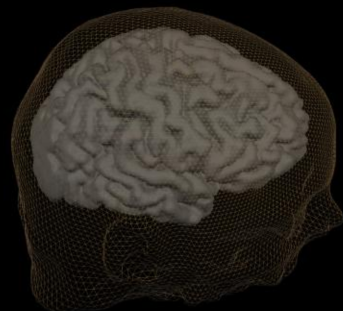
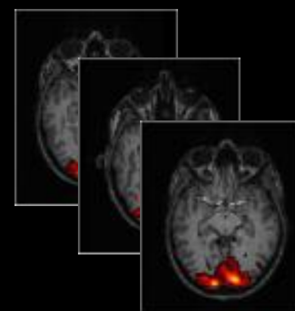
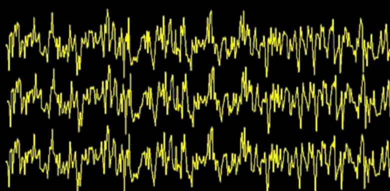
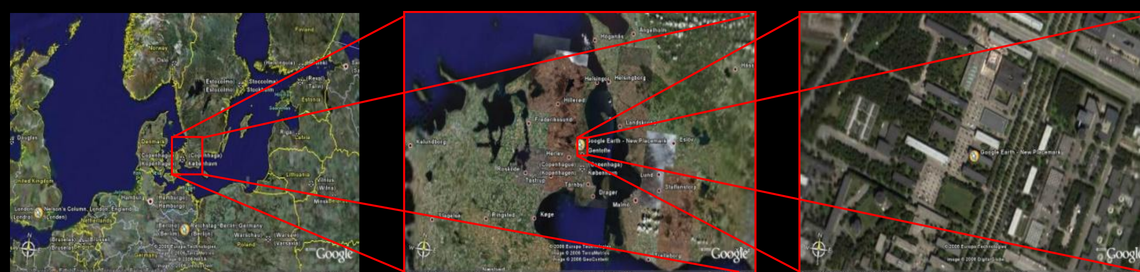
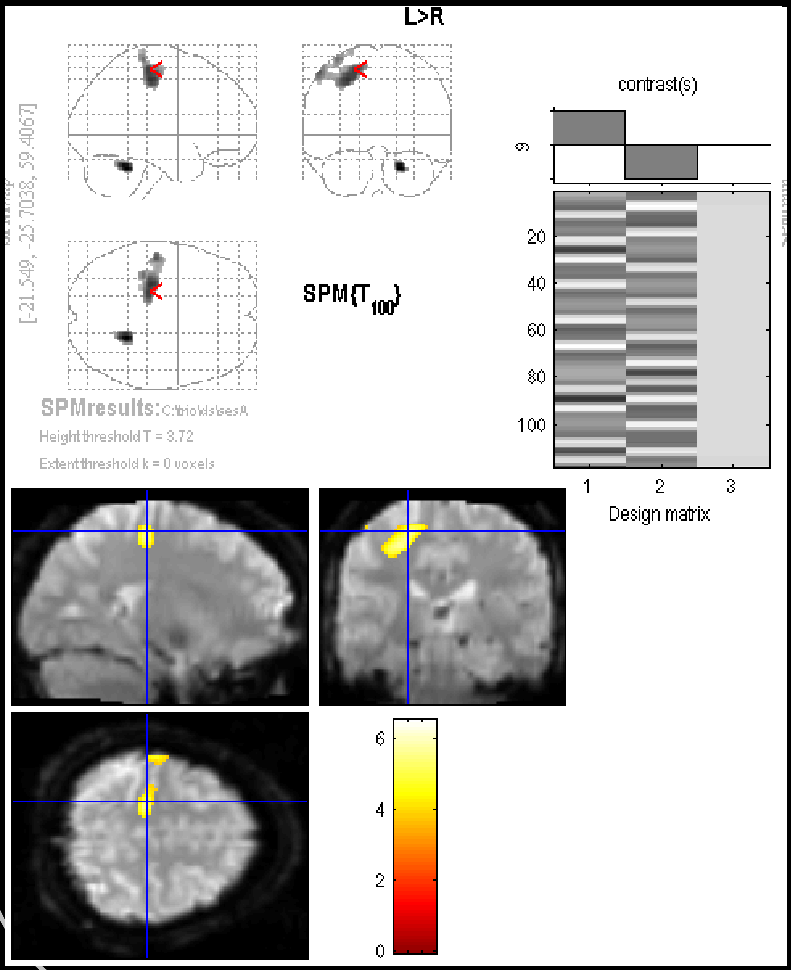
Shape and latency modeling of neuroimaging data

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



Problems:

1) Multiple comparisons, i.e. many voxels tested.

2) What is the true number of independent tests, i.e. voxels are highly correlated

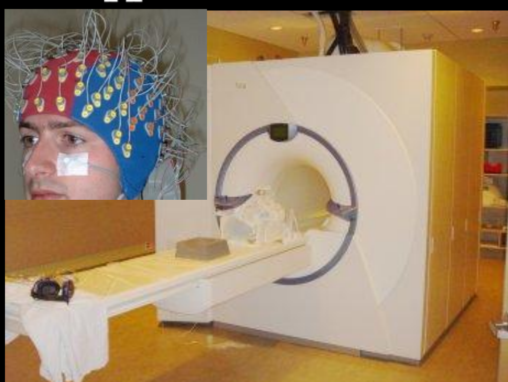
3) Data extremely noisy, i.e. low SNR rendering tests insignificant.

⇓

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

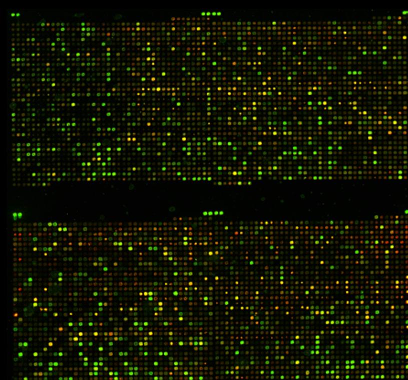
This problem is no different than the problems encountered in general in Modern Massive Datasets (MMDS)

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



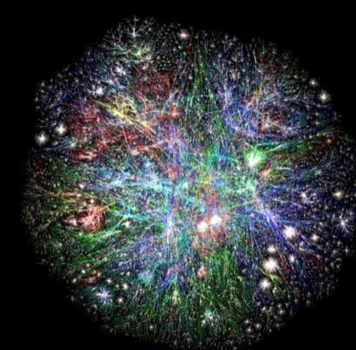
NeuroInformatics

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



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

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

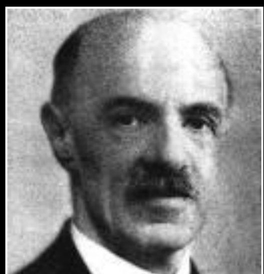
THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental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

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

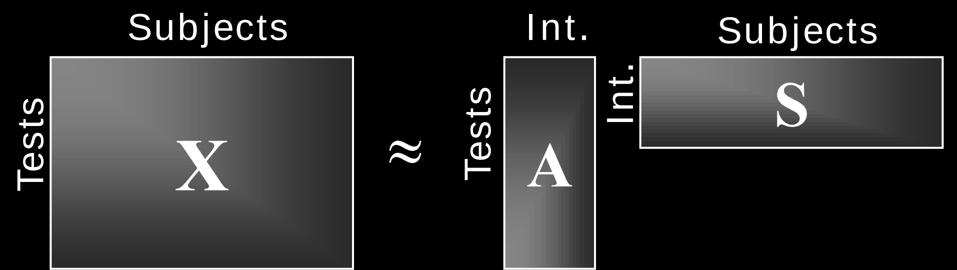
So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have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

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)

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

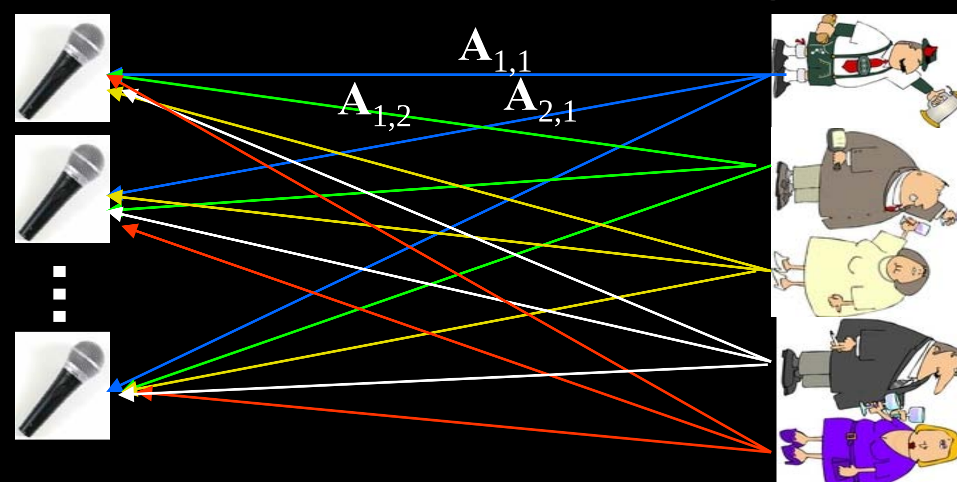
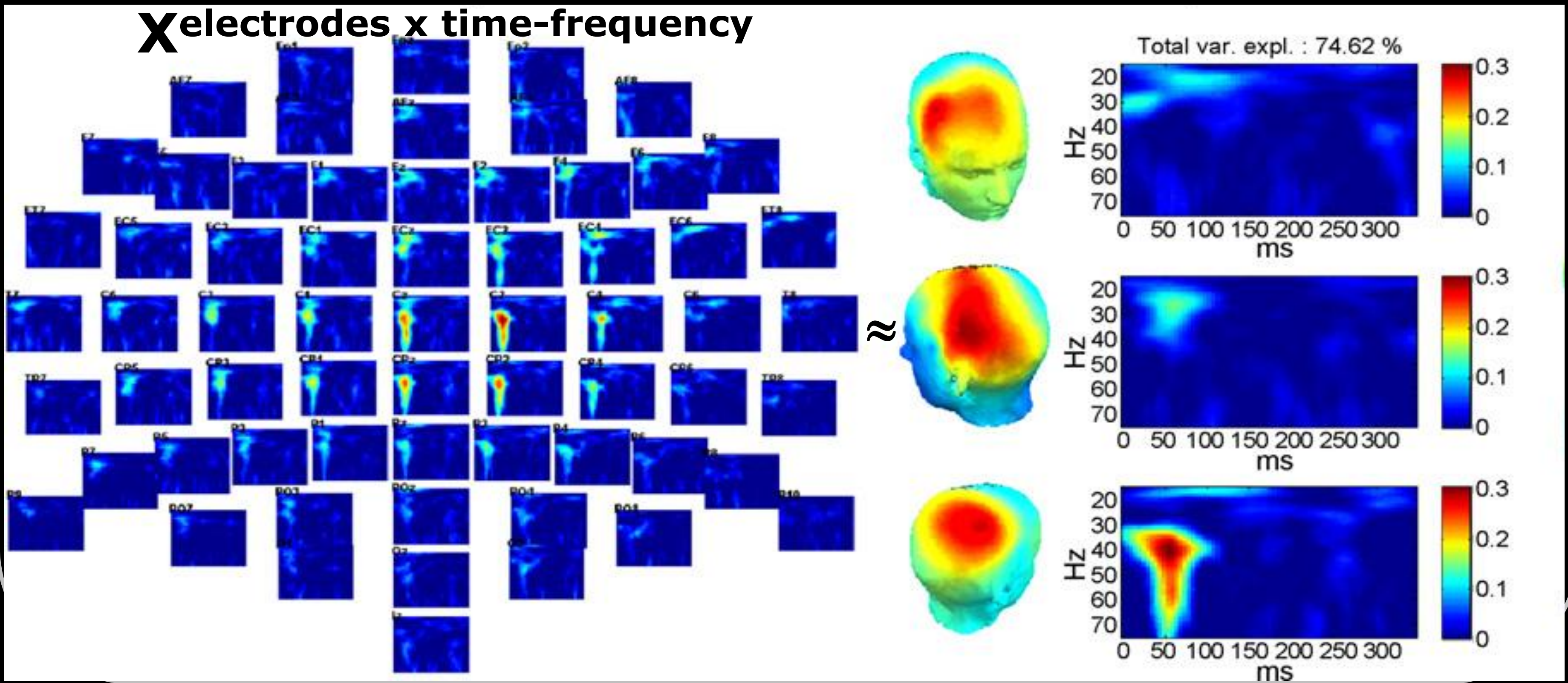
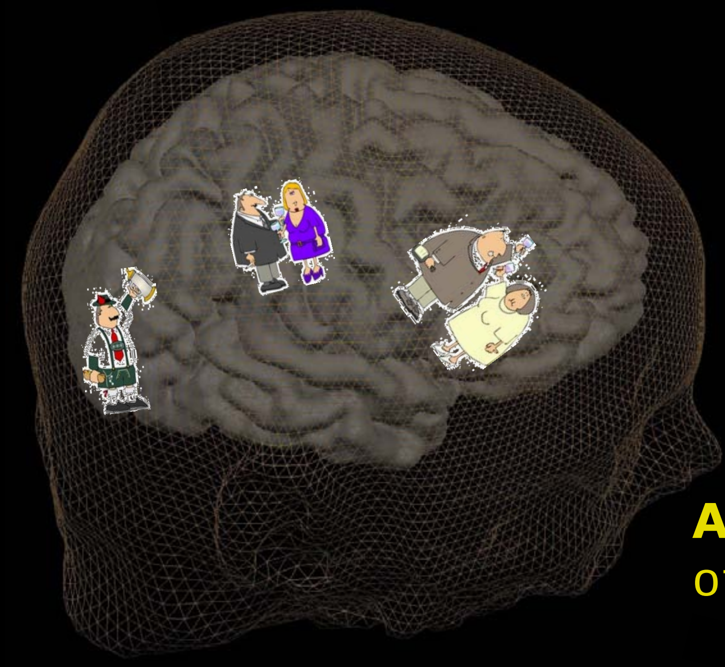


