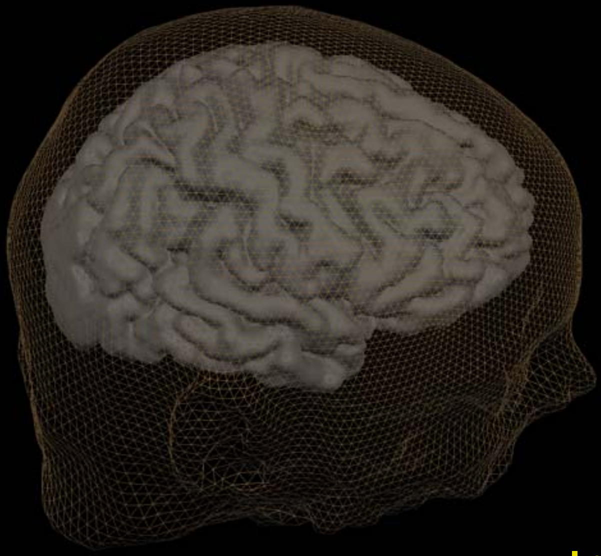
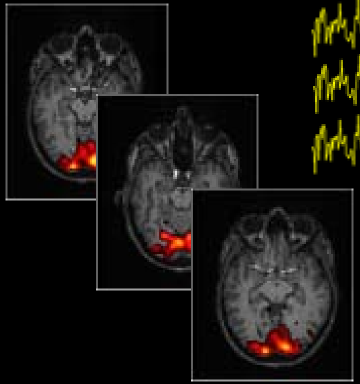
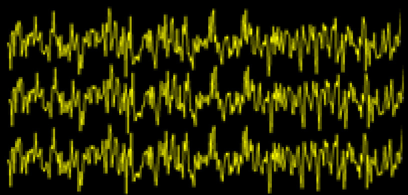
Unsupervised Multi-way Decompositions

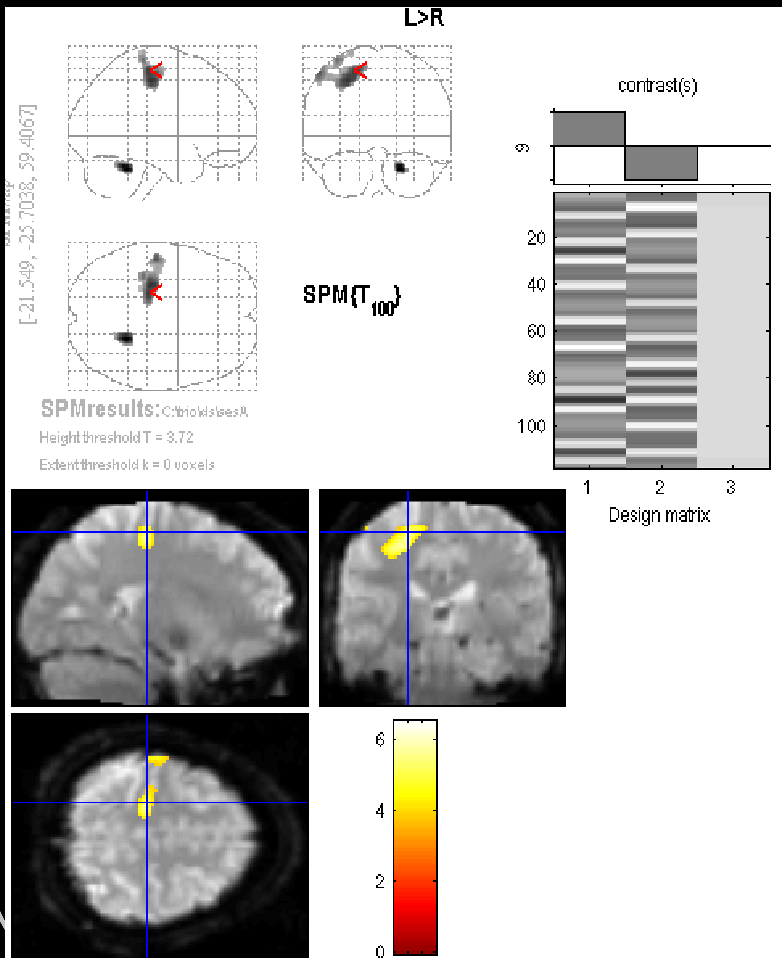






Morten Mørup
Informatics and Mathematical Modeling
Cognitive Systems
Technical University of Denmark
Acknowledgement: Kristoffer Hougaard Madsen and Lars Kai Hansen

SPM and univariate statistical analysis



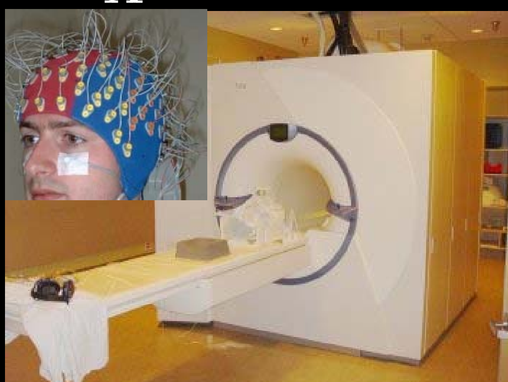
Problems:

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 (independent) sources in the data (beyond GLM)

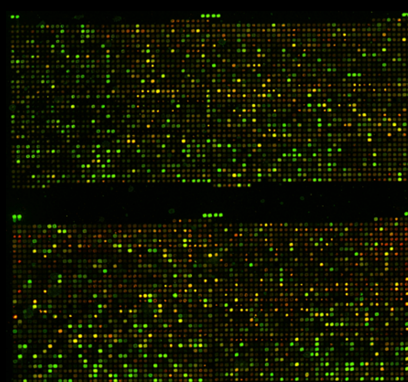
This problem is no different than the problems encountered in general in Modern Massive Datasets (MMDS)

$\mathbf{X}^{Space \times Time}$



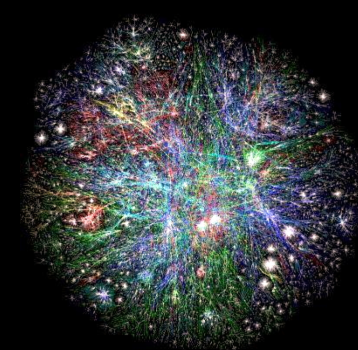
NeuroInformatics

$\mathbf{X}^{Gene\ seq. \times Samples}$



BioInformatics

$\mathbf{X}^{Webpages \times Webpages}$



ComplexNetworks

$\mathbf{X}^{Term \times Document}$

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

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

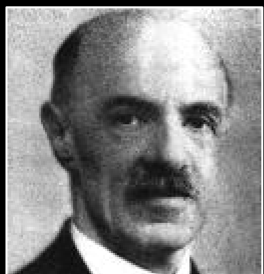
THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental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

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

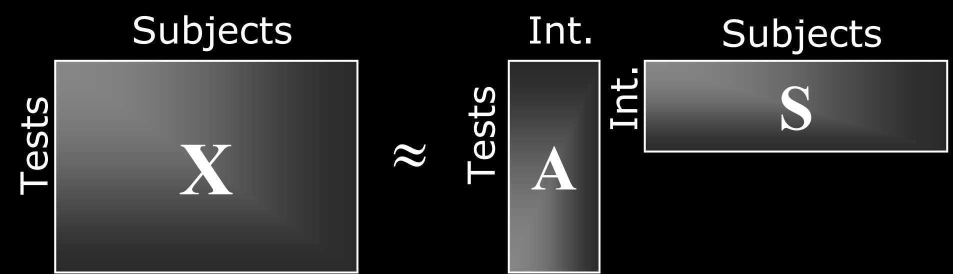
So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have 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!!

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)

