Decomposition methods for explorative neuroimaging

Lars Kai Hansen

DTU Informatics
Technical University of Denmark

Co-workers:
Morten Mørup, Kristoffer Madsen, Finn Å. Nielsen, Mads Dyrholm, Stephen Strother, Rasmus Olsson, Thomas Kolenda, Niels Mørch, Ulrik Kjems, Sidse Arnfred, Benny Lautrup.







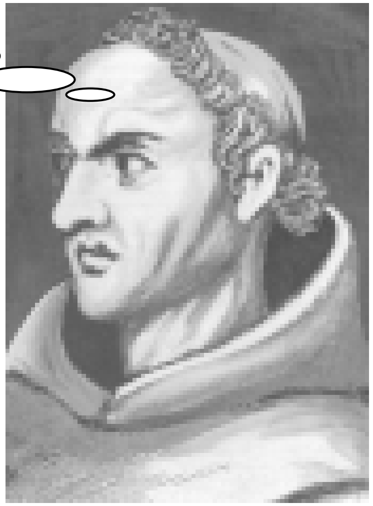



Cimbi
Center for integrated molecular brain imaging

DTU

Do not multiply causes!

OUTLINE



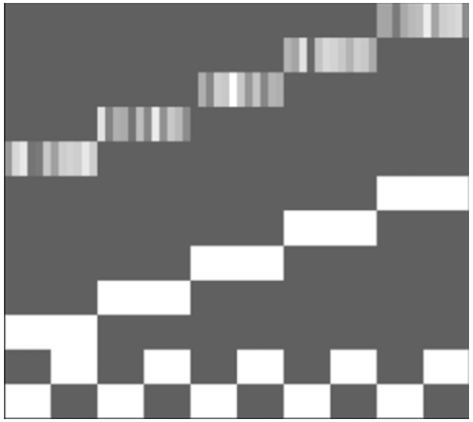
- Hypothesis testing vs knowledge discovery
  - Generalizability
- Factor models - Linear hidden variable representations
- Principal component analysis (PCA)
  - Understanding the limits to learning in high-dimensional data
  - Heuristics to heal overfitting in poor SNR's: Re-scaling projections
- Independent component analysis (ICA)
  - Statistical modeling and generalization
  - Generalizations: Convolutive mixing, shift
- More generalizations
  - NMF
  - Multiway modeling

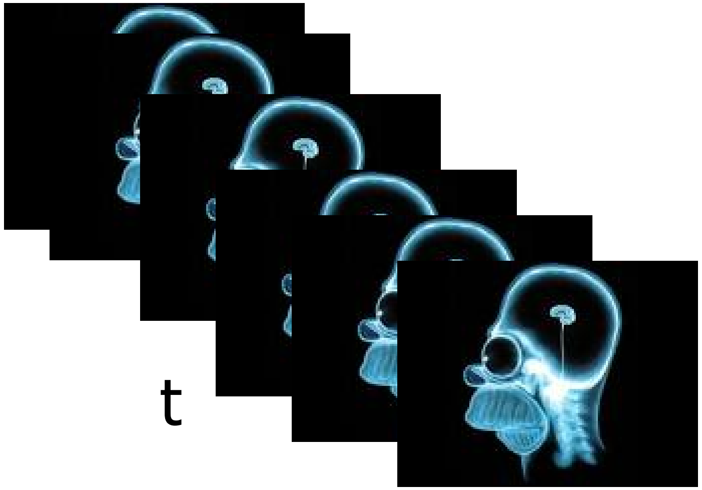
Multivariate neuroimaging models

Neuroimaging aims at extracting the mutual information between stimulus and response.

- Stimulus: Macroscopic variables, ”design matrix” ... *s(t)*
- Response: Micro/meso-scopic variables, the neuroimage ... *x(t)*
- Mutual information is stored in the joint distribution … *p(x,s).*

Often s(t) is assumed known….unsupervised methods consider s(t) or parts of s(t) ”hidden”…..





Cimbi Center for integrated molecular brain imaging
Lars Kai Hansen
IMM, Technical University of Denmark
DTU