Illustration of Factor Analysis on frequency transformed EEG

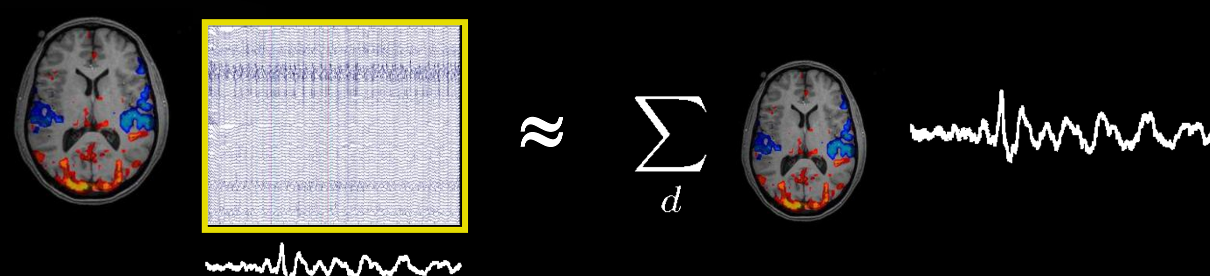


The EEG/MEG/fMRI Party problem



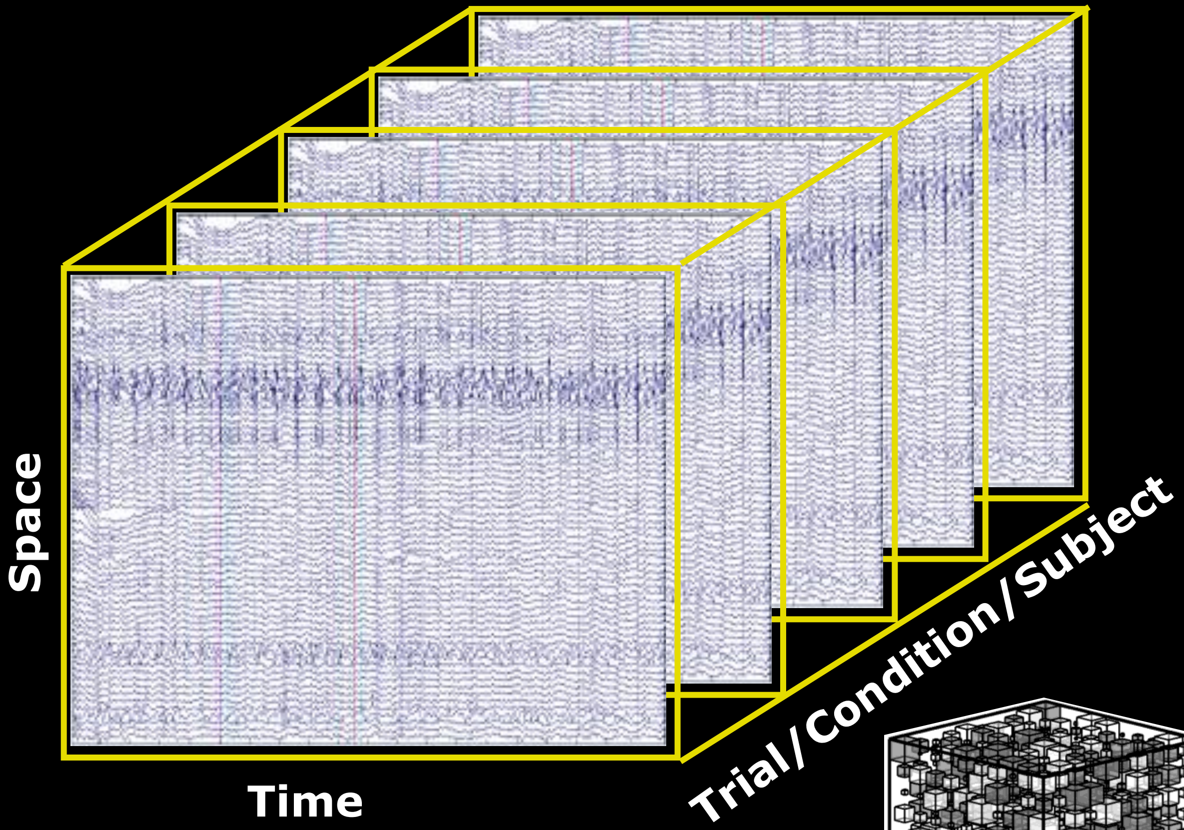
$$\mathbf{X}^{\text{Voxel}\times\text{Time}} \approx \sum_d \mathbf{a}_d^{Voxel}\mathbf{b}_d^{\text{Time}}$$

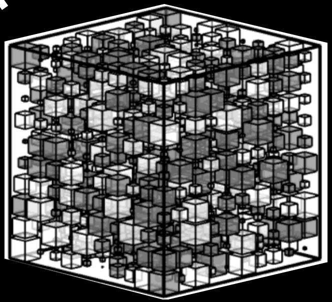
Assumption: Data **instantaneous** mixture of temporal signatures. (PCA/ICA/NMF)



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

From 2-way to multi-way analysis





Multi-subject analysis

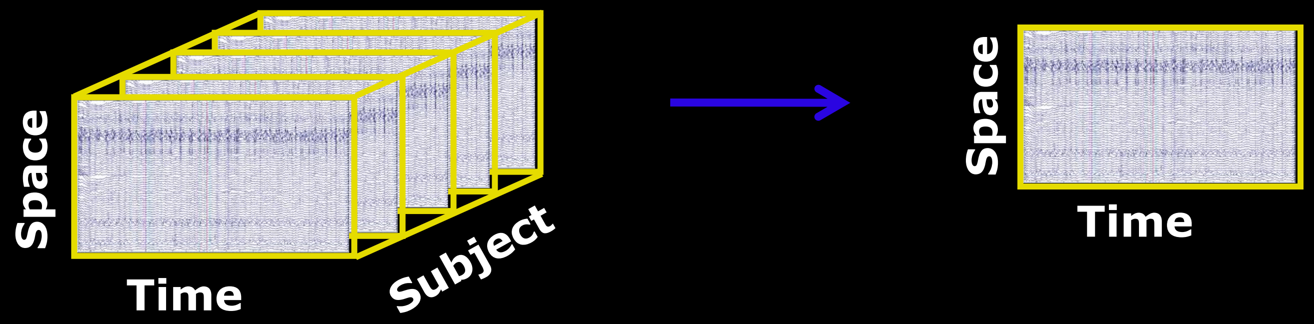
At least four possibilities:

- Pre-average data
- Separate analysis
- Data concatenation
- **Tensor/multi-way models**

Pre-averaging

Simply average data over subjects prior to analysis

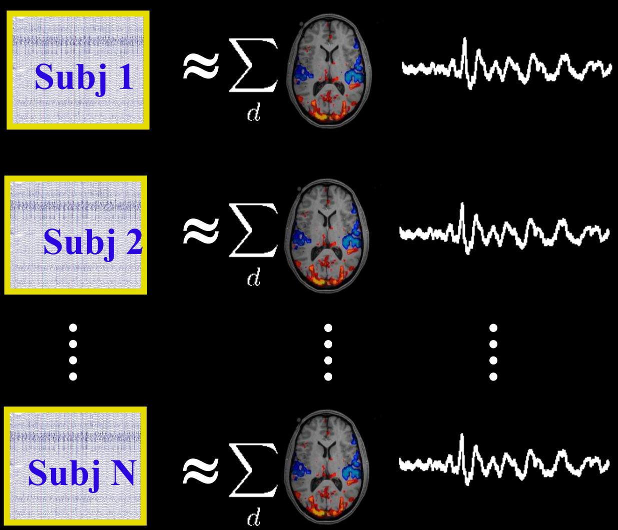
- Common spatial profiles
- Common time profiles
- Model must generalise in both space and time over subjects



Separate analysis

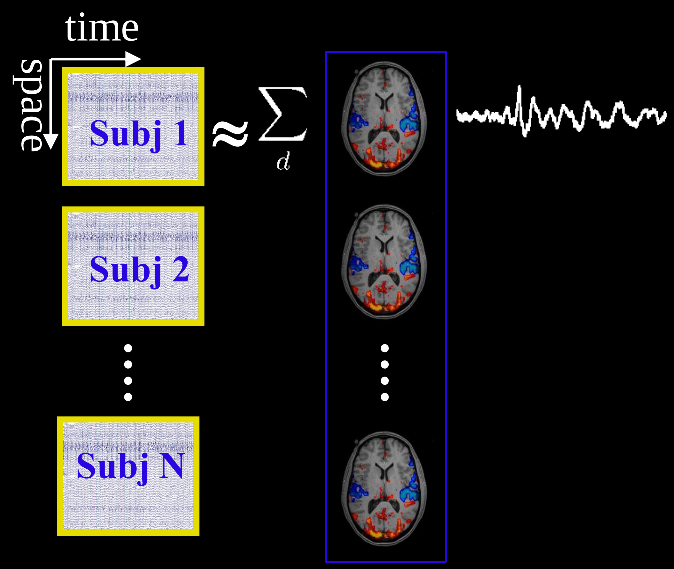
Run analysis separately for each subject

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

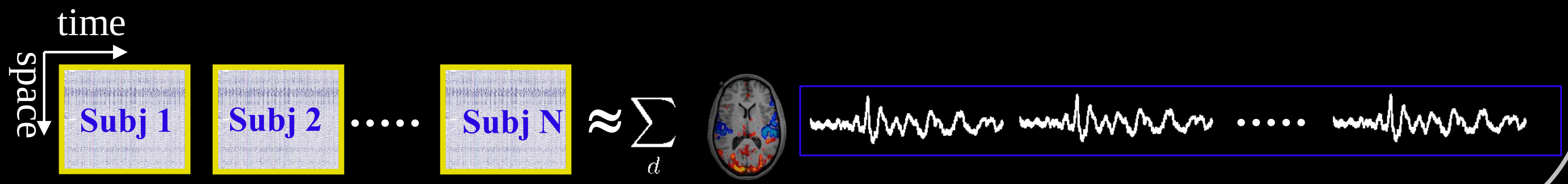


Concatenation of multi-way data to 2-way

(identical time series varying spatial maps)