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

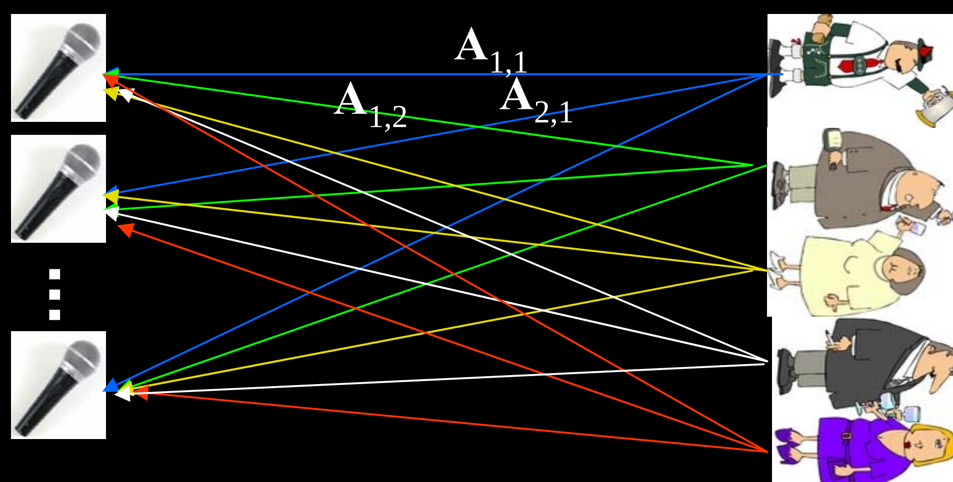
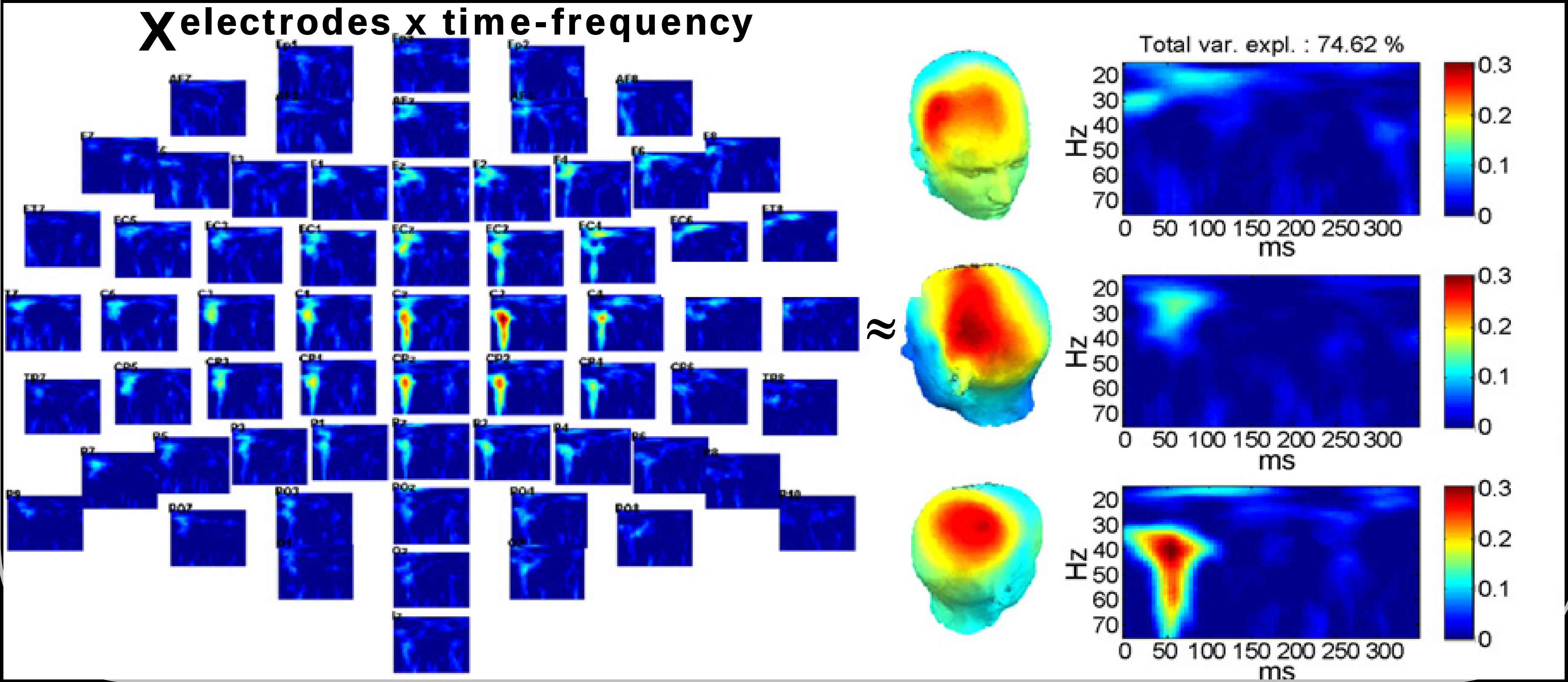
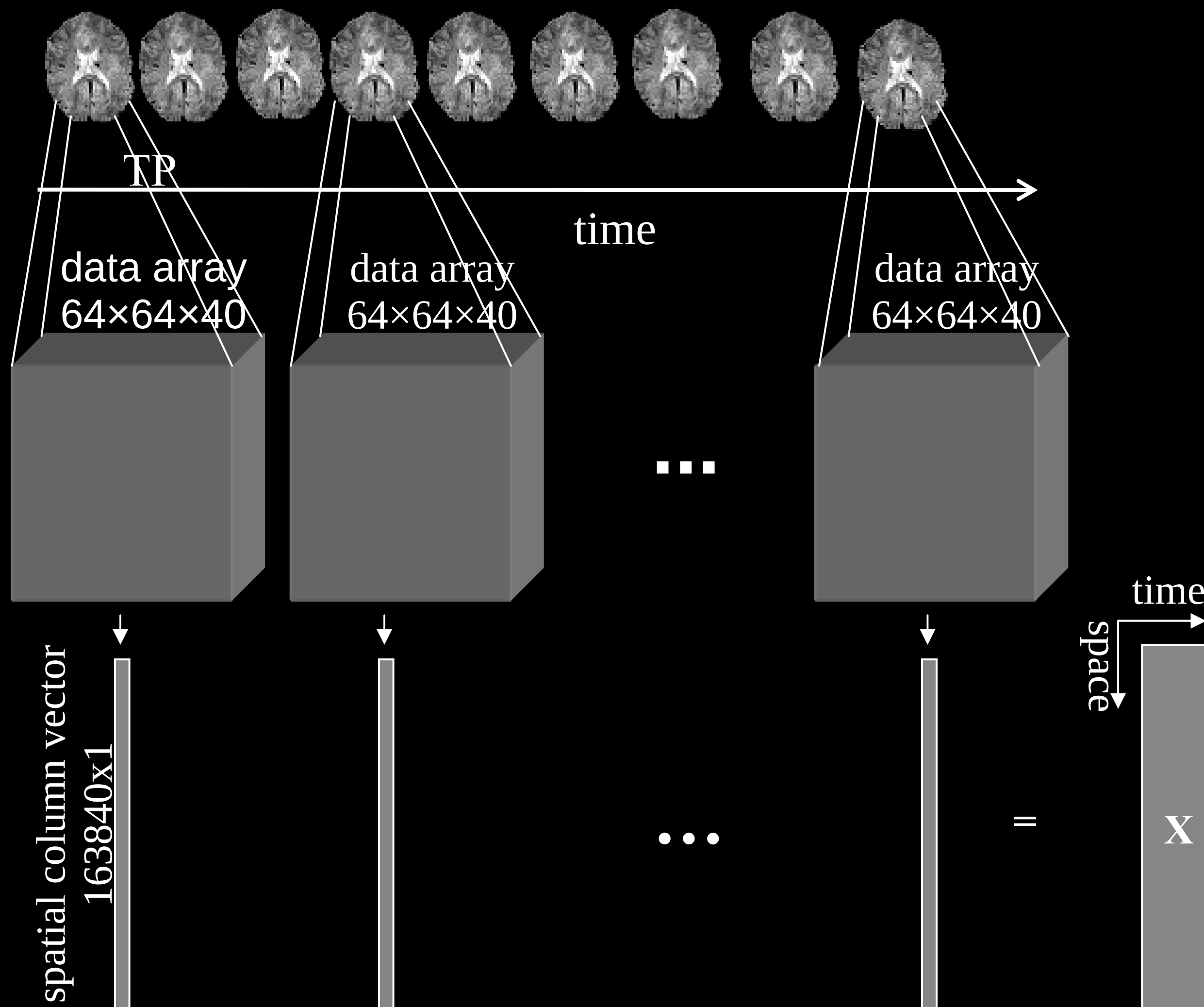


Illustration of Factor Analysis on frequency transformed EEG

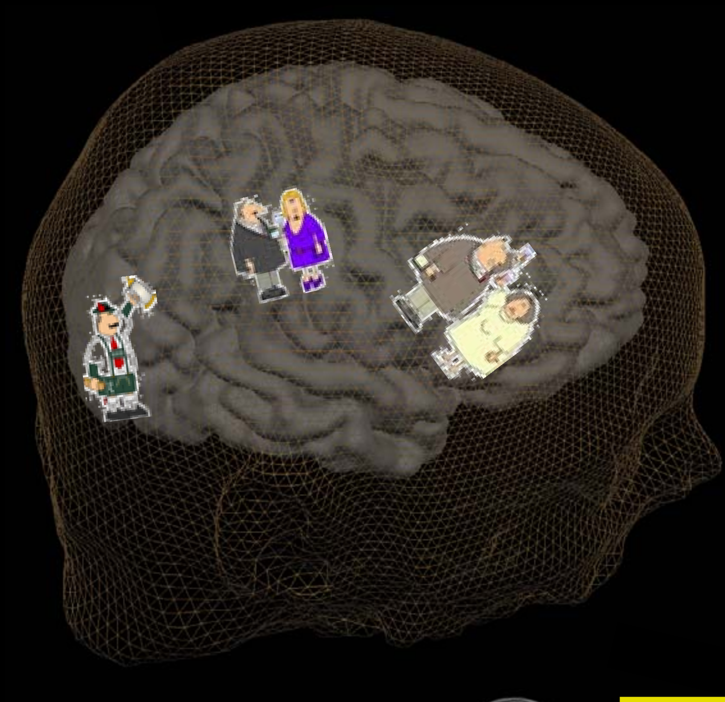


fMRI in matrix fom



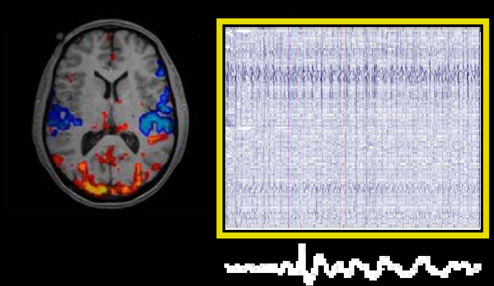
data matrix (*X*) of space×time

The fMRI Party problem



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

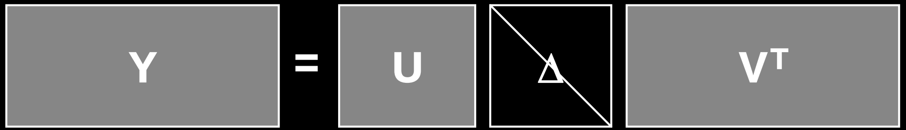
Assumption: Data **instantaneous** mixture of temporal signatures. (PCA/ICA/NMF)



$$\approx \sum_d$$

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

Singular Value Decomposition (SVD)



$$Y = U \Delta V^\top$$

$$U^\top U = I, \quad V^\top V = I$$

Δ diagonal

- Unique (up to permutation of components)
- Equivalent to PCA
- Convex optimization problem
  (one global solution – easy to find)
- Sort components according to singular values
- Truncate to obtain approximate model
- The orthogonality constraint is often not appropriate
- Spatial/temporal versions are equivalent

Louis L. Thurstone
(1887-1955)

“In a factor problem one is concerned about how to account for the observed correlations among all the variables in terms of the smallest number of factors and with the smallest possible residual error."