Multivariate neuroimaging models

- Univariate models (SPM, fMRI time series models etc)

$$p(x,s) = p(x \mid s)p(s) = \prod_j p(x_j \mid s) \cdot p(s)$$

SPM Central

- Multivariate models (PCA, PLS, ICA, ANN etc)

$$p(x,s) = p(s \mid x)\, p(x)$$

- Modeling from data (D) w. parameterized function families

$$p(s \mid x) \sim p(s \mid x, D) \sim p(s \mid x, \hat{\theta}), \qquad \hat{\theta} = \hat{\theta}(D)$$

$$p(x) \sim p(x \mid D) \sim p(x \mid \hat{\theta}),$$

Early multivariate neuroimage analyses

- **Principal component analysis**
  Regional analysis (Moeller and Strother, *JCBFM* 1991), Spatial (Friston et al., *JCBFM* 1993), Selection of subspace dimension (Hansen et al. *NeuroImage* 1999)
- **Linear multivariate models**
  Worsley et al. (*NeuroImage* 1997), Discriminants (Mørch et al. *IPMI* 1997)
- **Non-linear multivariate models**
  Artif. neural networks (Lautrup et al., 1995; Mørch et al., *IPMI* 1997)
  Independent components (McKeown et al., *PNAS* 1998)
  Spatio-temporal clustering (Ding et al. ISMR 1994, Toft et al, *HBM Conf. 1997*, Goutte et al., *NeuroImage* 1999) Meta analysis (Goutte et al. *HBM* 2001)
- **Plurality and similarity**
  ROC curves for multiple models (Lange et al., *NeuroImage* 1999)
  Consensus of activation maps (Hansen et al., *NeuroImage* 2001)

Generative model

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

$$p(\mathbf{x} \mid \mathbf{A}, \boldsymbol{\theta}) = \int p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}, \boldsymbol{\Sigma})\, p(\mathbf{s} \mid \boldsymbol{\theta})\, d\mathbf{s}$$

$$p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2}\, e^{-\frac{1}{2}(\mathbf{x}-\mathbf{A}\mathbf{s})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mathbf{A}\mathbf{s})}$$

Source distribution:
PCA: … normal
ICA: … other
IFA: … Gauss. Mixt.

PCA: $\boldsymbol{\Sigma} = \sigma^2 \cdot \mathbf{1},$

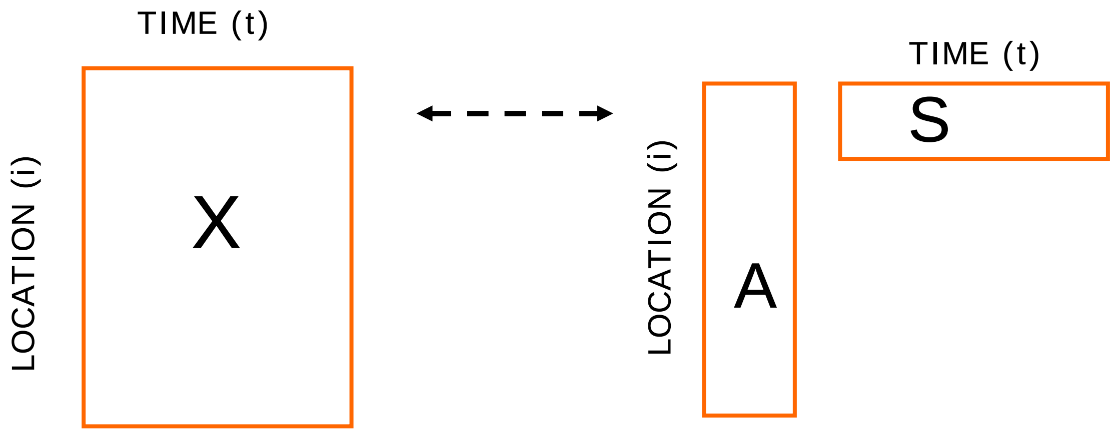
FA: $\boldsymbol{\Sigma} = \mathbf{D}$

S known:	GLM
$(1-A^{-1})$ sparse:	SEM
S,A positive:	NMF

Højen-Sørensen, Winther, Hansen,
Neural Comp (2002), Neurocomputing (2002)