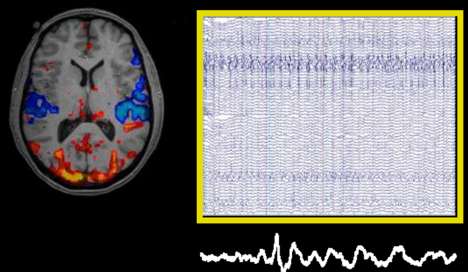


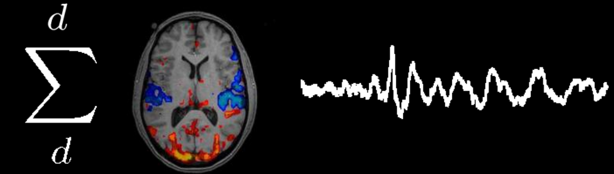
(identical spatial map, varying time series)



Multilinear modelling

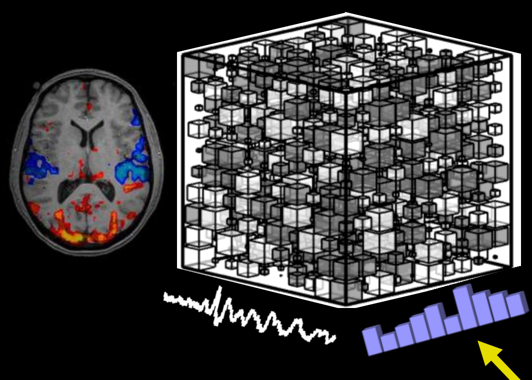
Bilinear Model: $\mathbf{X}^{\mathrm{Voxel}\times\mathrm{Time}} \approx \sum_d \mathbf{a}_d^{Voxel}\mathbf{b}_d^{\mathrm{Time}}$





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

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

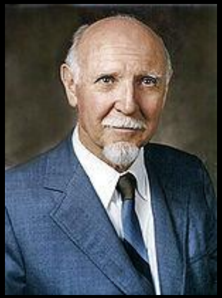


Assumption: Data **instantaneous** mixture of temporal signatures that are expressed to **various degree** over the Subjects/trials
(Canoncial Decomposition, Parallel Factor (CP))

(weighted averages over the trials)

History of multi-way decomposition

- **Hitchcock 1927 (not the filmmaker!)**
  Generalized 2-way rank to n-way (i.e. proposed the CP-model,) as well as introduced the notion of n-mode rank
- **Cattell 1944**
  Parallel Proportional Profiles (to resolve rotational indeterminacy in factor analysis)
- **Harshman and Carrol & Chang 1970**
  Independently proposed the PARAFAC and CanDecomp models (CP model, see later slides)



Cattell: Also very famous for 16 personality factor model and the 16PF Questionnaire

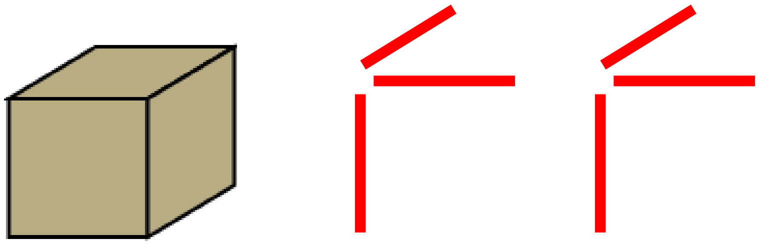
Harshman

Carrol

Many ways of writing the CP model

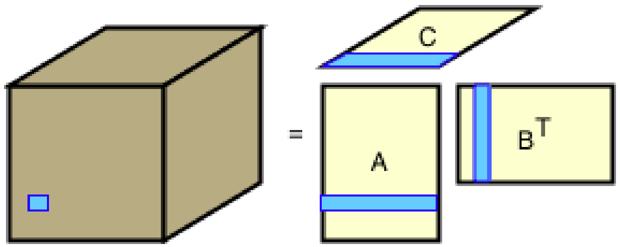
- **Outer product form**

$$\mathcal{X} \approx \sum_{i=1}^{r} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$$



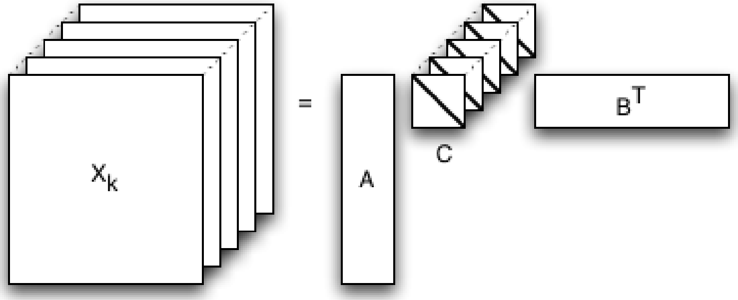
- **Scalar form**

$$x_{ijk} \approx \sum_{i=1}^{r} a_{ir} b_{jr} c_{kr}$$



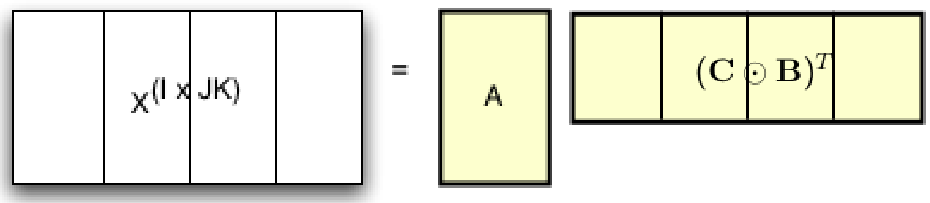
- **Tensor slice form**

$$\mathbf{X}_k \approx \mathbf{A}\ \mathrm{diag}(\mathbf{c}_{k:})\mathbf{B}^T$$



- **Matrix form**

$$\mathbf{X}_{(1)} \approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T$$



Bilinear decomposition not unique

$$\mathbf{X} \approx \mathbf{A}\mathbf{B}^\top = \mathbf{A}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{B}^\top = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}^\top$$

Multi-linear decomposition is in general unique!!

$$\begin{aligned} \boldsymbol{X}_{(:,:,k)} \approx \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{C}_{k,:})\boldsymbol{B}^T &= (\boldsymbol{A}\boldsymbol{T})\underbrace{(\boldsymbol{T}^{-1}\,\mathrm{diag}(\boldsymbol{C}_{k,:})\boldsymbol{Q})}(\boldsymbol{Q}^{-1}\boldsymbol{B}^T) \\ &= \widehat{\boldsymbol{A}}\,\mathrm{diag}(\widehat{\boldsymbol{C}}_{k,:})\widehat{\boldsymbol{B}}^T. \end{aligned}$$

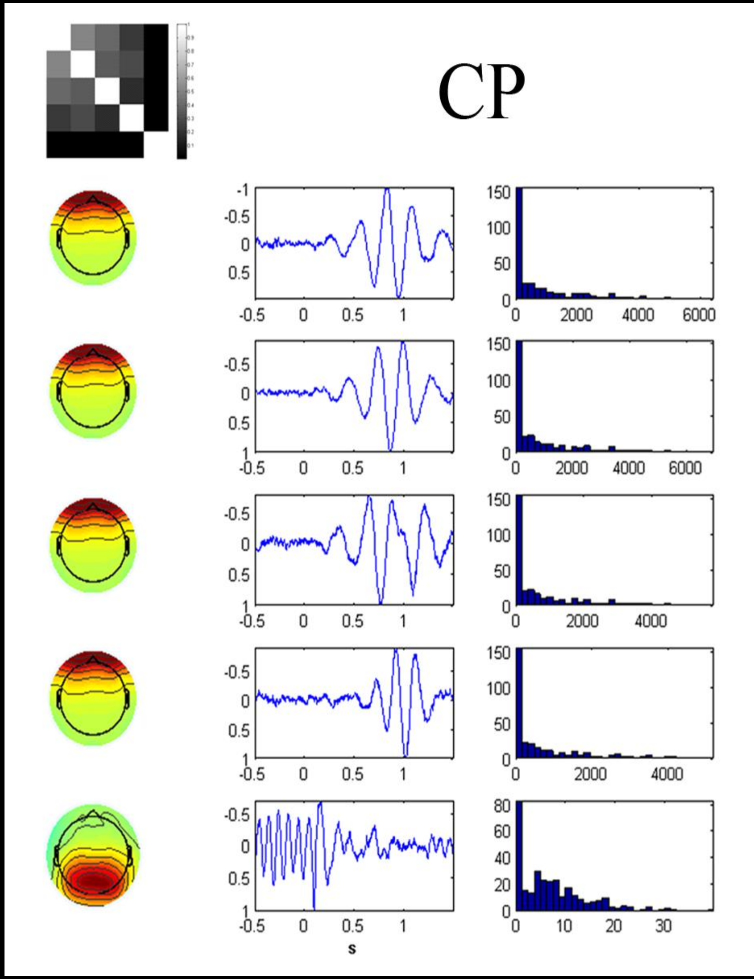
Kruskal (1976, 1977) derived the following uniqueness criterion generalized to N-ways arrays in (Sidiropoulos and Bro, 2000):

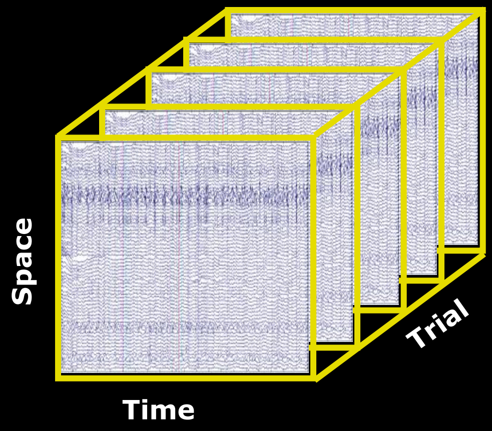
$$\begin{aligned} &\textit{3-way array:} \quad k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2D + 2 \\ &\textit{N-way array:} \quad \sum_n k_{\mathbf{A}^{(n)}} \geq 2D + N - 1 \end{aligned}$$

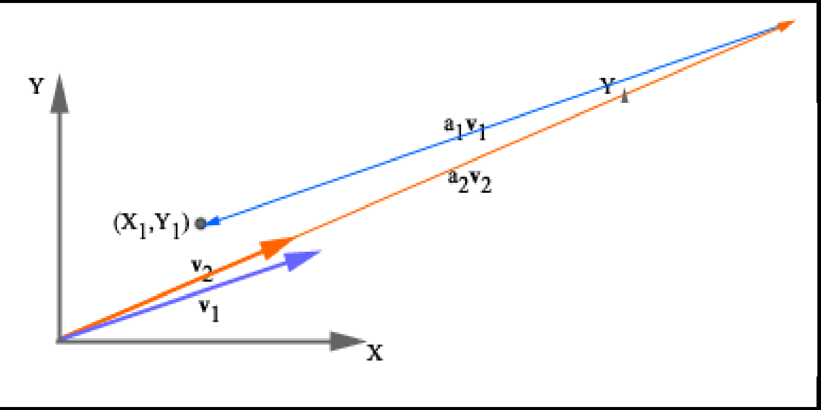
where $k_{\mathbf{A}}$ is the k-rank denoting the smallest subset of columns of $\mathbf{A}$ that is guaranteed to be linearly independent. Thus, $k_{\mathbf{A}} \leq rank(\mathbf{A})$.