Thurstone, 1947

This quote inspired in the 50-70’s the now classical psychometric rotation criteria such as:

Varimax, Quartimax, Orthomax

Goal of rotation criteria: A large loading in one factor be opposite small loadings of the remaining factors $\Rightarrow$ histogram of loadings should have high peak around zero and heavy tails (forming sparse distribution)

Independent Component Analysis
(A modern approach to the classic rotation problem)

InfoMAX / ML: Optimize distribution of sources assumed independent and non-gaussian (Bell & Sejnowski, 1995)


$$\log L = \sum_j \log f(\boldsymbol{Q}\mathrm{s}_j) + \log|\det(\boldsymbol{Q})|.$$

Optimize deviation from normality: For instance as measured by kurtosis (Comon, 1994, Girolami 1996, Pearlmutter 1996, Hÿvarinen 1997)

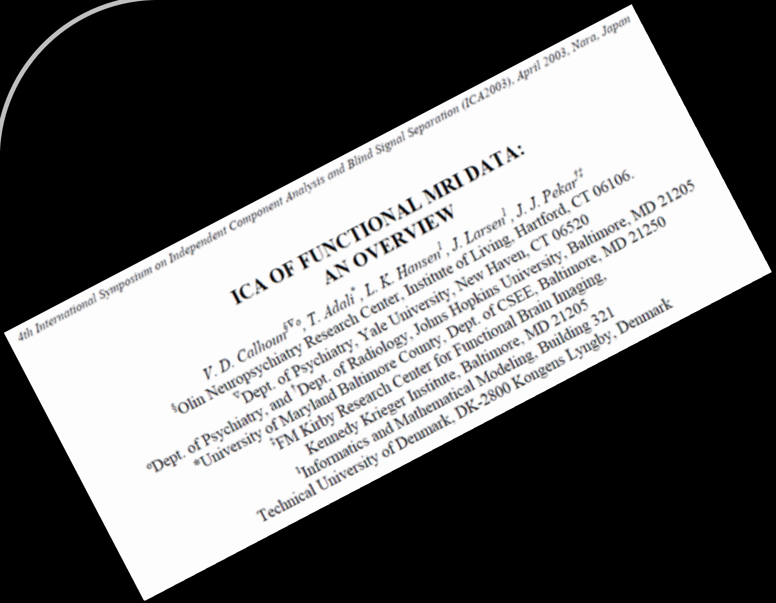
$$kurt(\boldsymbol{S}) = E[\boldsymbol{S}^{.4}] - 3E[\boldsymbol{S}^{.2}]^2$$

Jointly diagonalize some higher order moments, cumulants, autocorrelations (Comon, 1994, Molgedey & Schuster 1994)

$$\mathcal{C}^{\mathbf{X}}_{i,j,k,l} \approx \sum_d \mathbf{A}_{i,d}\mathbf{A}_{j,d}\mathbf{A}_{k,d}\mathbf{A}_{l,d}\mathcal{C}^{\mathbf{S}}_{d,d,d,d}$$

infoMAX/ML based on sparse priors and maximization of *kurt(**S**)* equivalent to the former rotation criteria.

DTU



4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

ICA OF FUNCTIONAL MRI DATA: AN OVERVIEW

V. D. Calhoun[§∇∘], T. Adali[*], L. K. Hansen[ℓ], J. Larsen[ℓ], J. J. Pekar[†‡]
[§]Olin Neuropsychiatry Research Center, Institute of Living, Hartford, CT 06106.
[∇]Dept. of Psychiatry, Yale University, New Haven, CT 06520
[∘]Dept. of Psychiatry, and [†]Dept. of Radiology, Johns Hopkins University, Baltimore, MD 21205
[*]University of Maryland Baltimore County, Dept. of CSEE, Baltimore, MD 21250
[‡]FM Kirby Research Center for Functional Brain Imaging,
Kennedy Krieger Institute, Baltimore, MD 21205
[ℓ]Informatics and Mathematical Modeling, Building 321
Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

ICA on fMRI

(Example of single subject analysis)

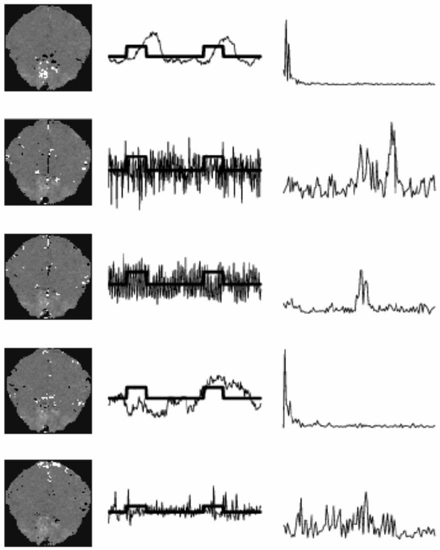


Figure 2.1: Images (left), time courses (center), and amplitude spectrum (right) from a five component temporal ICA analysis of a visual stimulation fMRI data set

Stimuli full-checkerboard (8Hz), each trial consist of 10 seconds pause 10 seconds stimuli and 10 seconds pause. Data acquired at 3 Hz.

Two other important factor analytic type approaches

Sparse Coding

Bruno A. Olshausen

Nature, 1996

Emergence of simple-cell receptive field properties by learning a sparse code for natural images

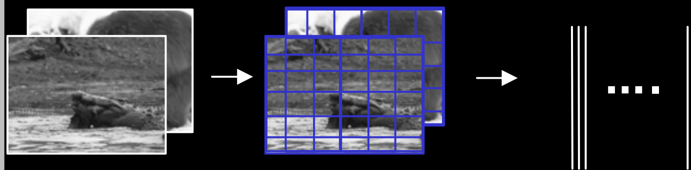
Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

David J. Field

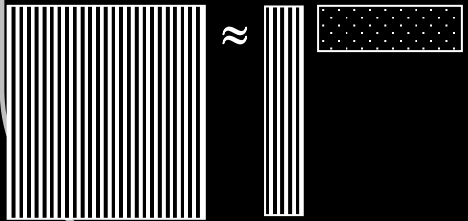
$$\underbrace{\text{argmin}_{A,S}\, D(X, AS)}_{\text{Preserve Information}} + \lambda \underbrace{sp(S)}_{\text{Preserve Sparsity (Simplicity)}}$$

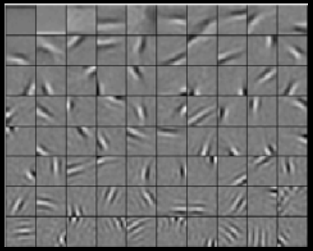
Tradeoff parameter



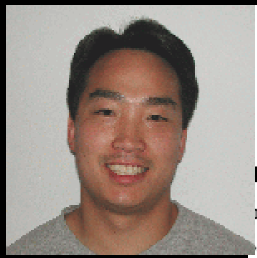
Sparse coding:
A corresponds to Gabor like features resembling simple cell behavior in V1 of the brain!

$$C(\mathbf{A},\mathbf{S}) = \frac{1}{2}\|\mathbf{X}^{pixels\times patches} - \mathbf{A}^{pixels\times feat.}\mathbf{S}^{feat.\times patches}\|_F^2 + \lambda|\mathbf{S}|_1$$





Non-negative Matrix Factorization



Daniel D. Lee

Nature 1999

non-negative matrix factorization

Daniel D. Lee* & H. Sebastian Seung*†

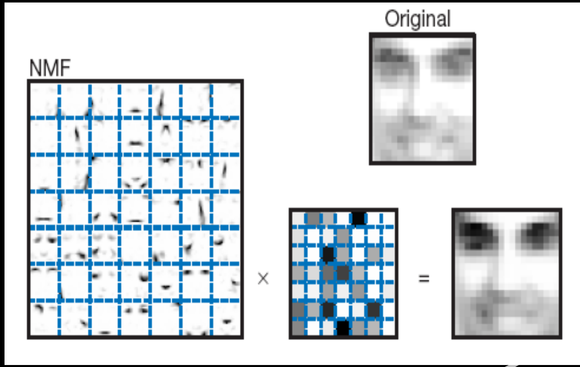
* Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA
† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Is perception of the whole based on perception of its parts? There

Sebastian Seung

$$X \approx AS,\ X \geq 0, A \geq 0, S \geq 0$$

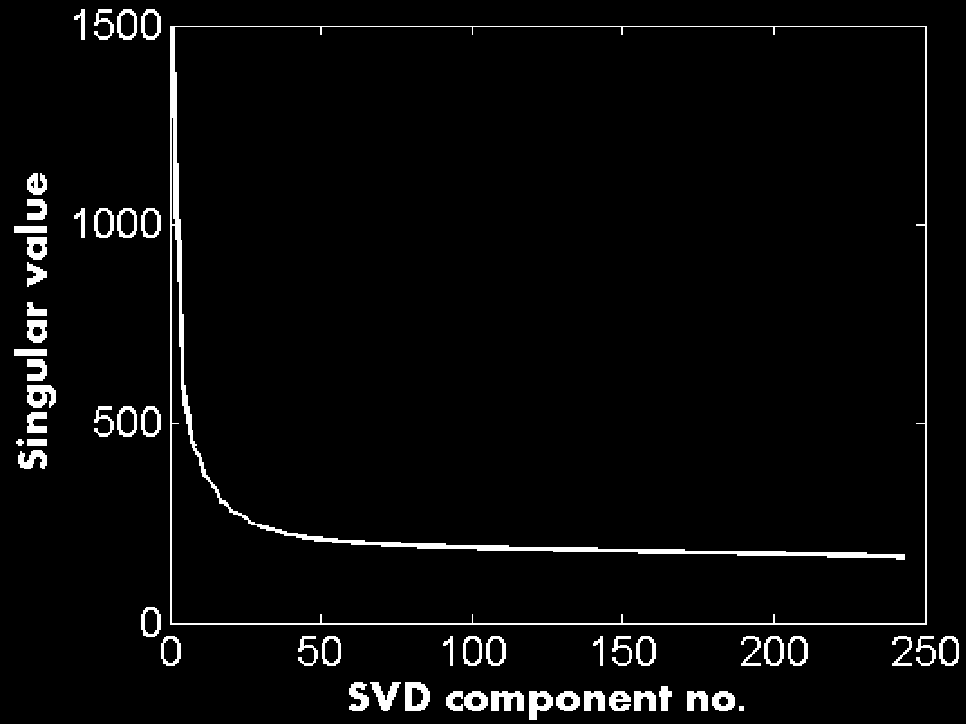
Non-negative Matrix Factorization:
gives Part based representation, i.e. the whole described by its constituting parts



Important challenge in Unsupervised Learning:
How many components adequately model the data

?

- The L-curve approach



- Bayesian Learning and Automatic Relevance Determination (Bayesian PCA)

William of Ockham

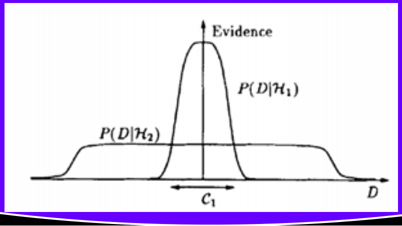
The explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory.