Factor models

- Represent a datamatrix by a low-dimensional approximation



$$X(i,t) \approx \sum_{k=1}^{K} A(i,k) S(k,t)$$

Matrix factorization: SVD / PCA, NMF, Clustering

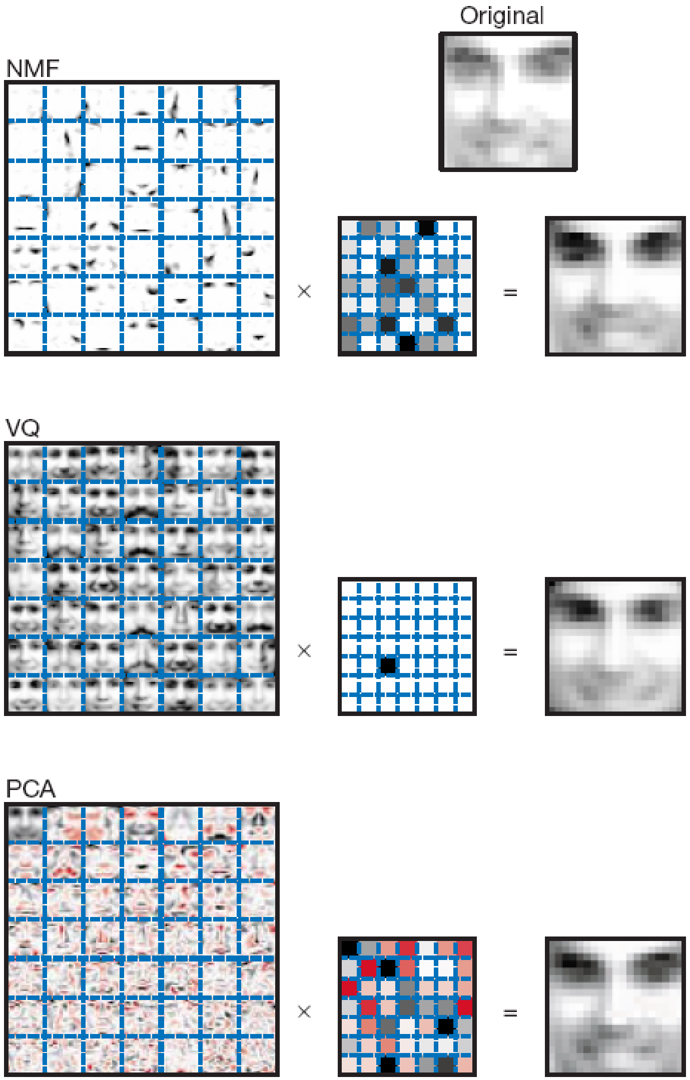


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting of $n = 19 \times 19$ pixels, and constituting an $n \times m$ matrix V. All three find approximate factorizations of the form $V \approx WH$, but with three different types of constraints on W and H, as described more fully in the main text and methods. As shown in the 7×7 montages, each method has learned a set of $r = 49$ basis images. Positive values are illustrated with black pixels and negative values with red pixels. A particular instance of a face, shown at top right, is approximately represented by a linear superposition of basis images. The coefficients of the linear superposition are shown next to each montage, in a 7×7 grid, and the resulting superpositions are shown on the other side of the equality sign. Unlike VQ and PCA, NMF learns to represent faces with a set of basis images resembling parts of faces.

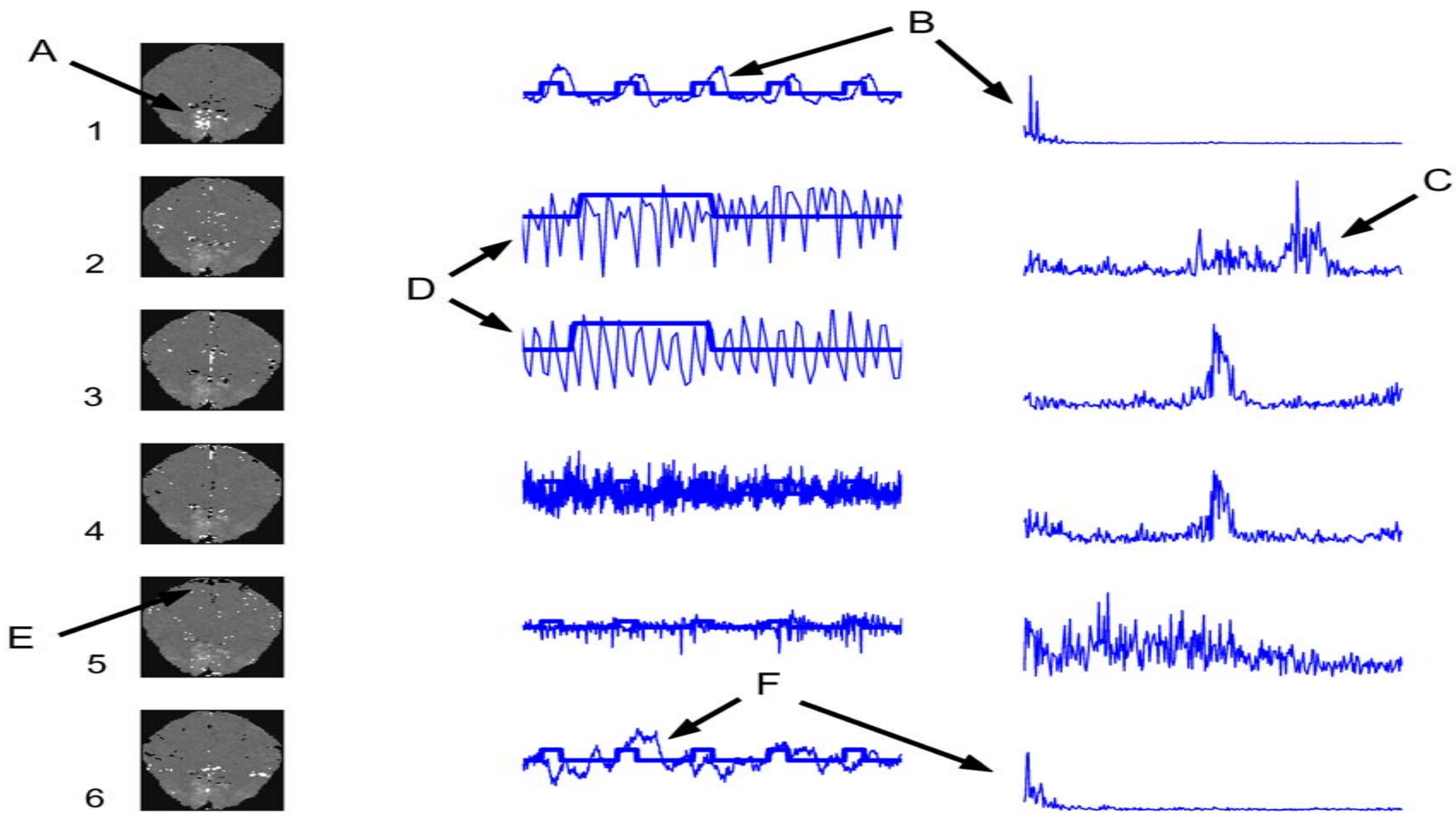
Learning the parts of objects by non-negative matrix factorization