"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

Unfortunately, Violation of multi-linearity causes degeneracy

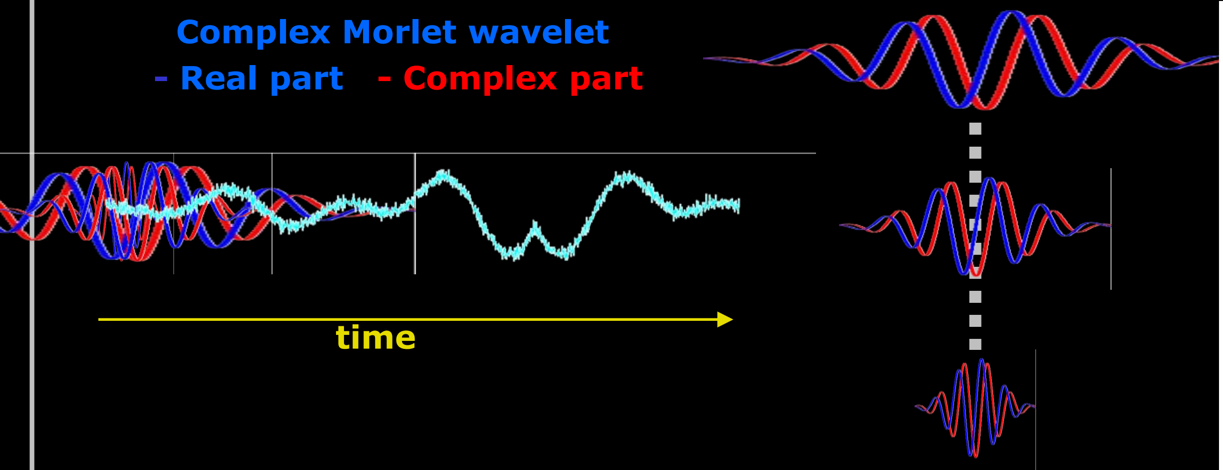


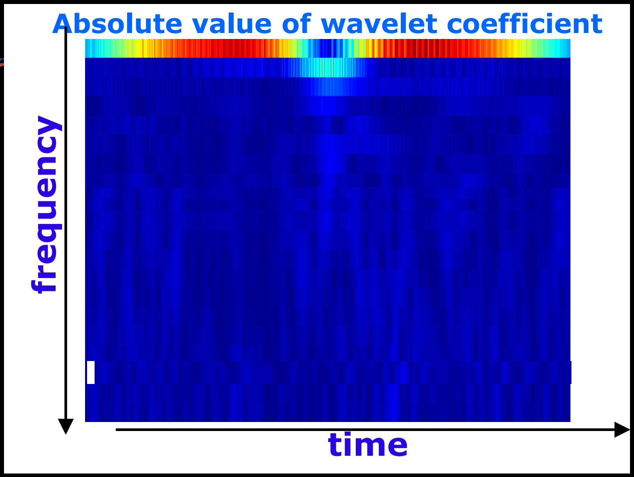


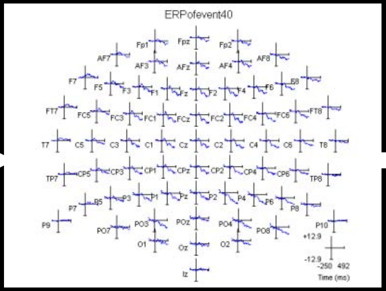


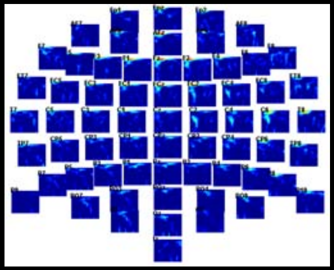
Common Fixes: Impose orthogonality, reguarlization or non-negativey constraints by analyzing Data transformed to a time-frequency domain representation

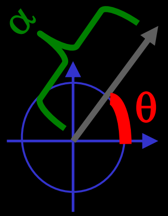
Time-frequency representation of EEG through wavelet transformation





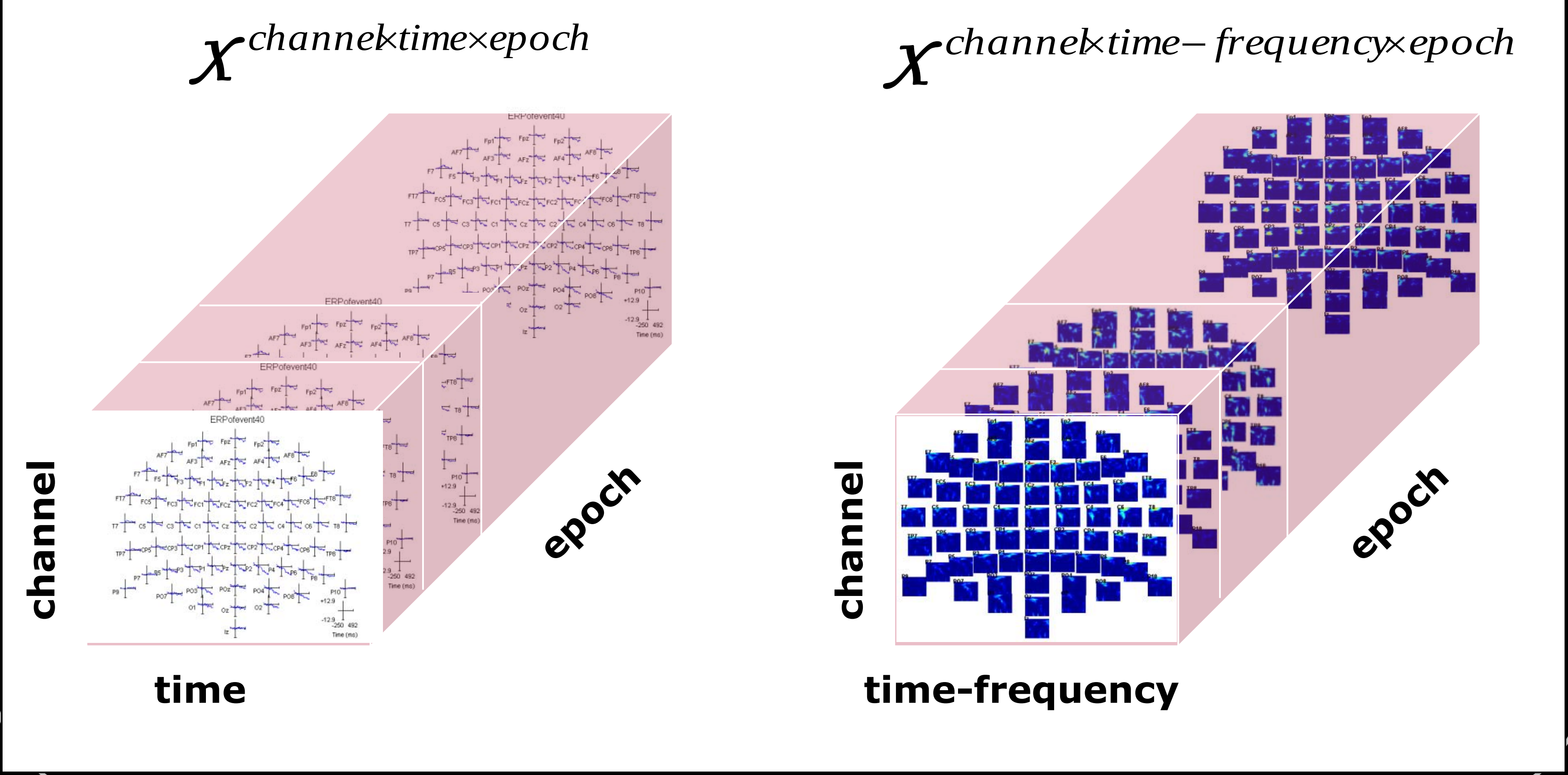






Captures frequency changes through time

Wavelet transformed event related data



Measures of the event related ERP in the time-frequency domain

$$ERPS(c,f,t) = \frac{1}{N}\sum_{n}^{N} |X(c,f,t,n)|^2 \tag{3}$$

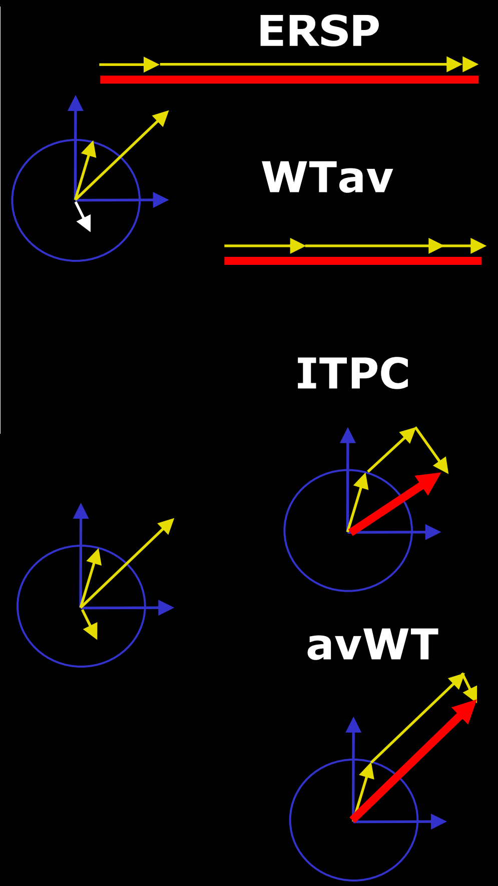
$$WTav(c,f,t) = \frac{1}{N}\sum_{n}^{N} |X(c,f,t,n)| \tag{4}$$

While the ERSP is a measure of the average power over epochs at given channel-frequency-time points the WTav is the average amplitude of the oscillation.

$$ITPC(c,f,t) = \frac{1}{N}\sum_{n}^{N} \frac{X(c,f,t,n)}{|X(c,f,t,n)|} \tag{5}$$

$$avWT(c,f,t) = \frac{1}{N}\sum_{n}^{N} X(c,f,t,n) \tag{6}$$

While the amplitude of the ITPC also named the phase locking index measures the phase consistency over epochs, the avWT corresponds to the wavelet transformed Evoked Potential (EP).



Measures of the event related ERP in the time-frequency domain (cont.)

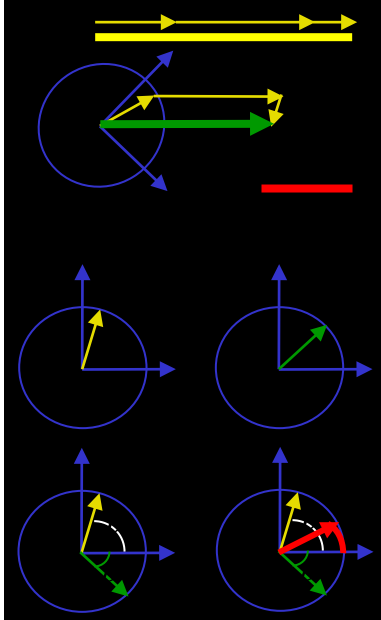
From the WTav and avWT the induced activity, i.e. everything that is not phase locked to the event can be estimated as

$$INDUCED(c,f,t) = WTav(c,f,t) - |avWT(c,f,t)| \tag{7}$$

Finally, the evoked response phase coherence (ERPCOH), i.e. how consistent the phase of a given oscillatory activity at channel c', frequency f' and time t' is to the activity at channel c, frequency f and time t, is given by:

$$ERPCOH_{c',f',t'}(c,f,t) = \frac{1}{N}\sum_{n}^{N} \frac{X(c,f,t,n)X^*(c',f',t',n)}{|X(c,f,t,n)||X(c',f',t',n)|} \tag{8}$$

where X^* denotes the complex conjugate.



www.ERPWAVELAB.org

Features:

Wavelet analysis

Data visualization

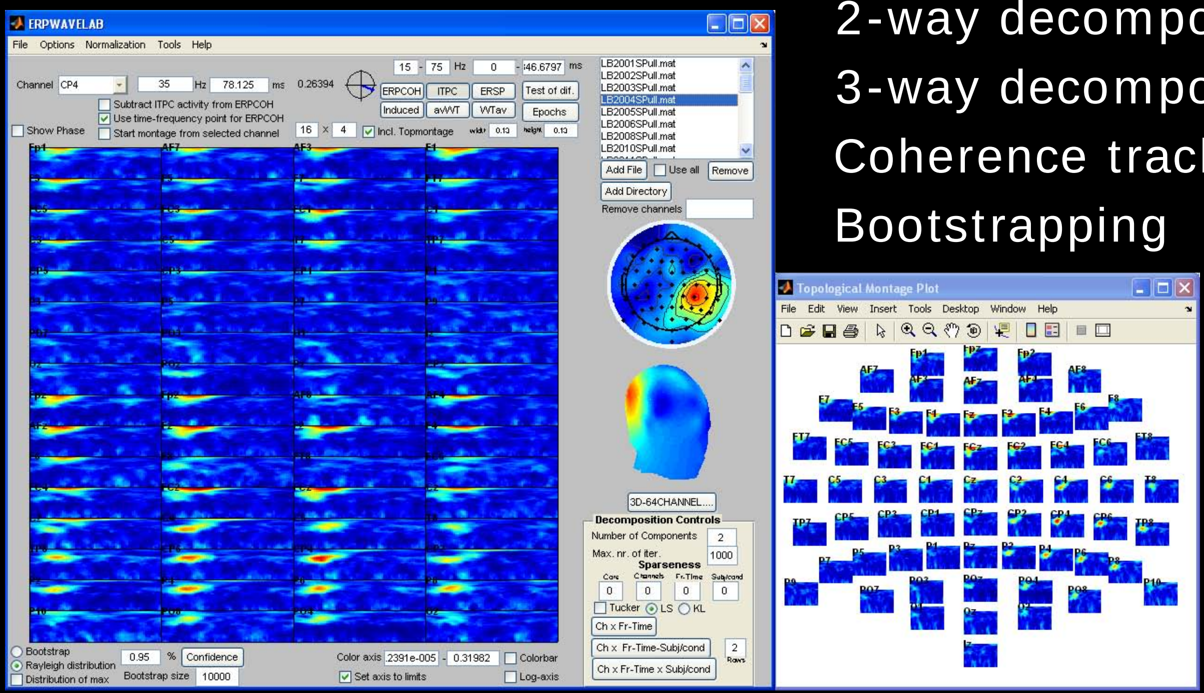
Artifact Rejection

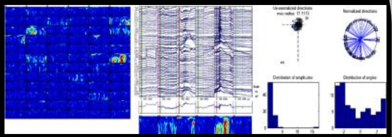
2-way decomposition

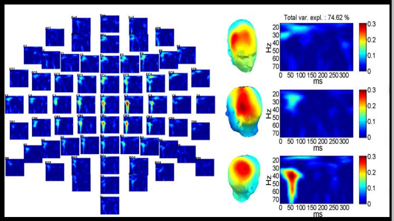
3-way decomposition

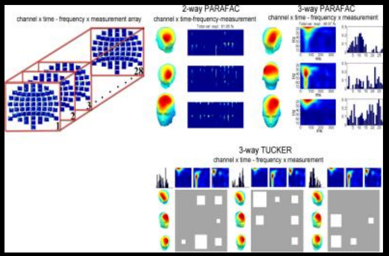
Coherence tracking

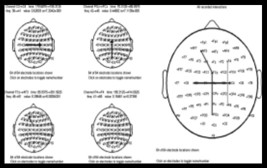
Bootstrapping

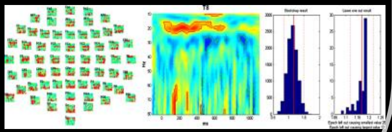






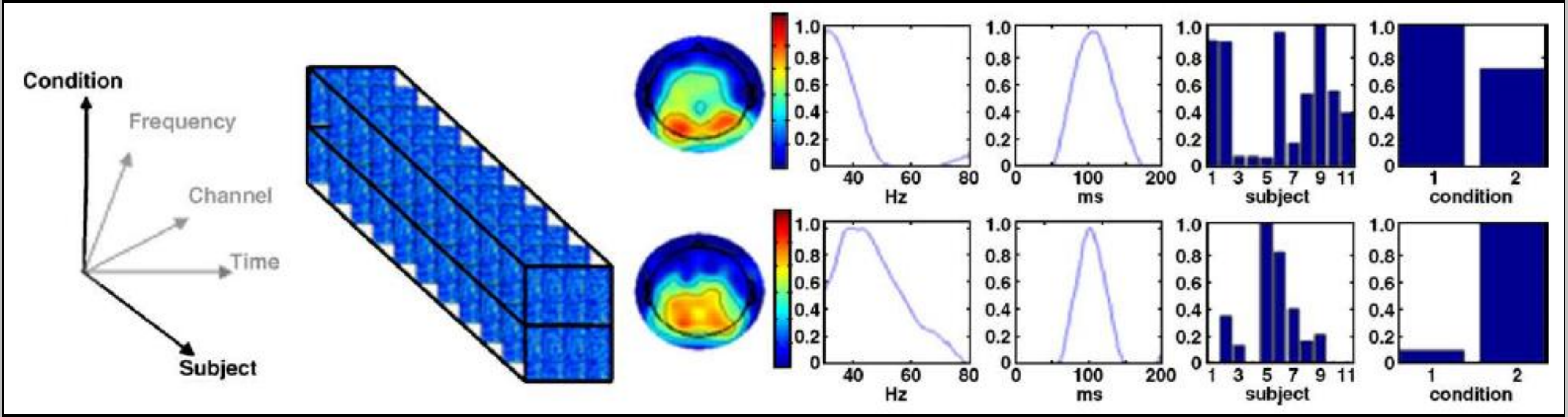






(Mørup et al, Journ. of Neurosc. Meth. 2007)

CP model extracts consistent activation allowing for subject/trial/condition dependent weights
(i.e. ”clever averaging”)



(Mørup et al., NeuroImage 2006)

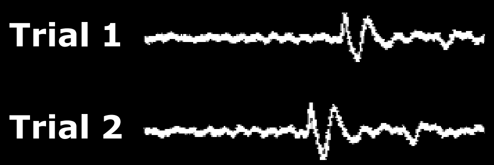
Degeneracy often a result of multi-linear models being to restrictive

Trilinear model can encompass:

- Variability in strength over repeats

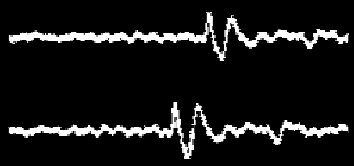
However, other common causes of variation are:

- Delay Variability



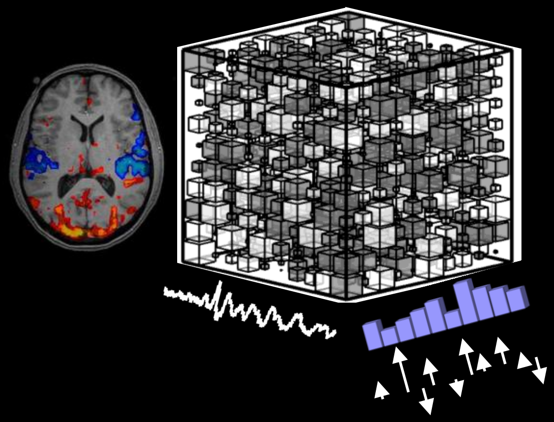
- Shape Variability

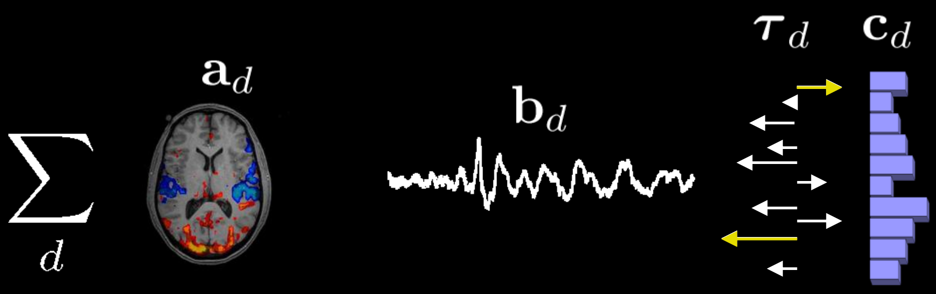
Modelling Delay Variability

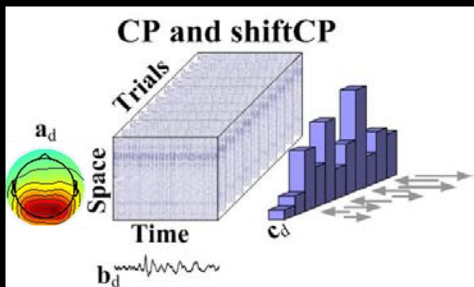


Shifted CP:

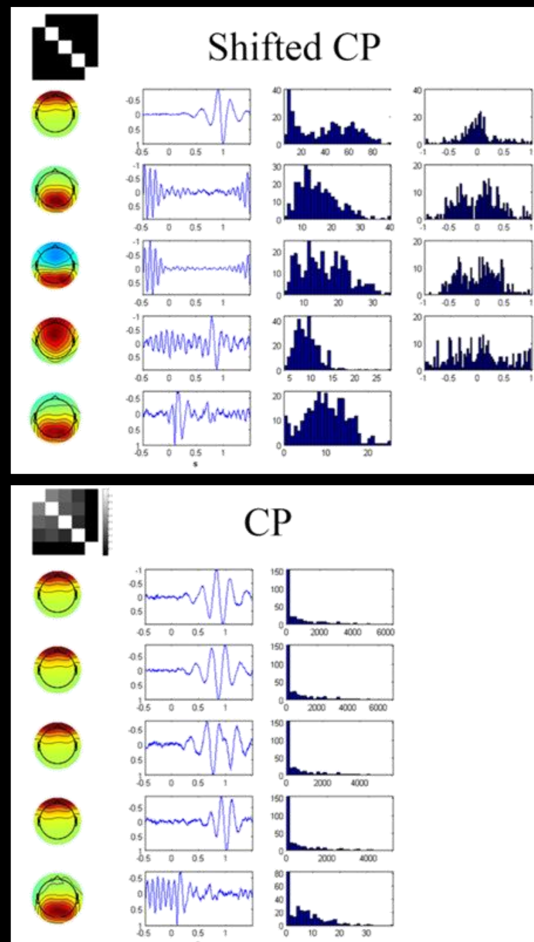
$$x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t - \boldsymbol{\tau}_{k,d}) c_{k,d}$$

