ij} = \mathbf{z}_i \boldsymbol{\eta} \mathbf{z_j}$$

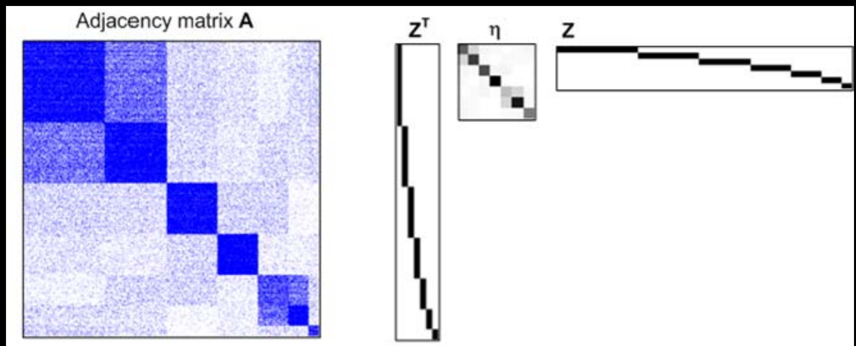
Bipartite



$$\pi_{ij} = \mathbf{z}_i^{(1)} \boldsymbol{\eta} \mathbf{z_j}^{(2)}$$

Multi-graph





The IRM statistical generative model for graphs:

$$\begin{aligned} \boldsymbol{Z} &\sim CRP(\alpha) \\ \eta_{ab} &\sim Beta(\gamma, \gamma) \\ A_{ij} &\sim Bernouilli(\pi_{ij}) \end{aligned}$$

Tucker

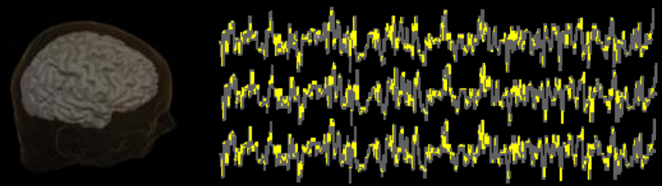
$$\pi_{ijk} = \sum_{lmn} \eta_{lmn} z_{lj}^{(1)} z_{mi}^{(2)} z_{nk}^{(3)}$$