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

NATURE | VOL 401 | 21 OCTOBER 1999 | www.nature.com

ICA: Assume S(k,t), S(k',t) statistically independent



(McKeown, Hansen, Sejnowski, Curr. Op. in Neurobiology (2003)

Generalizability

Do not multiply causes!

- Generalizability is defined as *the expected performance on a random new sample* ... the mean performance of a model on a ”fresh” data set is an unbiased estimate of generalization

- Typical loss functions:

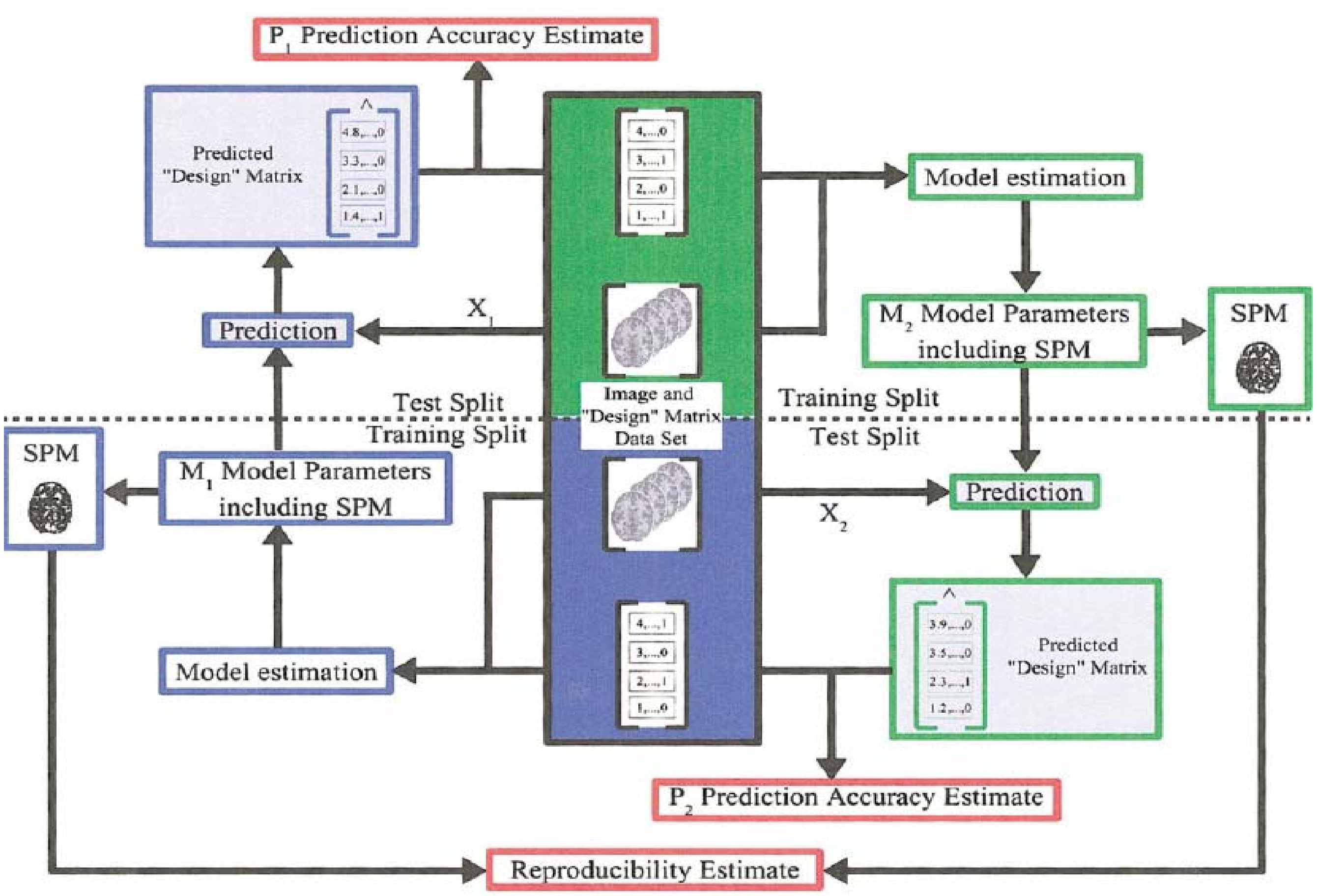
$$\langle -\log p(\mathbf{s} \mid \mathbf{x}, D) \rangle, \qquad \langle -\log p(\mathbf{x} \mid D) \rangle,$$