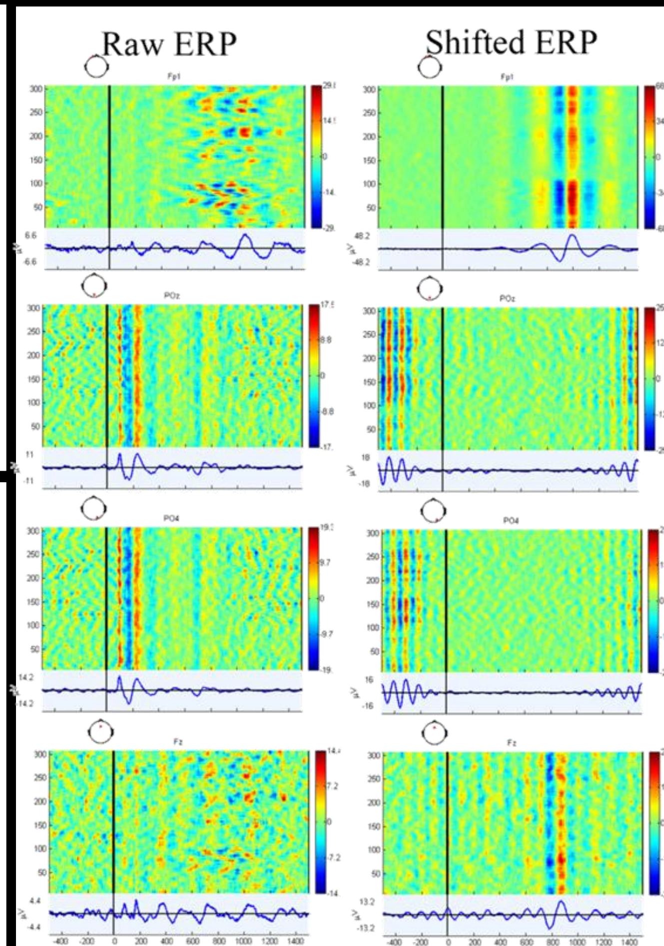




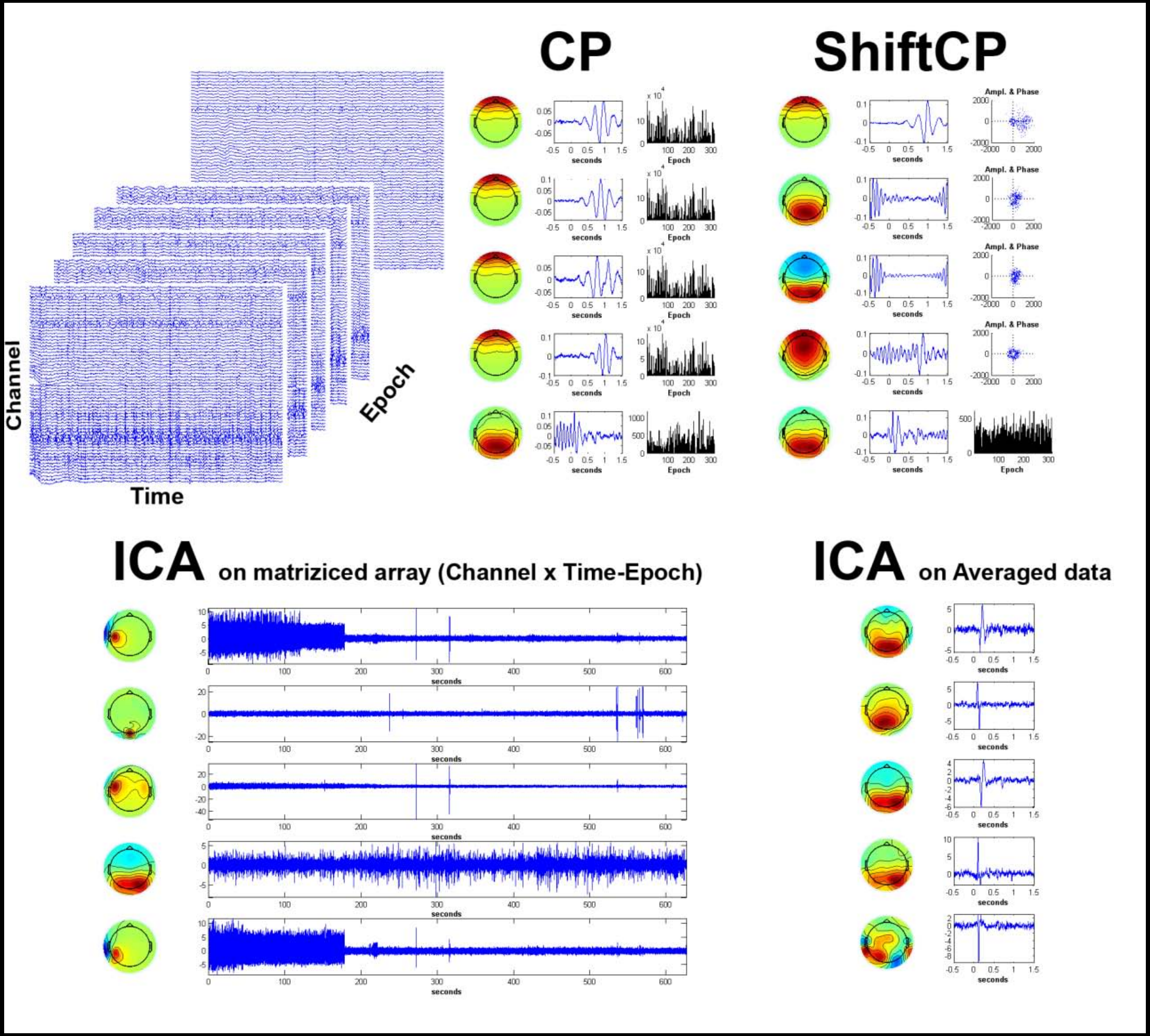


$$x_{i,k}(t) \approx \sum_d a_{i,d} b_d(t - \tau_{k,d}) c_{k,d}$$



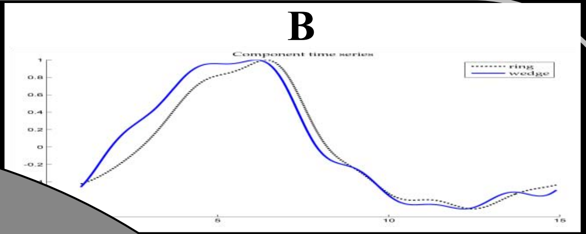


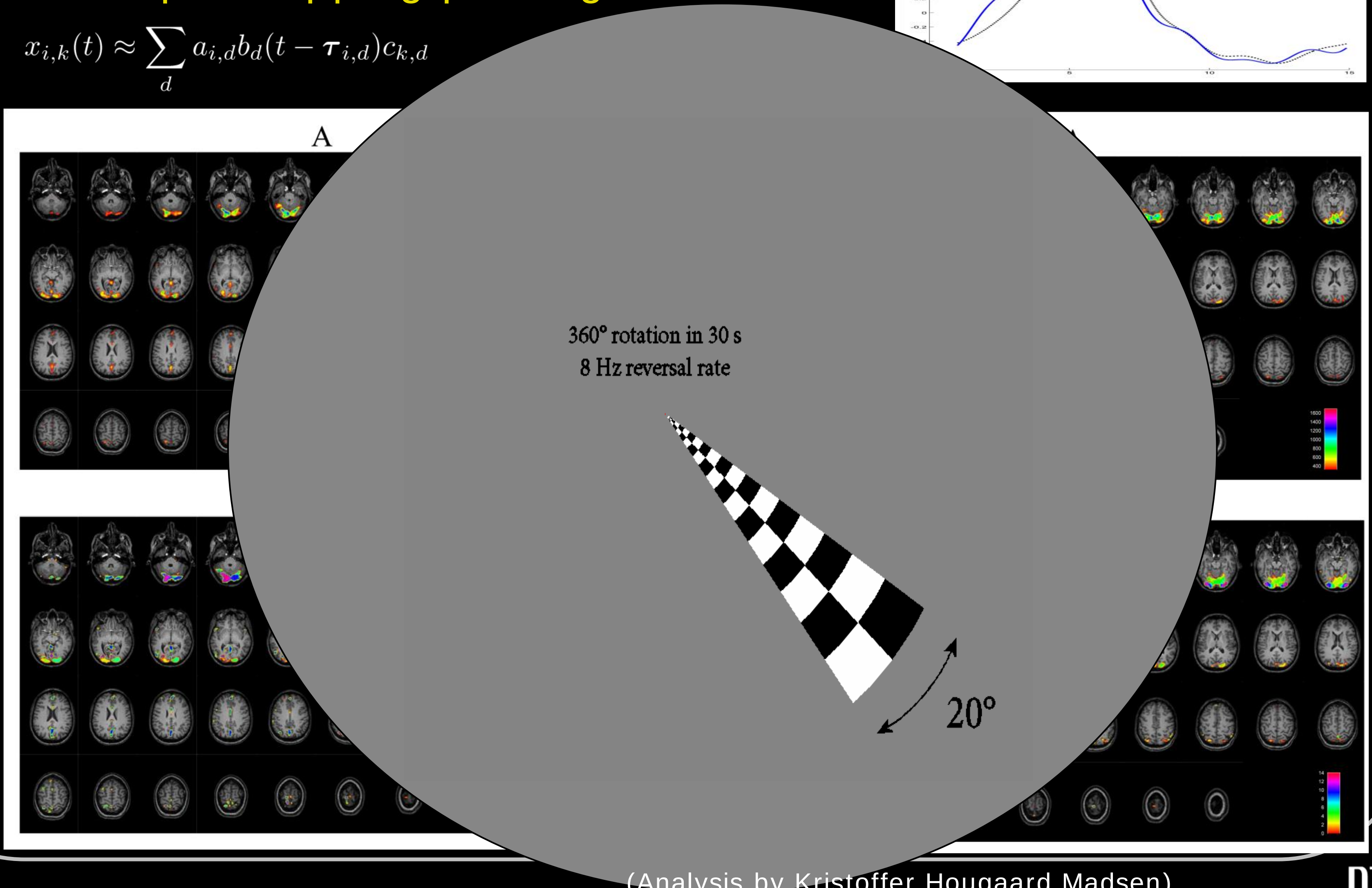
(Mørup et al.,
NeuroImage 2008)


CP
ShiftCP
Channel
Time
Epoch
ICA on matriziced array (Channel x Time-Epoch)
ICA on Averaged data
seconds
Epoch
Ampl. & Phase

Delay modelling of fMRI data from retinotopic mapping paradigm

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



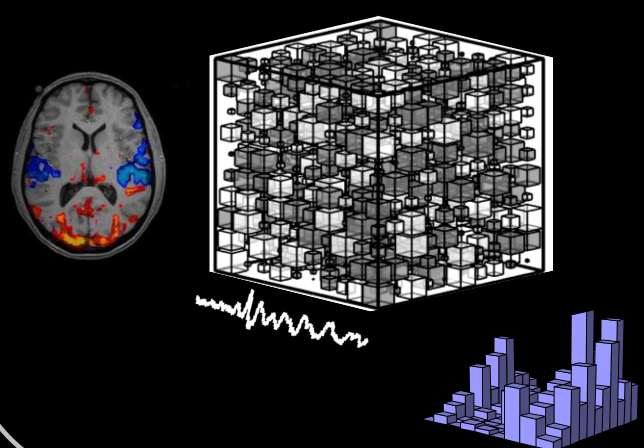


(Analysis by Kristoffer Hougaard Madsen)

Modeling Shape (and delay) Variability

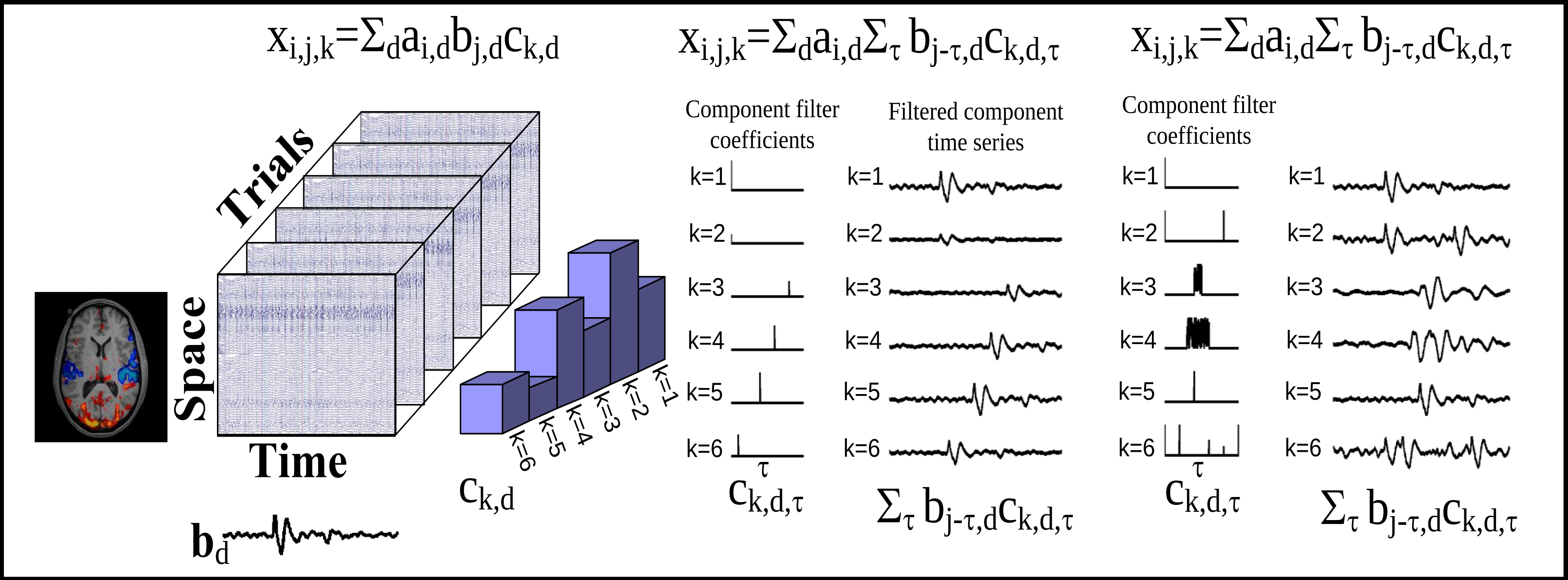
convolutive CP:

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



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

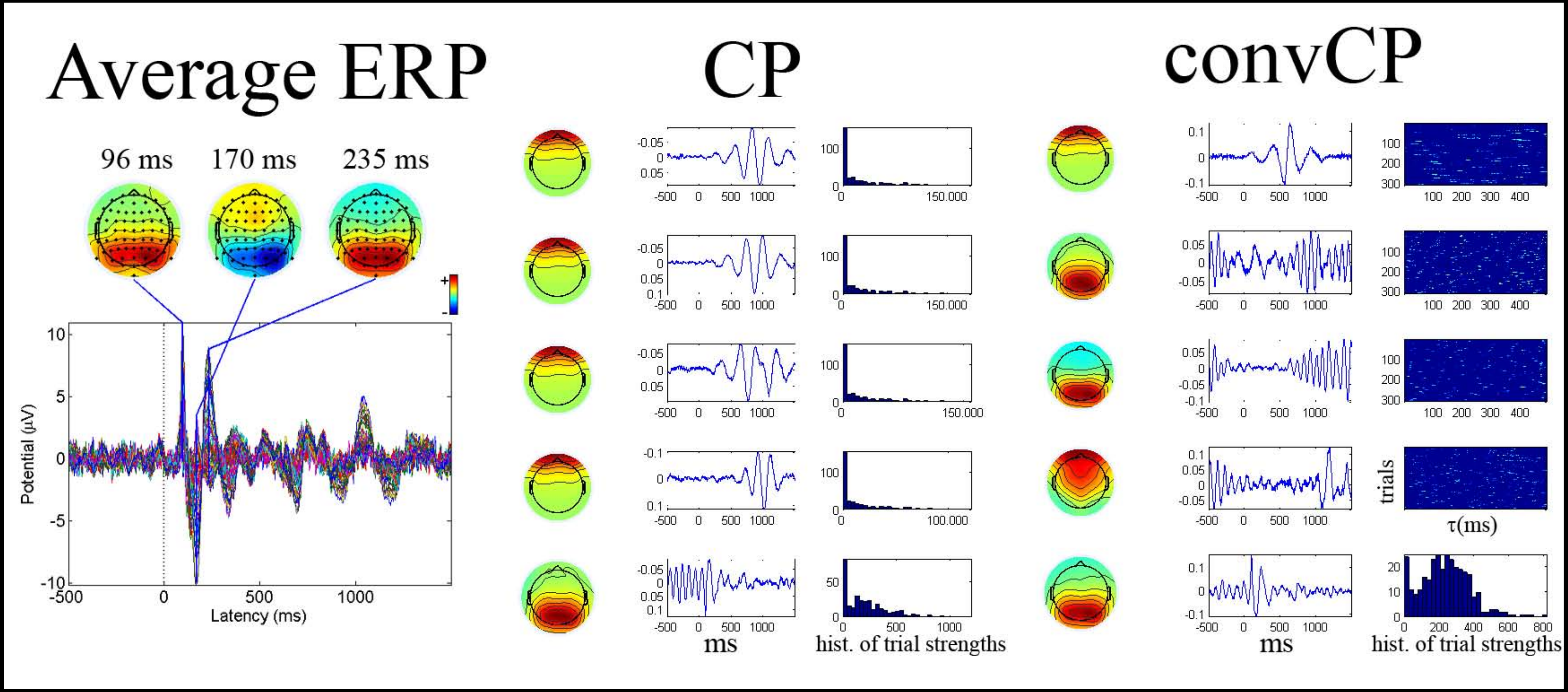
CP, ShiftCP and ConvCP



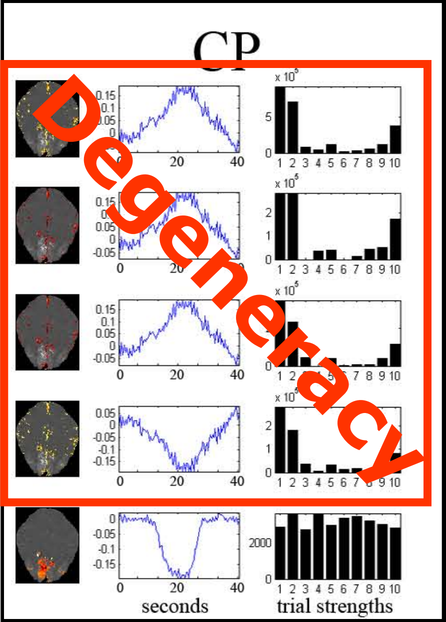
ConvCP: **Can model arbitrary number of component delays within the trials and account for shape variation within the convolutional model representation. Redundancy between what is coded in C and B resolved by imposing sparsity on C.**

(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)

Convolutive bi-linear model form a Latent Causal Modeling framework

Channel Specific Input Functions

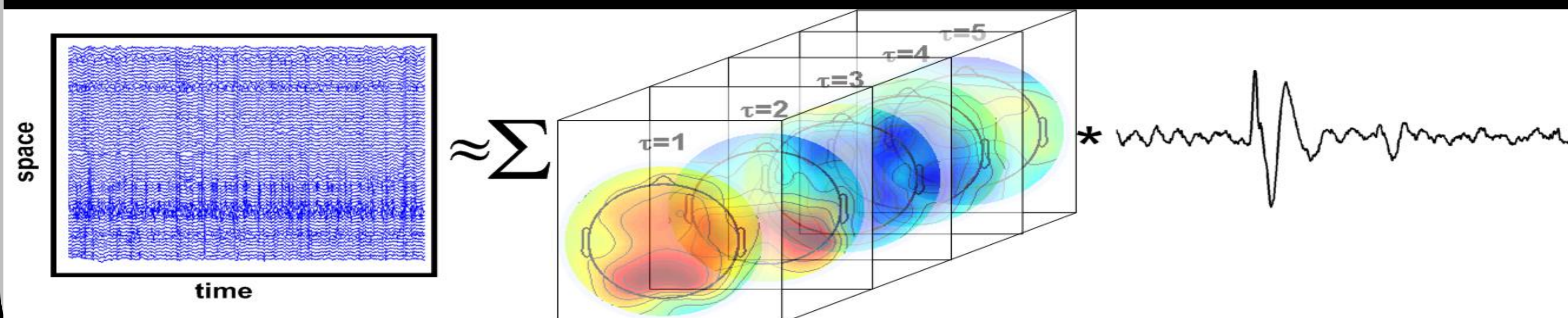
$$\textbf{DTF:}\quad x_i(t) = \sum_{\tau=0}^{\Upsilon-1}\sum_{j=1}^{J} h_{i,j}(\tau)e_j(t-\tau)$$

$$\textbf{SLCM:}\quad x_i(t) = \sum_{\tau=0}^{\Upsilon-1}\sum_{d=1}^{D} a_{i,d}(\tau)s_d(t-\tau)+\varepsilon_i(t).$$

Noise

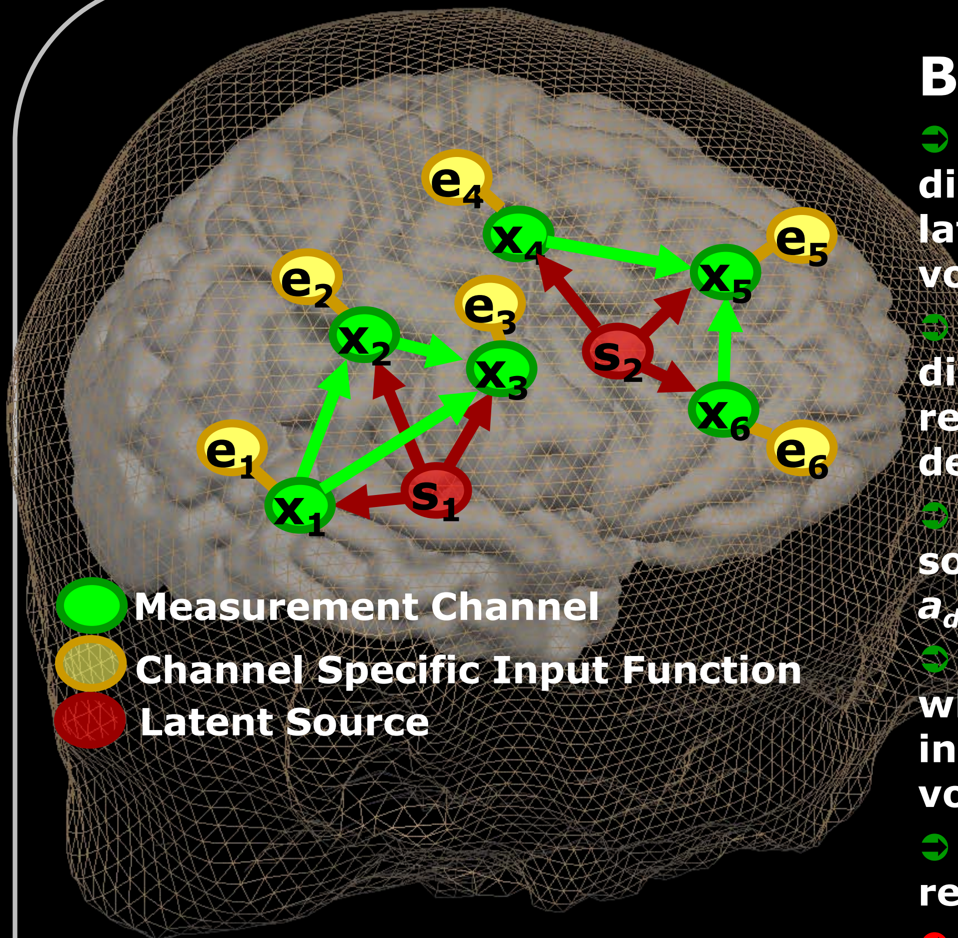
Transfer Functions

Latent Sources



$\mathbf{x}(t) = \sum_\tau \mathbf{A}(\tau)\mathbf{s}(t-\tau) = \sum_\tau \mathbf{A}(\tau)\mathbf{Q}\mathbf{Q}^{-1}\mathbf{s}(t-\tau) = \tilde{\mathbf{A}}(\tau)\tilde{\mathbf{s}}(t-\tau)$ → **We impose sparsity on** $\mathbf{A}(\tau)$

(Mørup et al., Nips workshop on Connectivity Inference in Neuroimaging 2009)



Benefits of SLCM over DTF:

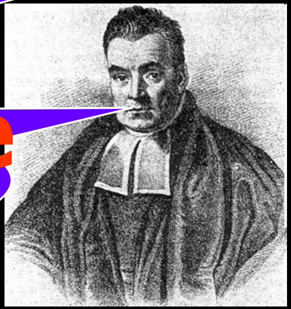
- **SLCM can potentially perform dimensionality reduction resulting in fewer latent sources than observed measurement voxels/channels.**
- **Constraints on the causal relations can be directly imposed on $A(\tau)$ such as sparsity and restricting the transfer function to specific delays.**
- **Spatial regions that are caused by the d^{th} source $s_d(t)$ are automatically grouped in $a_d(\tau)$.**
- **SLCM can handle instantaneous mixing whereas DTF is hard to interpret in case of instantaneous propagation between voxels/channels.**
- **SLCM can naturally handle overcomplete representations, i.e. I>>T.**
- **The estimation of SLCM is a non-convex problem!**

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?

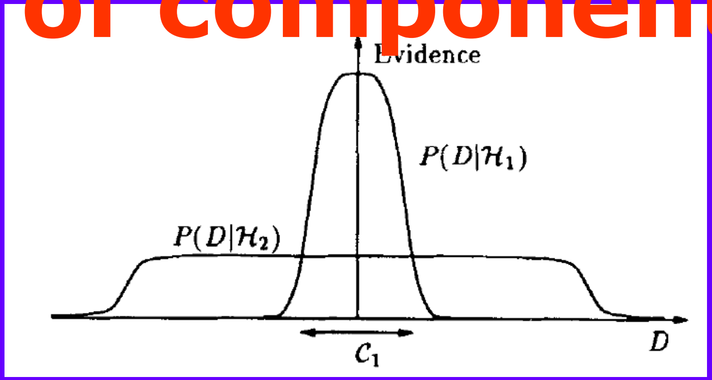


Thomas Bayes

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.