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

Thomas Bayes

David J.C. MacKay

Bayesian learning embodies Occam's razor, i.e. Complex models are penalized. The horizontal axis represents the space of possible data sets *D.* Bayes rule rewards models in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalized probability distribution on *D.*



Cross-validation

Anal Bioanal Chem (2008) 390:1241–1251
DOI 10.1007/s00216-007-1790-1

REVIEW

Cross-validation of component models: A critical look at current methods

R. Bro · K. Kjeldahl · A. K. Smilde · H. A. L. Kiers

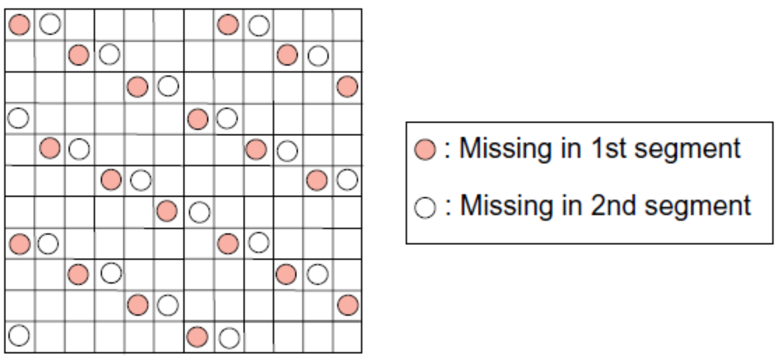


Fig. 1 The pattern of missing values used in Wold cross-validation for K=7

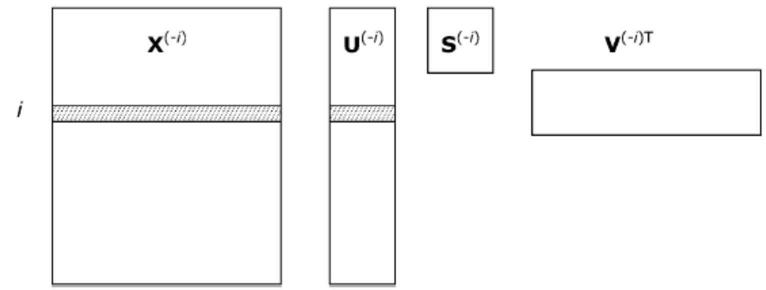


Fig. 2 Eastment–Krzanowski (EK) cross-validation. Data for one of two submodels is obtained by leaving out element x_{ij}. Likewise, the jth column of $\mathbf{X}$ is left out, whereby $\mathbf{X}^{(-j)}$, $\mathbf{U}^{(-j)}$, $\mathbf{S}^{(-j)}$ and $\mathbf{V}^{(-j)}$ are obtained

Split data into X^{train} and X^{test}, learn model parameters on X^{train} and use these parameters to predict X^{test}

Problem facing unsupervised learning: X^{train} and X^{test} hardly ever independent sets (i.e. noise correlated, rendering missing values not truly missing in the training set)

Other approaches

- Laplace approximation to the model evidence
- Bayesian Information Criterion (BIC) / Minimum Description Lenght (MDL)
- Akaike's Information Criterion (AIC)
- Final Prediction Error (FPE)

For all the above approaches a penalty term for model complexity is introduced based on some kind of asymptotic theory

$$L(\boldsymbol{X}|\mathcal{M}) + C(\mathcal{M})$$

Varimax

ICA

NMF

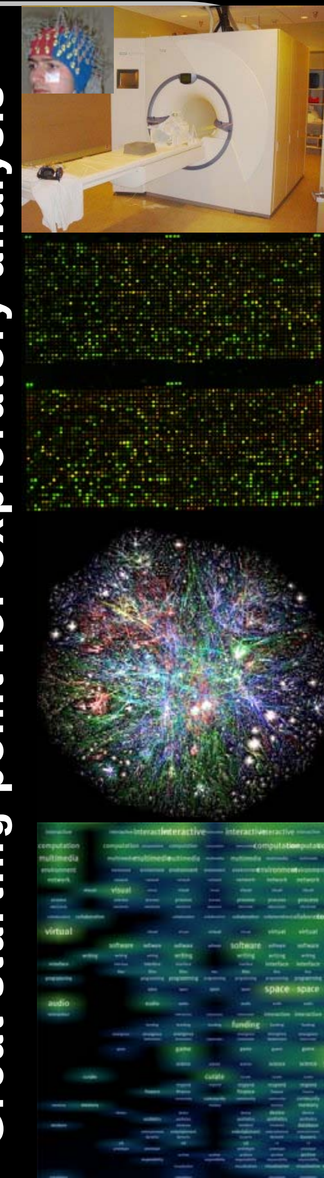
Sparse Coding

Model Selection

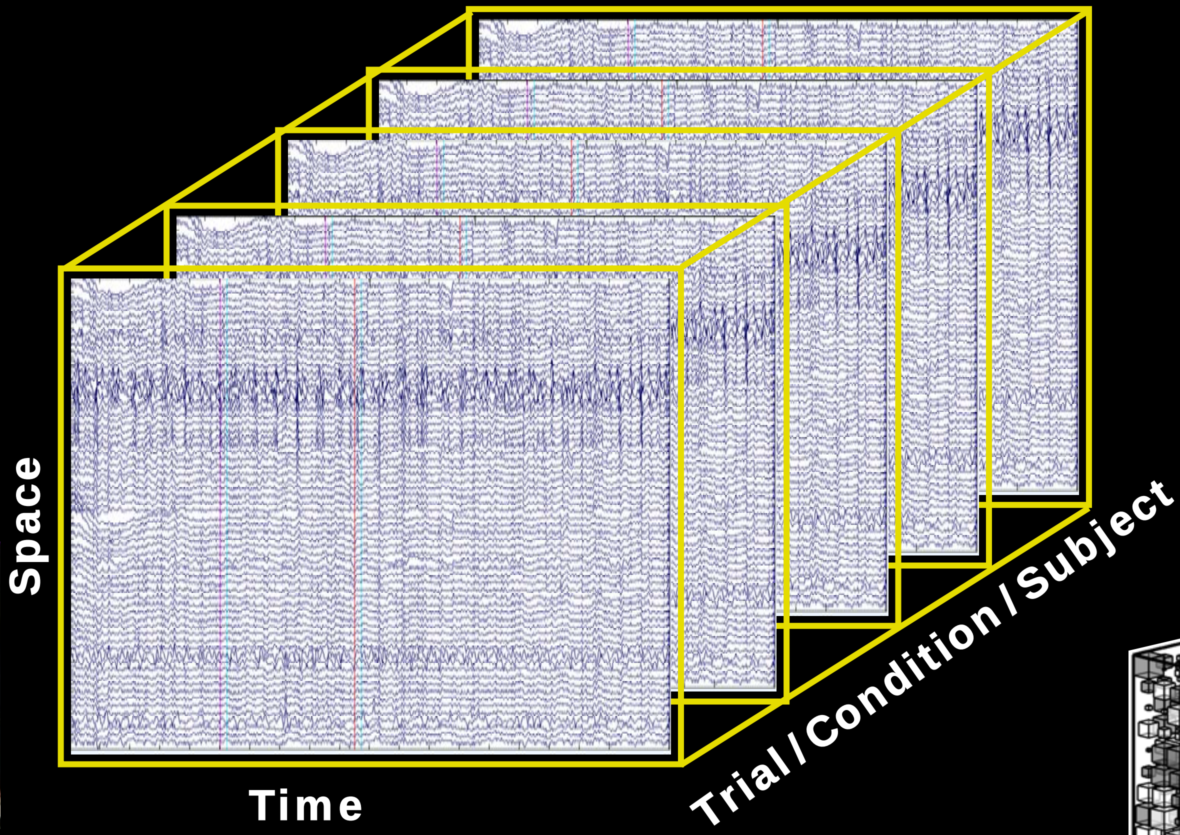
William of Ockham (1288-1347)

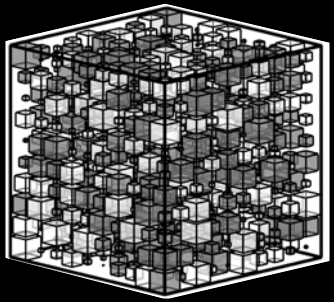
Principle of parsimony

Great starting point for exploratory analysis



From 2-way to multi-way analysis





Multi-subject analysis

- At least four possibilities:
  - Pre-average data
  - Separate analysis
  - Data concatenation
  - Tensor models

Pre-averaging

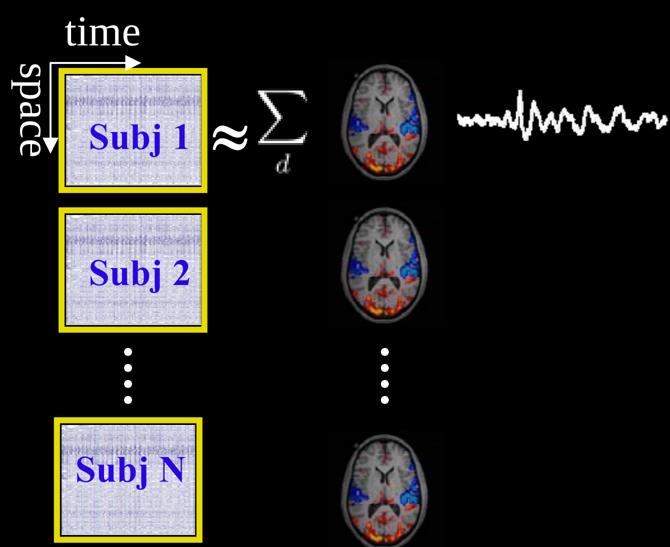
- Simply average data over subjects prior to analysis
  - Common spatial profiles
  - Common time profiles
  - Model must generalise in both space and time over subjects

Separate analysis

- Run analysis separately for each subject
  - Separate spatial maps for each subject
  - Separate time series for each subject
  - Cluster components after analysis to establish correspondence
  - Many parameters

Concatenation of multi-way data to 2-way

(identical time series varying spatial maps)

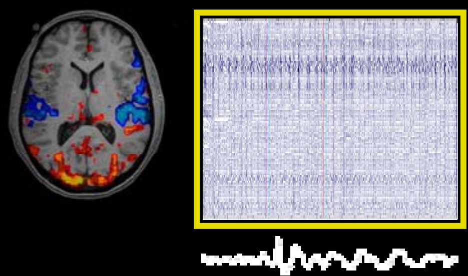


(identical spatial map, varying time series)

Multilinear modelling

Bilinear Model:

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

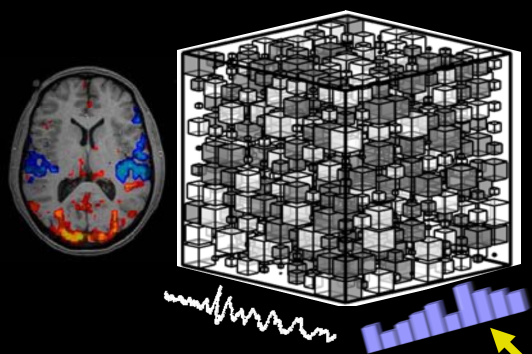


$$\sum_d$$

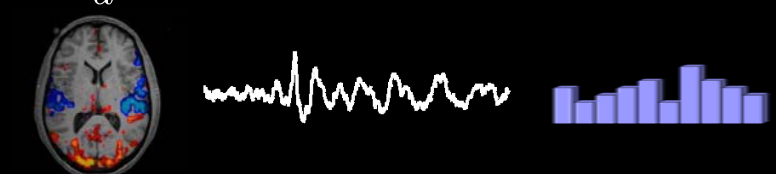
Assumption: Data **instantaneous** mixture of temporal signatures.
(PCA/ICA/NMF)

Trilinear Model:

$$\mathbf{X}^{\text{Voxel}\times\text{Time}\times Trial} \approx \sum_d \mathbf{a}_d^{Voxel}\mathbf{b}_d^{\text{Time}}\mathbf{c}_d^{Trial}$$



$$\sum_d$$

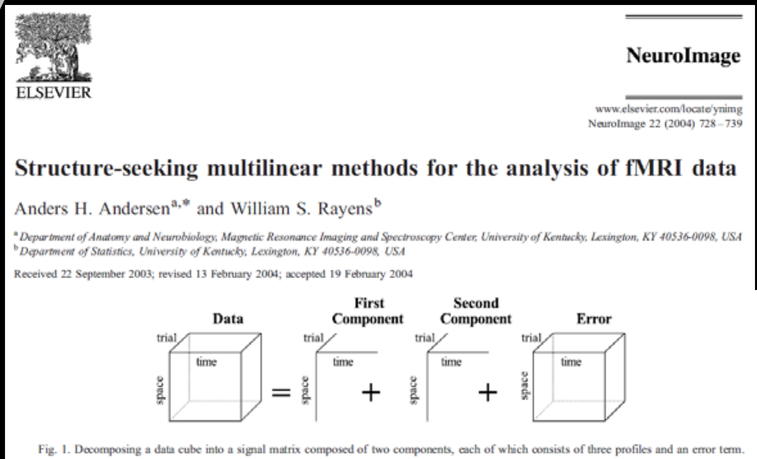


Assumption: Data **instantaneous** mixture of temporal signatures
that are expressed to **various degree** over the trials
(Canoncial Decomposition, Parallel Factor (CP))

(weighted averages over the trials)

"A surprising fact is that the nonrotatability characteristic can hold even when the number of factors extracted is greater than every dimension of the three-way array." - Kruskal 1976

Examples of Multiway analysis of fMRI and EEG



ELSEVIER
NeuroImage
www.elsevier.com/locate/ynimg
NeuroImage 22 (2004) 728 – 739

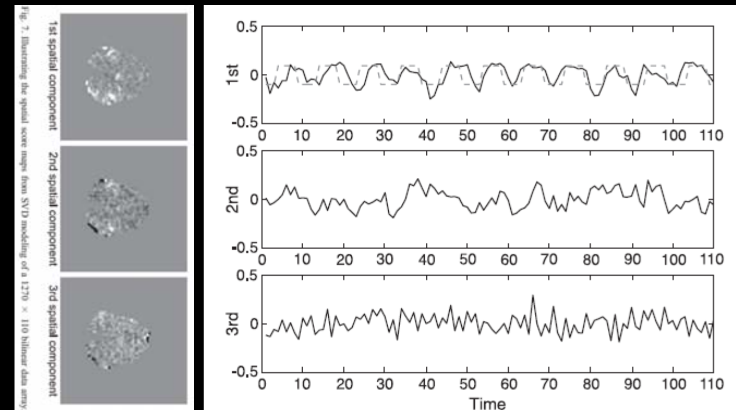
Structure-seeking multilinear methods for the analysis of fMRI data

Anders H. Andersen[a,*] and William S. Rayens[b]

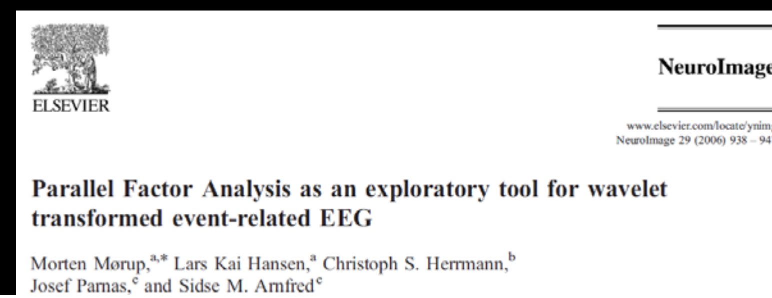
[a] Department of Anatomy and Neurobiology, Magnetic Resonance Imaging and Spectroscopy Center, University of Kentucky, Lexington, KY 40536-0098, USA
[b] Department of Statistics, University of Kentucky, Lexington, KY 40536-0098, USA

Received 22 September 2003; revised 13 February 2004; accepted 19 February 2004

Fig. 1. Decomposing a data cube into a signal matrix composed of two components, each of which consists of three profiles and an error term.



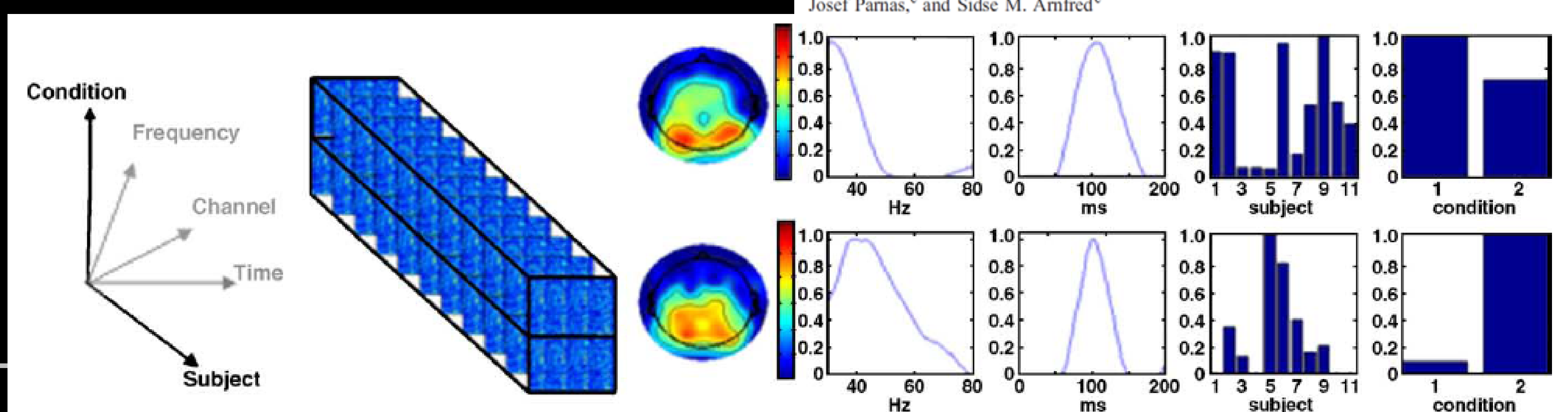
Extracts consisten activation alloving for subject / trial / condition dependent weights (i.e. "clever averaging")



ELSEVIER
NeuroImage
www.elsevier.com/locate/ynimg
NeuroImage 29 (2006) 938 – 947

Parallel Factor Analysis as an exploratory tool for wavelet transformed event-related EEG

Morten Mørup,[a,*] Lars Kai Hansen,[a] Christoph S. Herrmann,[b] Josef Parnas,[c] and Sidse M. Arnfred[c]



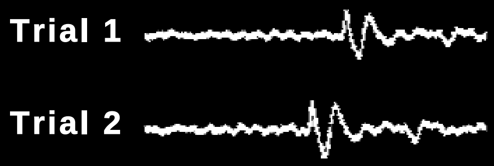
Unfortunately, multi-linear models are often to restrictive

Trilinear model can encompass:

- Variability in strength over repeats

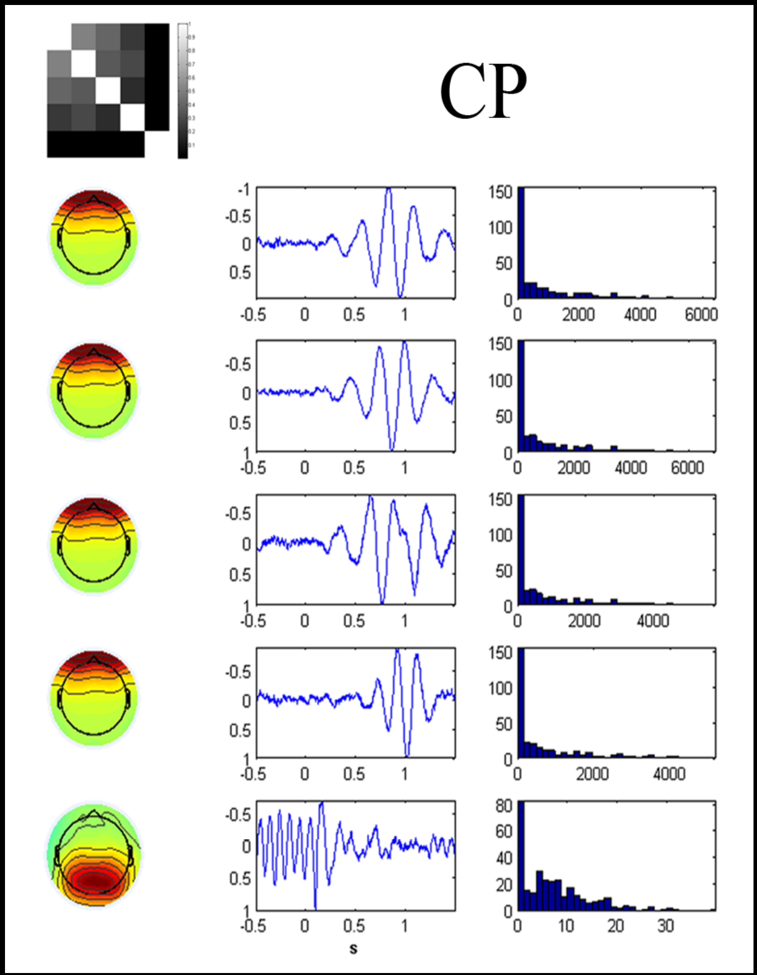
However, other common causes of variation are:

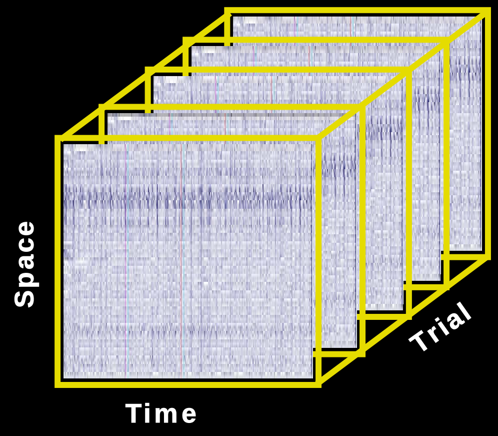
- Delay Variability

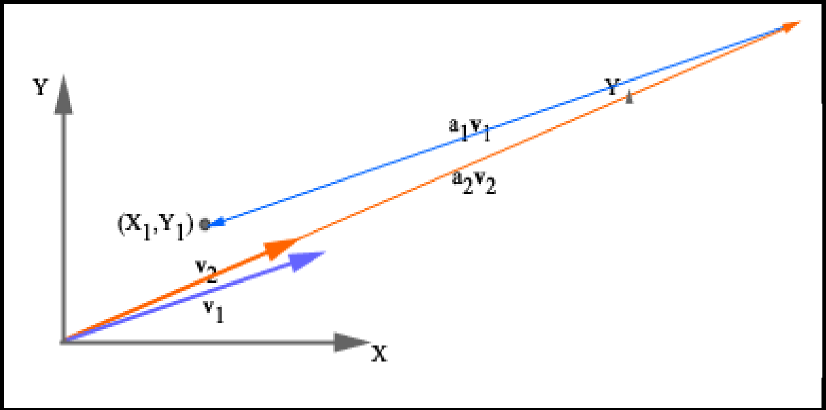


- Shape Variability

Violation of multi-linearity causes degeneracy



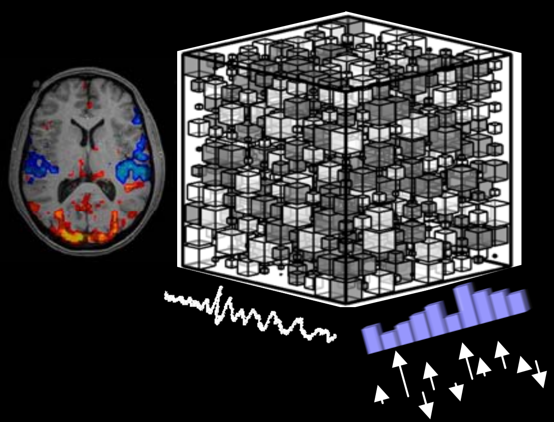


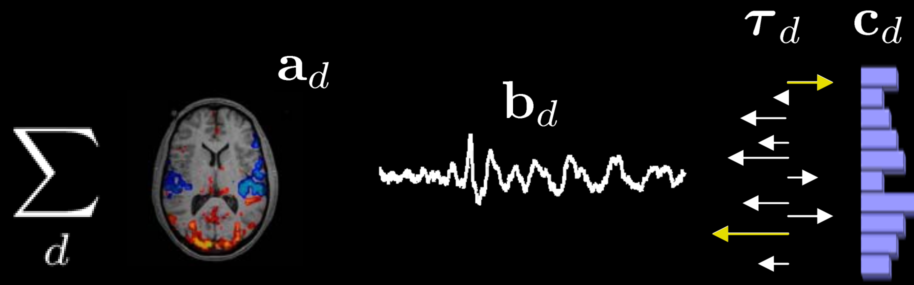


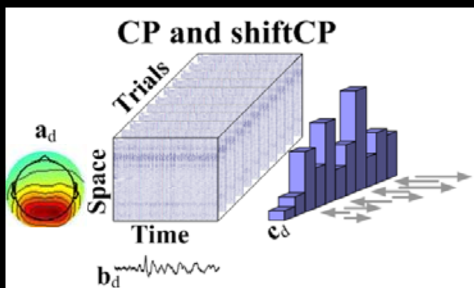
Modelling Delay Variability

Shifted CP:

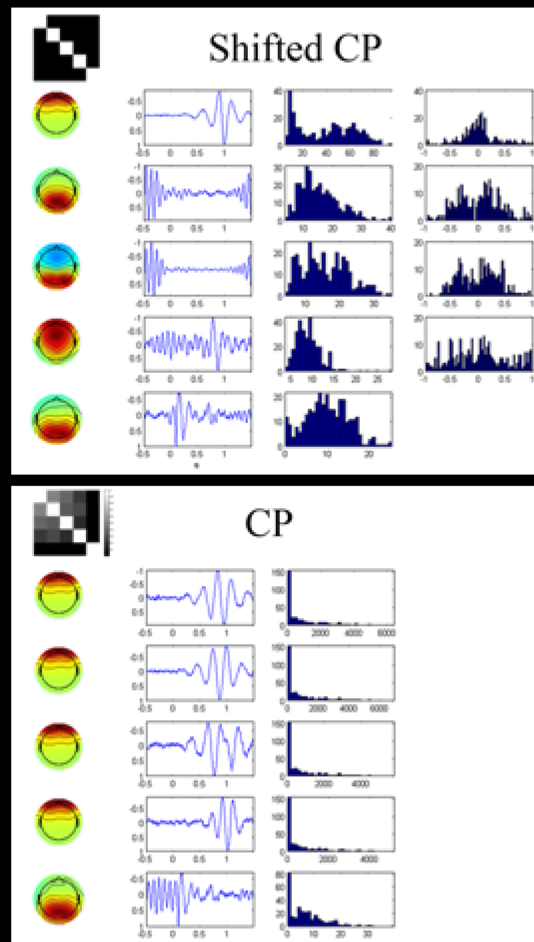
$$x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t - \boldsymbol{\tau}_{k,d}) c_{k,d}$$

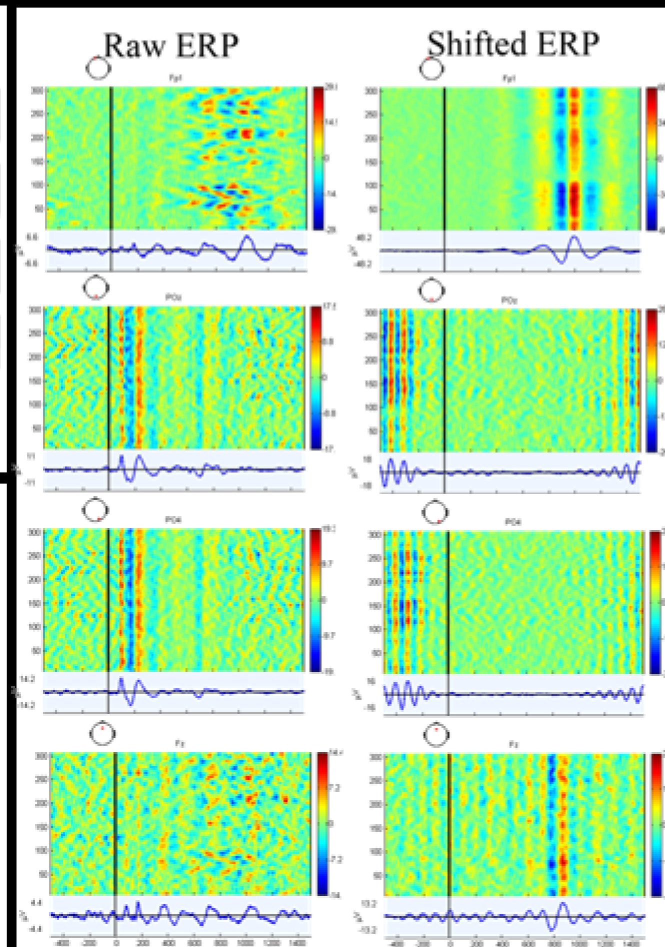






$$x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t - \boldsymbol{\tau}_{k,d}) c_{k,d}$$