$$\langle (\mathbf{s} - \hat{\mathbf{s}}(D))^2 \rangle, \qquad \left\langle \log \frac{p(\mathbf{s}, \mathbf{x} \mid D)}{p(\mathbf{s} \mid D)\, p(\mathbf{x} \mid D)} \right\rangle$$

- Note: Even for hidden variable models we can ”predict”!

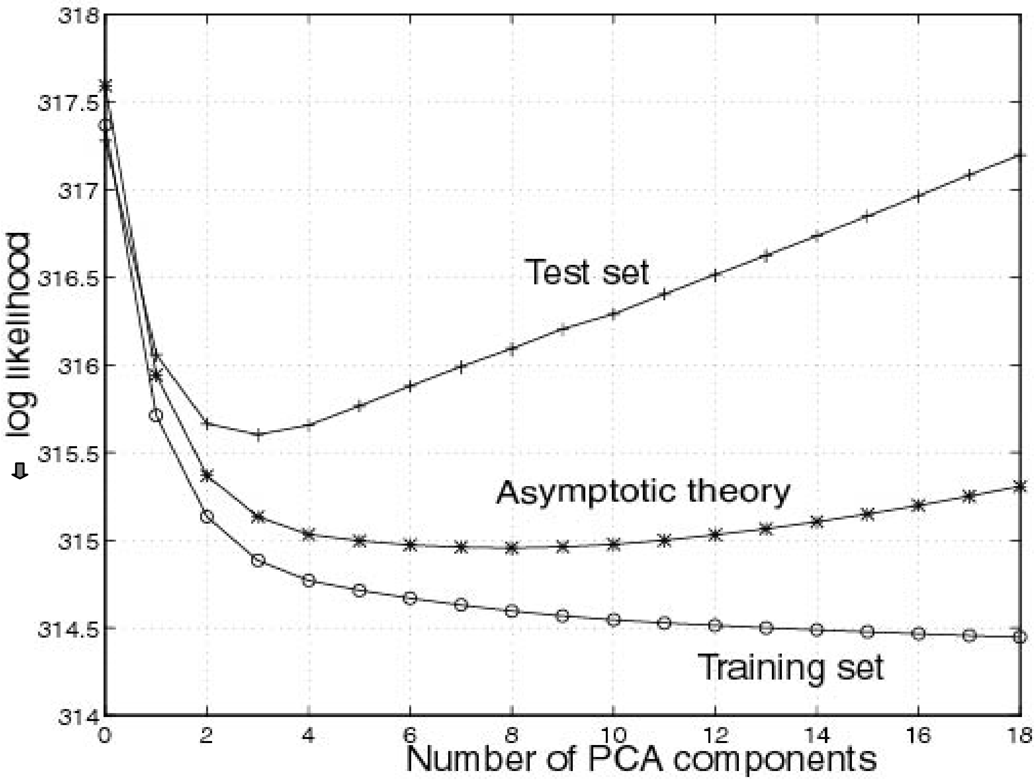
- Results can be presented as ”bias-variance trade-off curves” or ”learning curves”