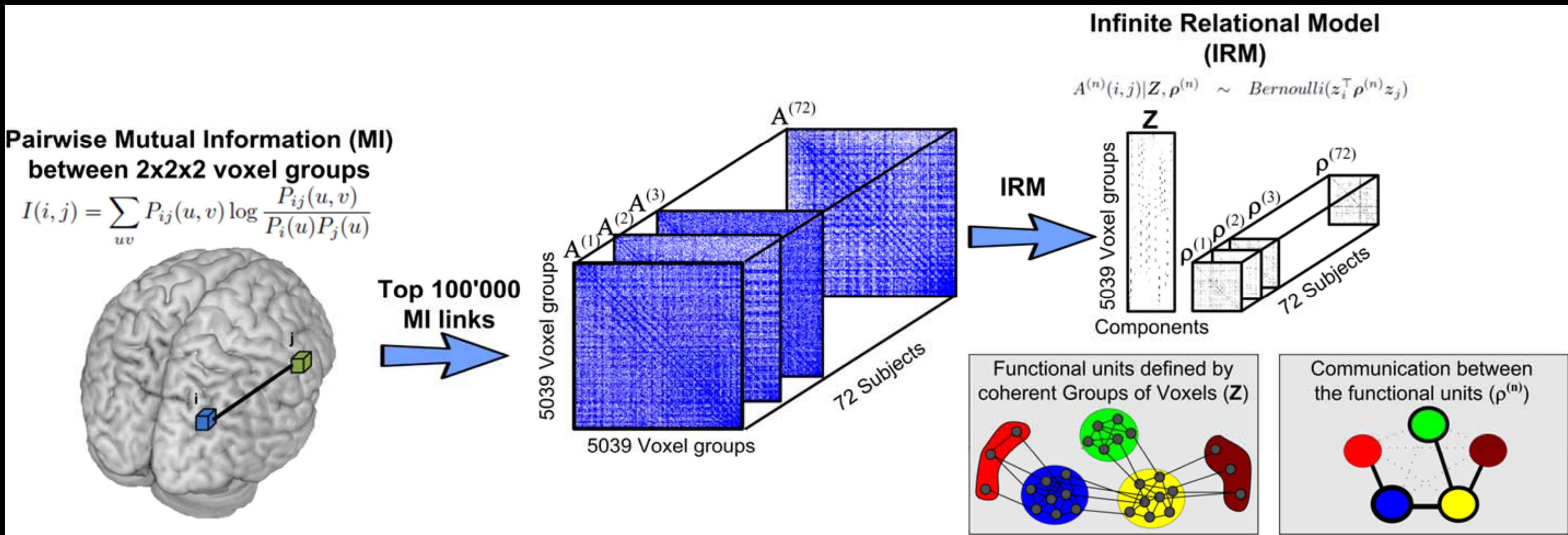
Tucker2

$$\pi_{ij}^{(k)} = \mathbf{z}_i^{(1)} \boldsymbol{\eta}^{(k)} \mathbf{z}_j^{(2)}$$

Potential symmetry constraints, i.e. $\mathbf{z}^{(1)} = \mathbf{z}^{(2)}$

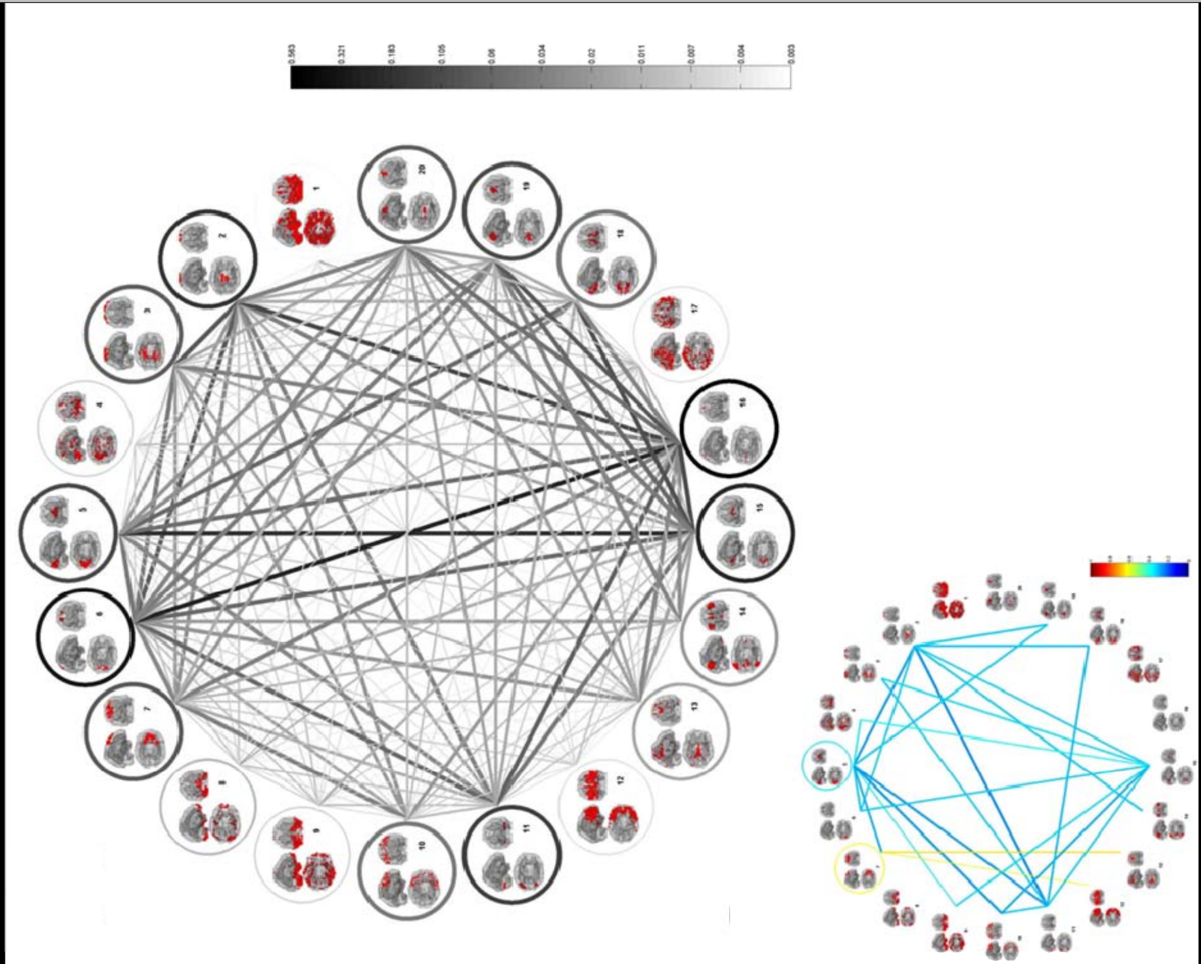
Modeling the consistent functional connectivity of the brain