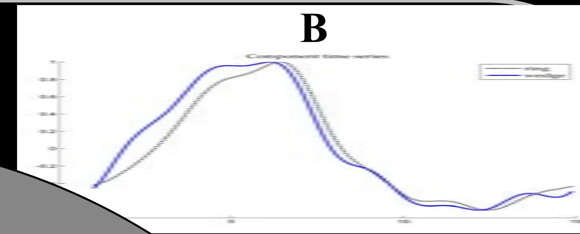


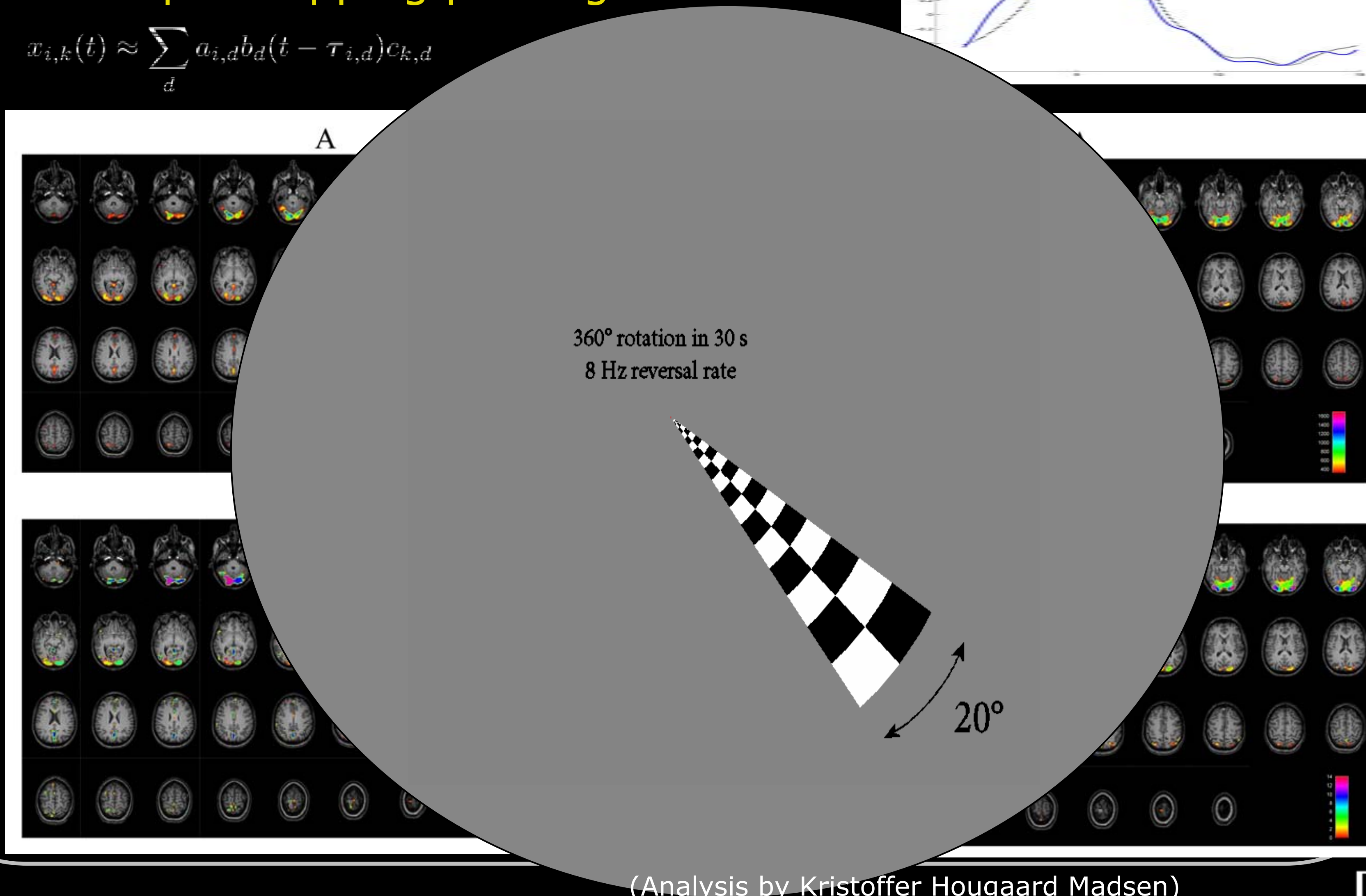


(Mørup et al., NeuroImage 2008)

Delay modelling of fMRI data from retinotopic mapping paradigm

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



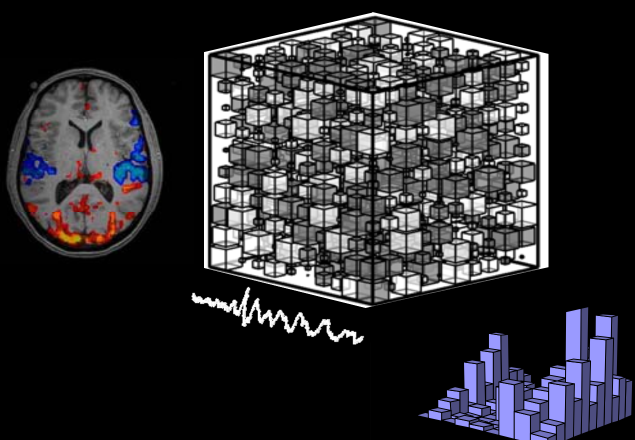


(Analysis by Kristoffer Hougaard Madsen)

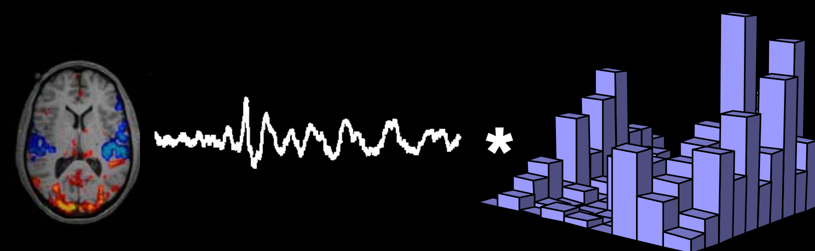
Modeling Delay and Shape Variability

convolutive CP:

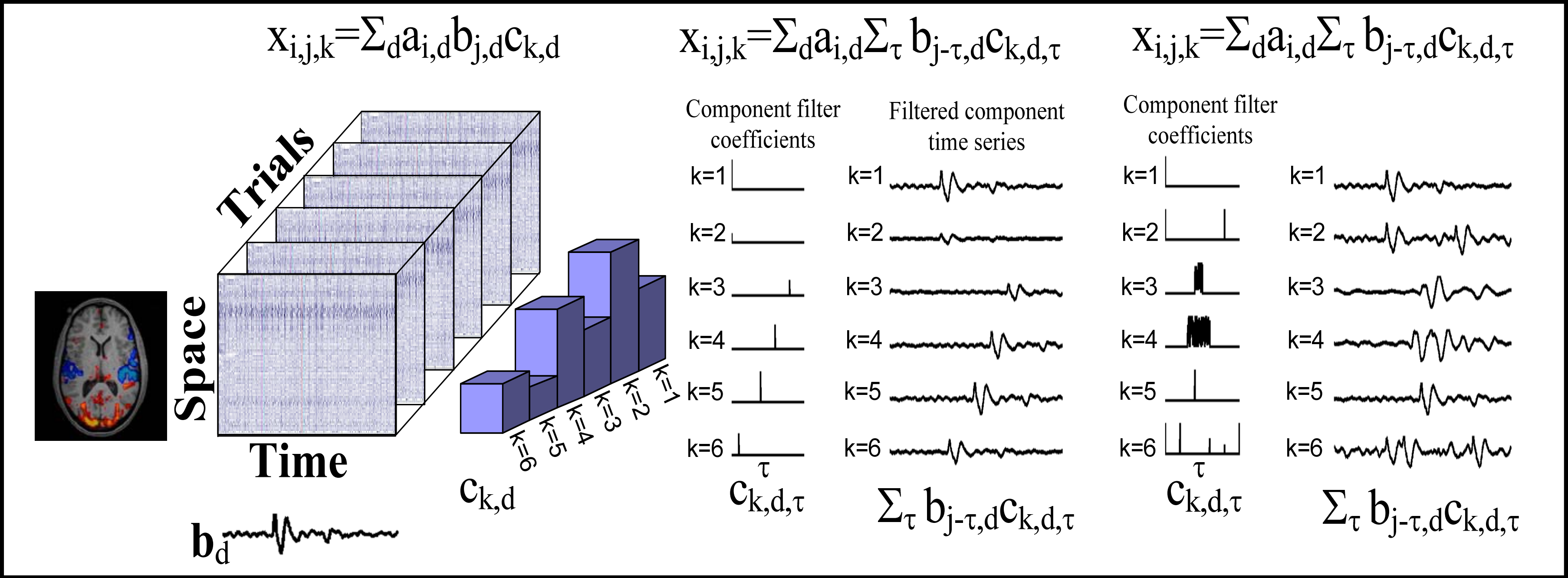
$$x_{i,k}(t) \approx \sum_{d,\tau} a_{i,d} b_d(t-\tau) c_{k,d}(\tau)$$



$$\sum_{d}$$

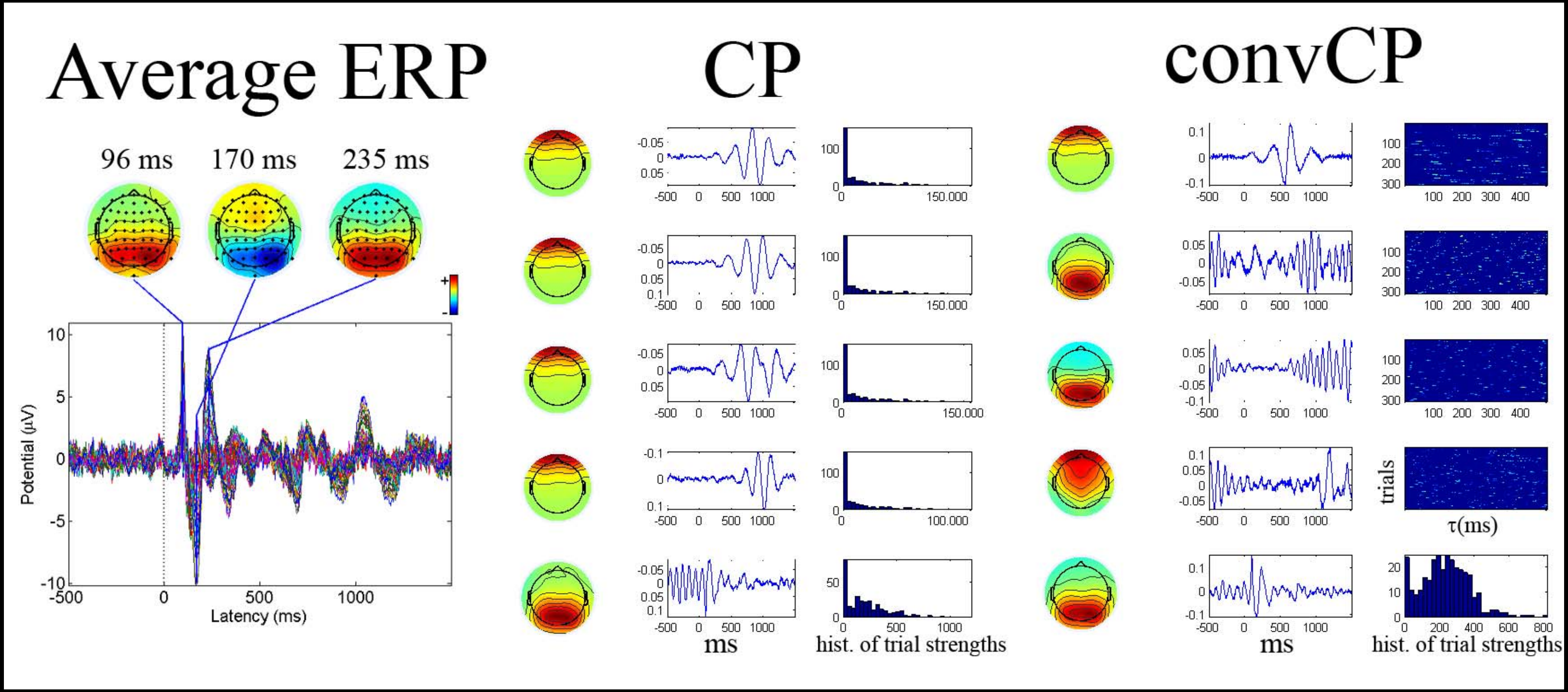


CP, ShiftCP and ConvCP

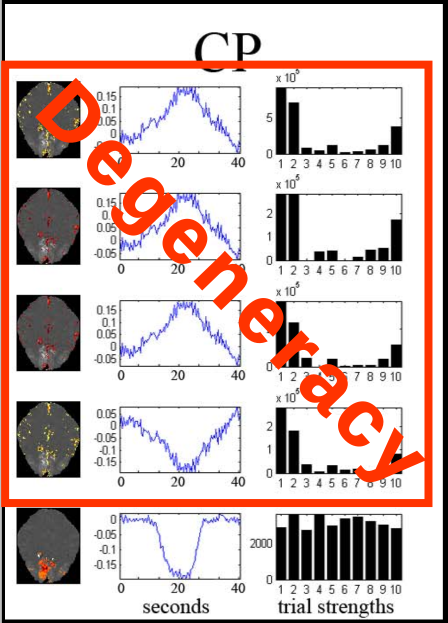


ConvCP: Can model arbitrary number of component delays within the trials and account for shape variation within the convolutional model representation