NPAIRS: Reproducibility of parameters



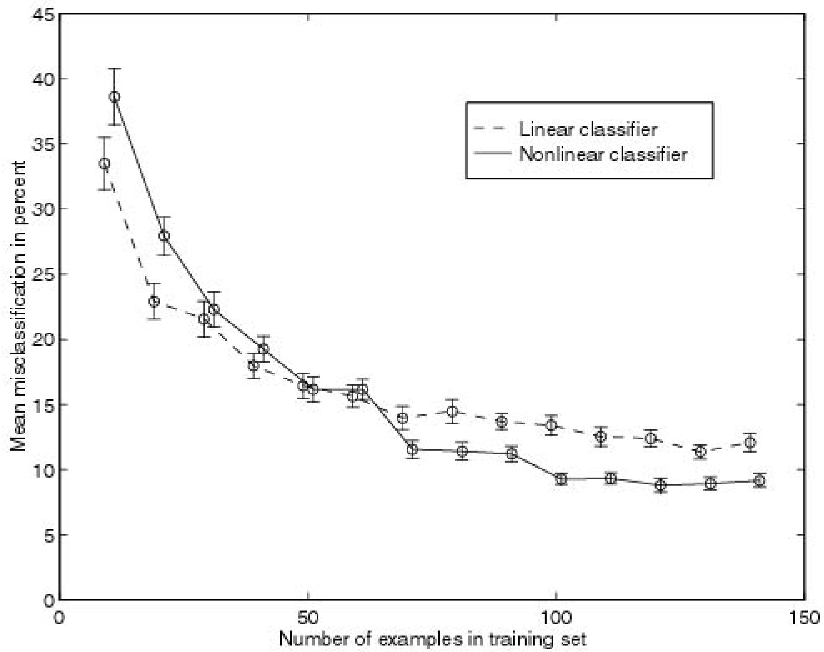
NeuroImage: Hansen et al (1999), Hansen et al (2000), Strother et al (2002), Kjems et al. (2002), LaConte et al (2003), Strother et al (2004)

Bias-variance trade-off as function of PCA dimension

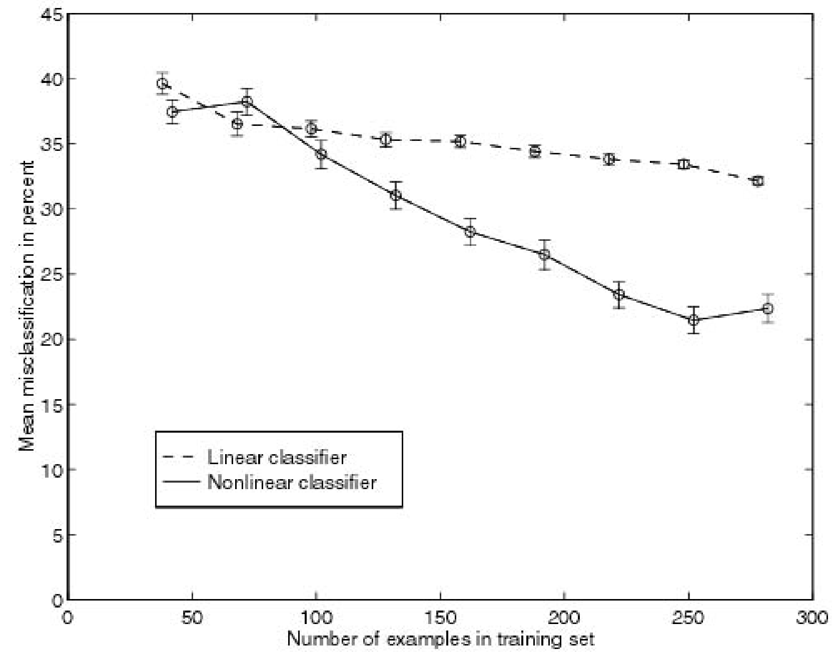


Hansen et al. *NeuroImage* (1999)

Learning curves multivariate models: Within group generalizability



PET



fMRI

Finger tapping, analysed by PCA dimensional reduction and Fisher LD / ML Perceptron. Mørch et al. *IPMI* (1997)...."first mind reading in fMRI"