Bayesian inference admit estimation of model order and degree of sparsity through Automatic Relevance Determination

$$\textbf{SLCM:} \quad x_i(t) = \sum_{\tau=1}^{\Upsilon-1} \sum_{d=1}^{D} a_{i,d}(\tau) s_d(t-\tau) + \varepsilon_i(t).$$

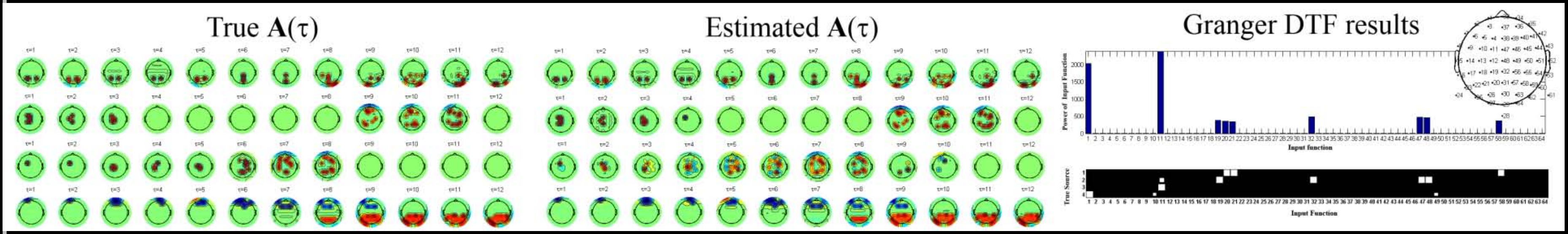
$$\begin{aligned}
\varepsilon_i(t) &\sim Normal(0, \sigma^2) \\
\sigma^{-2} &\sim Gamma(1, \kappa \|\boldsymbol{X}\|_F^2) \\
\boldsymbol{a}_d(\tau) &\sim Laplace(0, \beta_d) \\
\beta_d &\sim Gamma(1, \alpha) \\
s_d(t) &\sim \delta(1 - \textstyle\sum_t s_d(t)^2)
\end{aligned}$$

$$\log P(\boldsymbol{X}, \mathcal{A}, \boldsymbol{S}, \sigma^{-2}, \boldsymbol{\beta} | \kappa, \alpha) = \left\{ \begin{array}{l}
-\frac{\sigma^{-2}}{2} \sum_t^T \|\boldsymbol{x}(t) - \sum_\tau \boldsymbol{A}(\tau)\boldsymbol{s}(t-\tau)\|_F^2 \\
-\frac{1}{2} IT \log(\sigma^{-2}) - \kappa \|\boldsymbol{X}\|_F^2 \sigma^{-2} \\
+\sum_d I\Upsilon \log \beta_d - \beta_d(\alpha + \sum_i^I \sum_\tau^\Upsilon |a_{i,d}(\tau)|) \\
+const. \\
\quad s.t. \quad \sum_t s_d(t)^2 = 1
\end{array} \right.$$

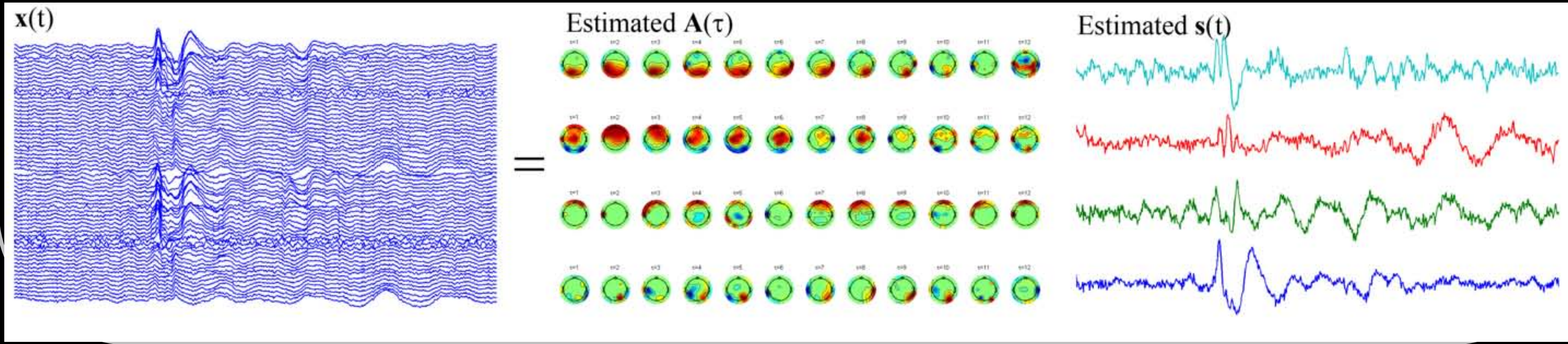
Regularization strength learned from data, i.e. $\beta_d^{\mathrm{MAP}} = \dfrac{I\Upsilon}{\alpha + \sum_i^I \sum_\tau^\Upsilon |a_{i,d}(\tau)|}$

SLCM analysis of synthetic and real EEG

Synthetic EEG data

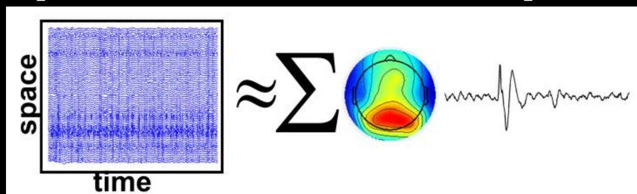


Real EEG data

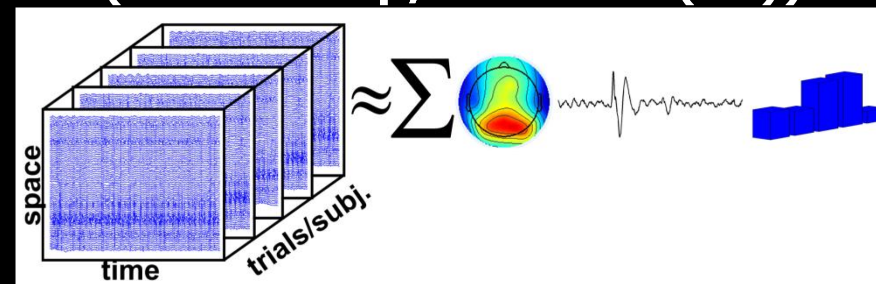


Summary of the ”tour de models”

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

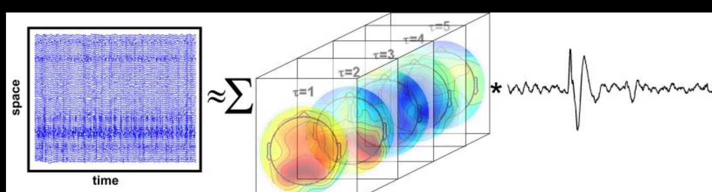


Multi-linear modelling (CandeComp/PARAFAC (CP))

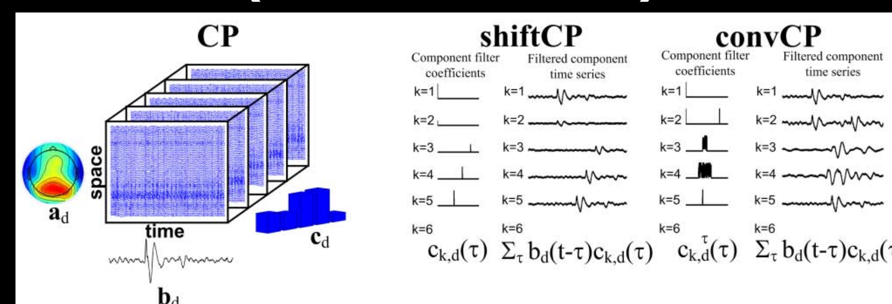


Extensions to model delay and shape changes

Convolutive Bi-linear modelling (related to Latent Causal Modeling)



Convolutive multi-linear modelling (shiftCP/convCP)

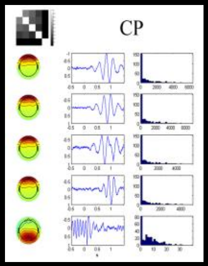


AIM of analysis

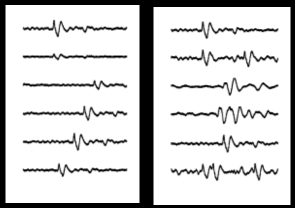
- Extract an efficient internal representation of the statistical structure implicit in the data
- Drive novel hypothesis for formal testing on validation data sets

Conclusion

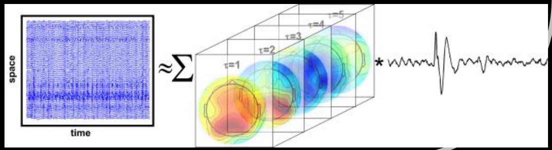
- Multi-linear modeling offers the ability to extract the consistent activity of neuroimaging data over repeats/subjects/conditions etc.



- However, violation of multi-linearity due to variability causes degeneracy
- Common causes of variability in neuroimaging data are delay and shape variation

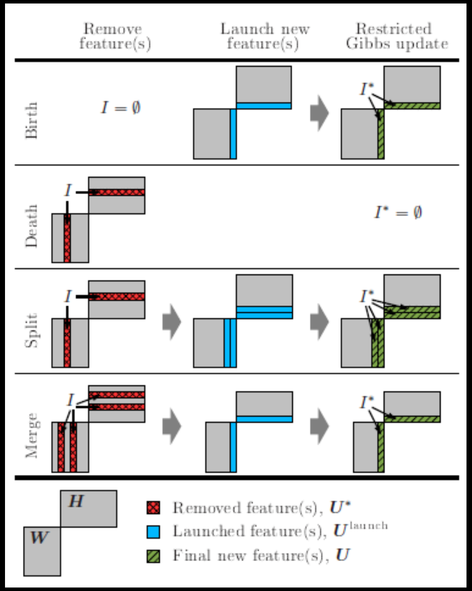


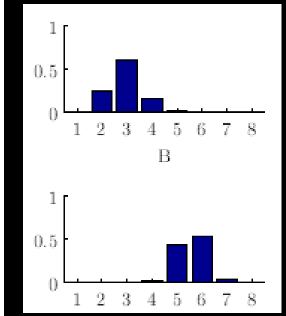
- Advancing the CP model to ShiftCP and ConvCP enables to address these types of variability.
- Modelling delay and shape changes is also relevant for bi-linear modelling and open doorways to address latent causal relations.

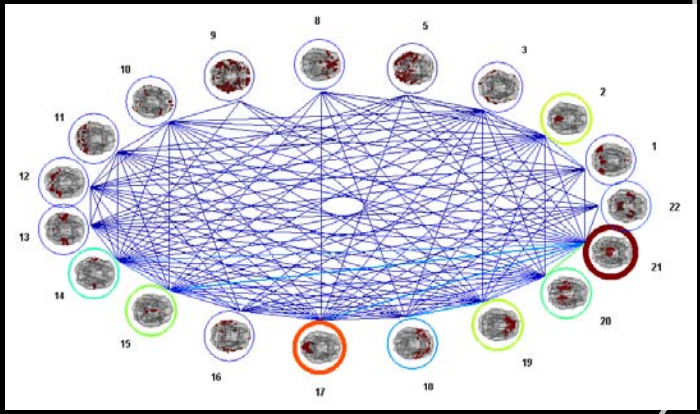


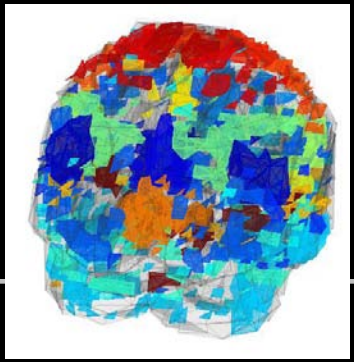
Current research

- Non-parametric efficient sampling approaches based on reversible jump MCMC for the described models.
  (See also Schmidt and Mørup, Infinite Non-negative Matrix Factorization, to appear EUSIPCO 2010)
- Analysis of neuroimaging data as complex networks using non-parametric community detection approaches.









DTU

Relevant papers

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. Schmidt, M Mørup, Infinite Non-negative Matrix Factorization, to appear Eusipco 2010

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

M. Mørup, Kristoffer H. Madsen, L.K. Hansen, Latent Causal Modeling of Neuroimaging Data, NIPS workshop on Connectivity inference in Neuroimaging data, 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

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