72 subjects: 42 multiple schlerosis and 30 normal subjects

(Mørup et al., to appear NIPS 2010)



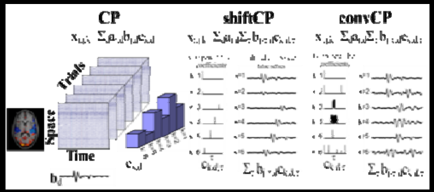
	Raw data	PCA	ICA	Degree	IRM
SVM	51.39	55.56	**63.89** ($p \leq 0.04$)	59.72	**72.22**($p \leq 0.002$)
LDA	59.72	51.39	**63.89** ($p \leq 0.05$)	51.39	**75.00**($p \leq 0.001$)
KNN	38.89	58.33	56.94	51.39	**66.67**($p \leq 0.01$)

(Mørup et al., to appear NIPS 2010)

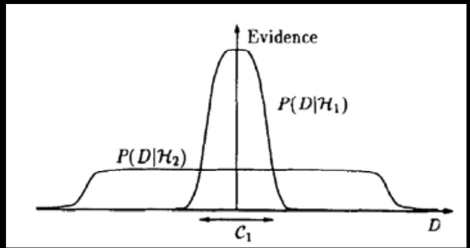
Summary

Multi-linear modeling offers the ability to explicitly extract the most consistent activity of neuroimaging data across repeats/subjects/conditions.

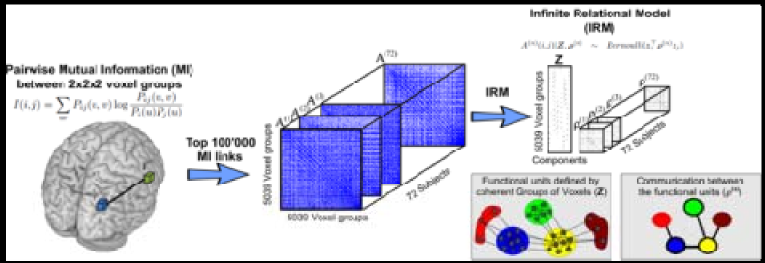
Common causes of variability in neuroimaging data are latency and shape changes-> **shiftCP and convCP**



Important problem in tensor decomposition is to adequately selected the number of components. **Bayesian learning** admits a general framework for model order selection and regularization tuning.



From neuroimaging data complex networks of functional connectivity can be derived. The **Infinite Relational Model** forms an efficient modeling framework for exploring consistent structures in these networks.



AIM of all the described analyses

- Extract an efficient internal representation of the statistical structure implicit in the data
- Drive novel hypothesis for formal statistical testing

Relevant papers

M. Mørup, K. H. Madsen, A. M. Dogonowski , L. K. Hansen, H. Siebner, Infinite Relational Modeling of Functional Connectivity in Resting State fMRI, to appear NIPS 2010

M. Mørup, Applications of tensor (multi-way array) factorizations and decompositions in data mining models in data mining, to appear Wiley DMKD 2010.

M.N. Scmidt, M. Mørup, Infinite Non-negative Matrix Factorization, EUSIPCO 2010

M. Mørup, L.K. Hansen, Automatic Relevance Determination for multi-way models, Journal of Chemometrics, 2009

M. Mørup, L.K. Hansen, S.M. Arnfred, L.-K. Lim, K.M. Madsen, Shift Invariant Multilinear Decomposition of Neuroimaging Data, NeuroImage vol. 42(4), pp.1439-50, 2008

M. Mørup, Kristoffer H. Madsen, L.K. Hansen Modeling trial based neuroimaging data, Nips workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis, 2008

M. Mørup, S.M. Arnfred, L.K.Hansen, Algorithms for Sparse Non-negative TUCKER, *Neural Computation*, vol. 20(8), pp. 2112-2131, 2008

Mørup, M., Hansen, L. K., Arnfred, S. M., ERPWAVELAB A toolbox for multi-channel analysis of time-frequency transformed event related potentials, *Journal of Neuroscience Methods*, vol. 161, pp. 361-368, 2007

M. Mørup, L. K. Hansen, C. S. Hermann, J. Parnas, S. M. Arnfred, Parallel Factor Analysis as an exploratory tool for wavelet transformed event-related EEG, *NeuroImage*, vol. 29(3), pp. 938-947, 2006