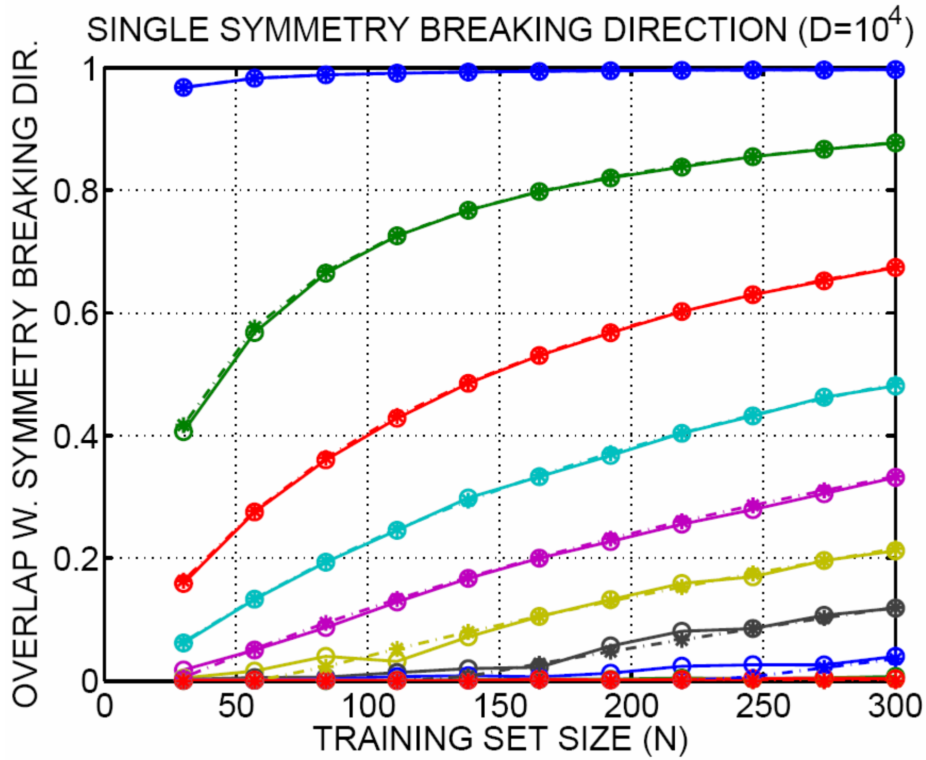
Modeling the generalizability of SVD

- Rich physics literature on ”retarded” learning
- **Universality**
  - Generalization for a ”single symmetry breaking direction” is a function of ratio of N/D and signal to noise S
  - For subspace models-- a bit more complicated -- depends on the component SNR’s and eigenvalue separation
  - For a single direction, the mean squared overlap $R^2 = \langle (u^T_1 * u_0)^2 \rangle$ is computed for N,D -> ∞

$$R^2 = \begin{cases} (\alpha S^2 - 1)/S(1+\alpha S) & \alpha > 1/S^2 \\ 0 & \alpha \le 1/S^2 \end{cases}$$

$$\alpha = N/D \qquad S = 1/\sigma^2 \qquad N_c = D/S^2$$

Hoyle, Rattray: Phys Rev E **75** 016101 (2007)



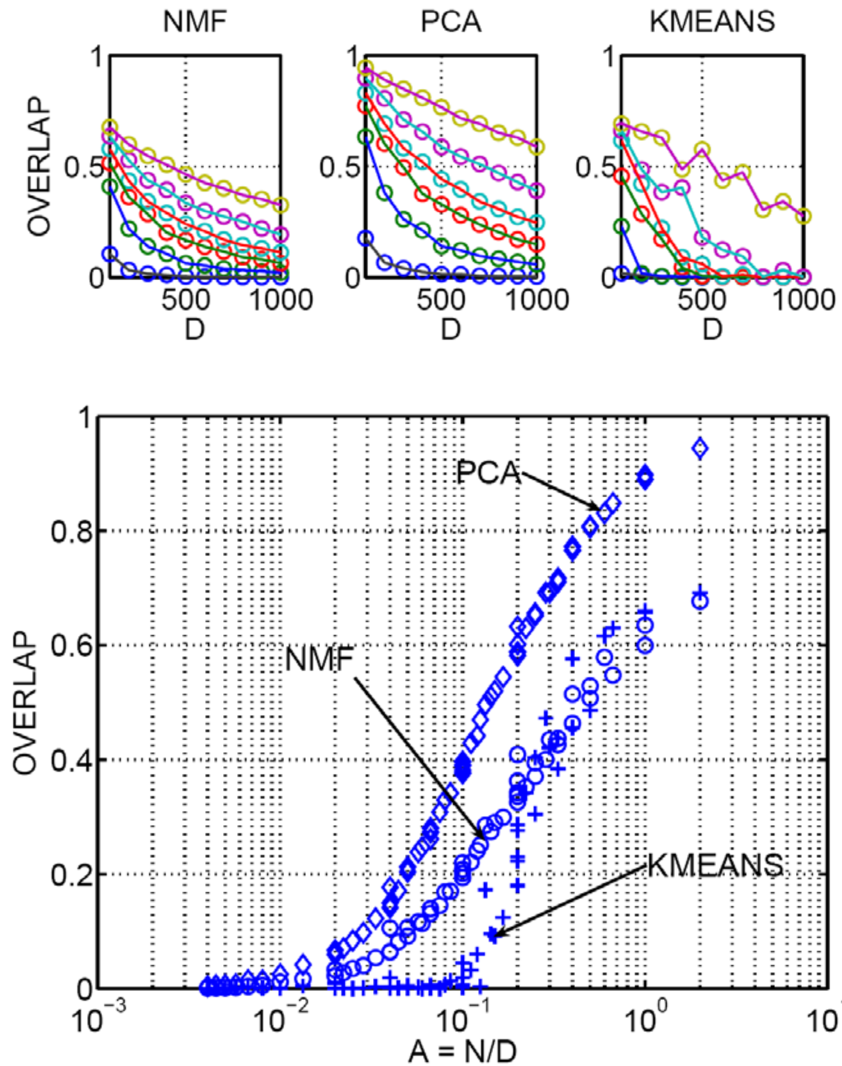
N_c = (0.0001, 0.2, 2, 9, 27, 64, 128, 234, 400, 625)