Convolutive Multi-linear decomposition



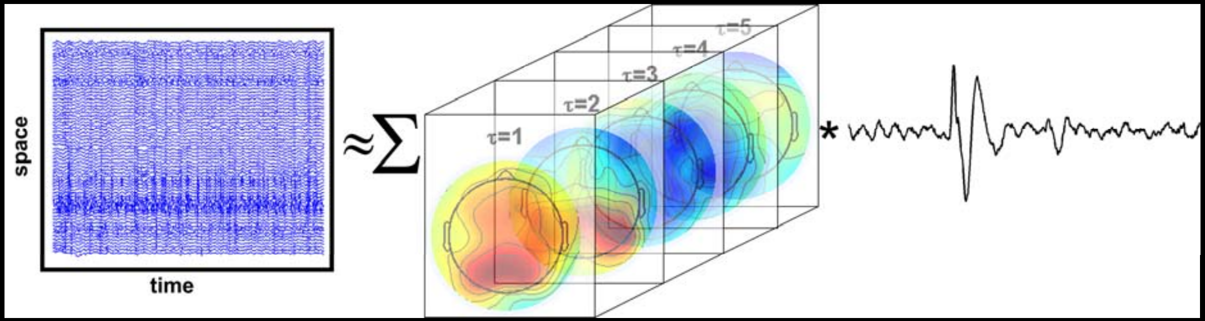
Analysis of fMRI data



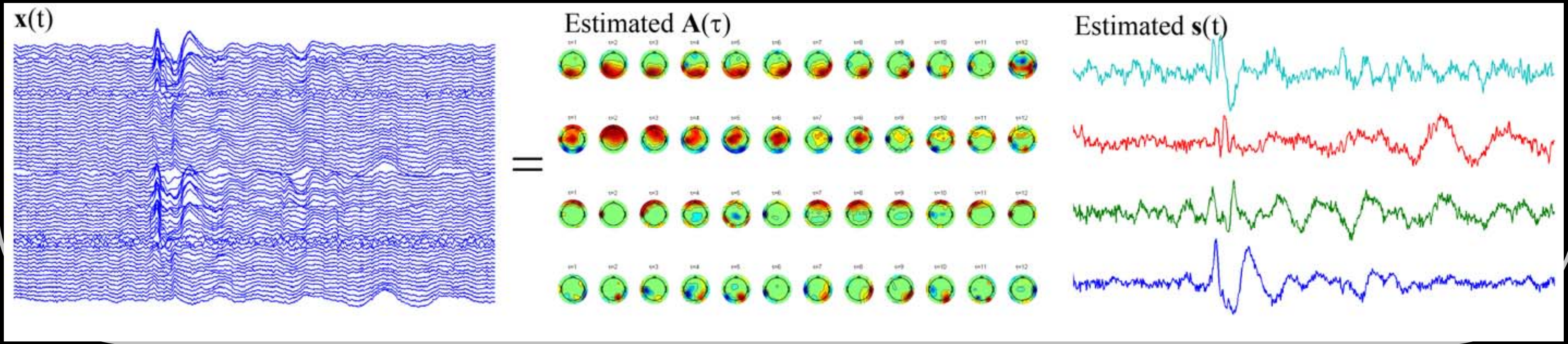
Each trial consists of a visual stimulus delivered as an annular full-field checkerboard reversing at 8 Hz.

Shape and delay modelling also relevant for bi-linear decomposition:
Convolutive Bilinear decomposition

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

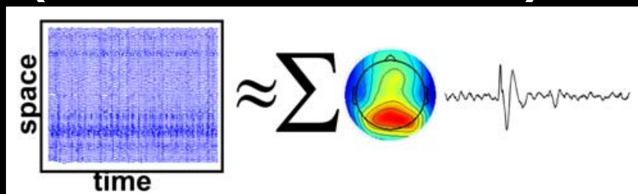


In fact the above model can be interpreted as a latent causal modelling framework

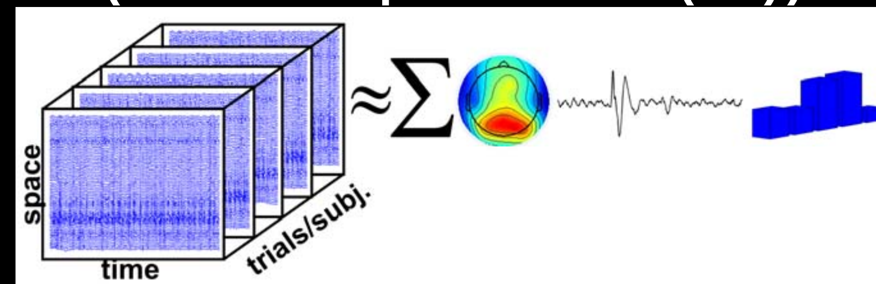


Summary of the ”tour de models”

Bi-linear modelling (ICA / SVD / PCA / NMF)

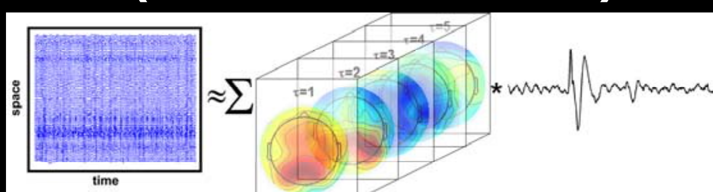


Multi-linear modelling (CandeComp / PARAFAC (CP))

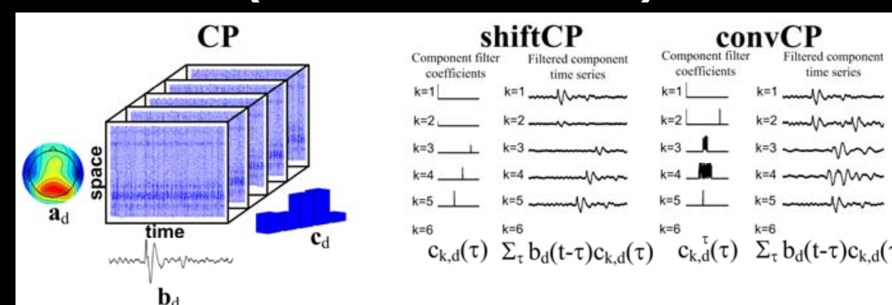


Extensions to model delay and shape changes

Convolutive Bi-linear modelling (convICA / convNMF)



Convolutive multi-linear modelling (shiftCP / convCP)



AIM of analysis

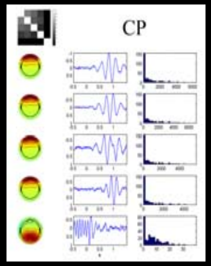
- Extract an efficient internal representation of the statistical structure implicit in the data
- Drive novel hypothesis for formal testing on validation data sets

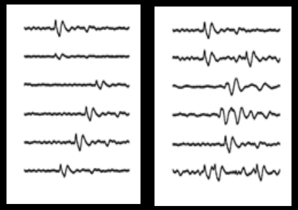
Conclusion

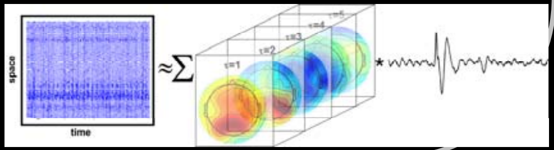
- Unsupervised learning is an important framework for multivariate analysis of neuroimaging data such as fMRI
- Bi-linear analysis ambiguous requiring additional assumption such as independence or sparsity (forming ICA and Sparse coding)

Conclusion

- Multi-linear modeling offers the ability to extract the consistent activity of neuroimaging data over repeats/subjects/conditions etc.
- However, violation of multi-linearity due to variability causes degeneracy
- Common causes of variability in neuroimaging data are delay and shape variation
- Advancing the CP model to ShiftCP and ConvCP enables to address these types of variability.
- Modelling delay and shape changes is also relevant for bi-linear modelling and open doorways to address latent causal relations.







Further reading

Anders H. Andersen and William S. Rayens. Structure-seeking multilinear methods for the analysis of fmri data. *NeuroImage*, 22:728–739, 2004.

M. Dyrholm, S. Makeig, and L. K. Hansen. Model structure selection in convolutive mixtures. In *6th International Conference on Independent Component Analysis and Blind Source Separation*, 2006.

Scott Makeig, Anthony J. Bell, Tzyy-Ping Jung, and Terrence J. Sejnowski. Independent component analysis of electroencephalographic data. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 145–151. The MIT Press, 1996.

M. Mørup, L. K. Hansen, C. S. Hermann, J. Parnas, and S. M. Arnfred. Parallel factor analysis as an exploratory tool for wavelet transformed event-related eeg. *NeuroImage*, 29(3):938–947, 2006.

M. Mørup, L.K. Hansen, and S. M. Arnfred. Erpwavelab a toolbox for multi-channel analysis of time-frequency transformed event related potentials. *Journal of Neuroscience Methods*, 161(361-368), 2007.

M. Mørup, L.K. Hansen, and S. M. Arnfred. Algorithms for sparse non-negative tucker. *Neural Computation*, 20(8):2112–2131, 2008.

M. Mørup, L.K. Hansen, S.M. Arnfred, L.-H. Lim, and K.H. Madsen. Shift-invariant multilinear decomposition of neuroimaging data. *NeuroImage*, 42(4):1439–1450, 2008.

Morten Mørup and Lars Kai Hansen. Automatic relevance determination for multi-way models. *Journal of Chemometrics*, 23:352–363, 2009.