σ = (0.01, 0.06, 0.12, 0.17, 0.23, 0.28, 0.34, 0.39, 0.45, 0.5)

Universality in PCA, NMF, Kmeans

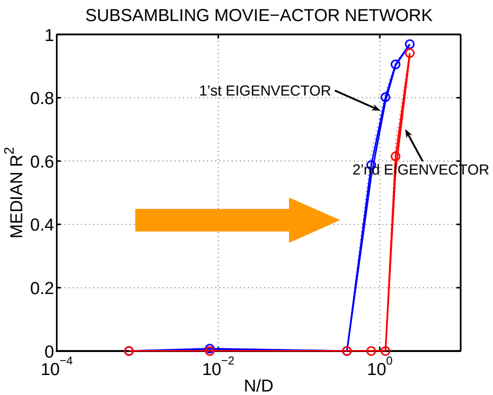
- Looking for universality by simulation
  - learning two clusters in white noise.
- Train K=2 component factor models.
- Measure overlap between line of sigth and plane spanned by the two factors.

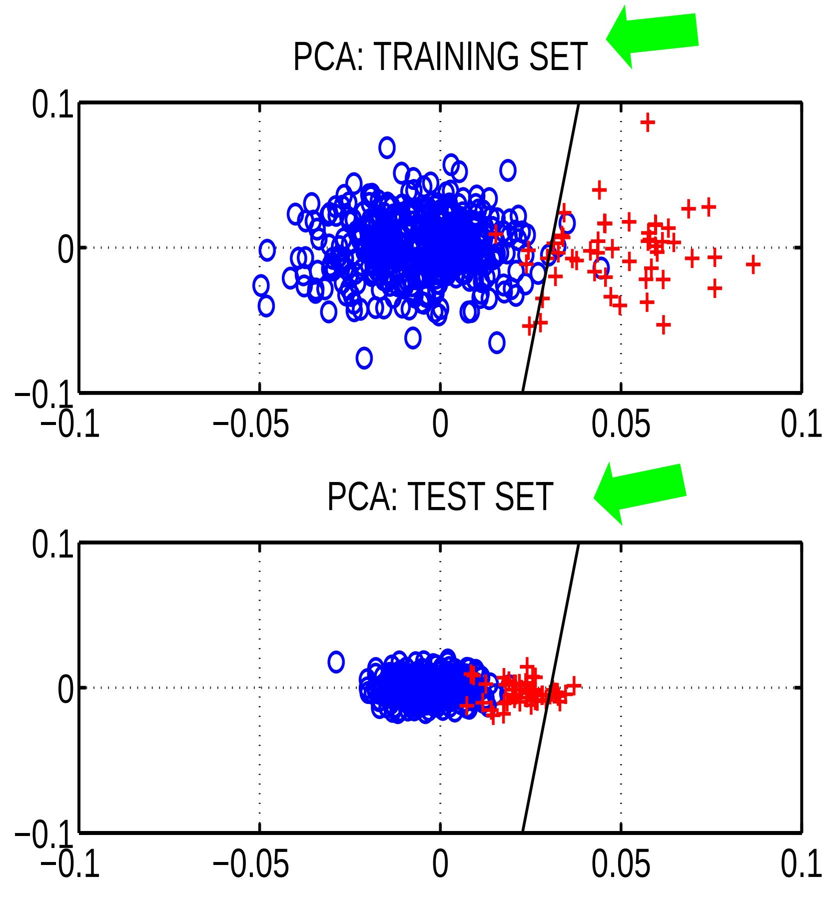
Experiment
Variable: N, D
Fixed: SNR



Restoring the generalizability of SVD

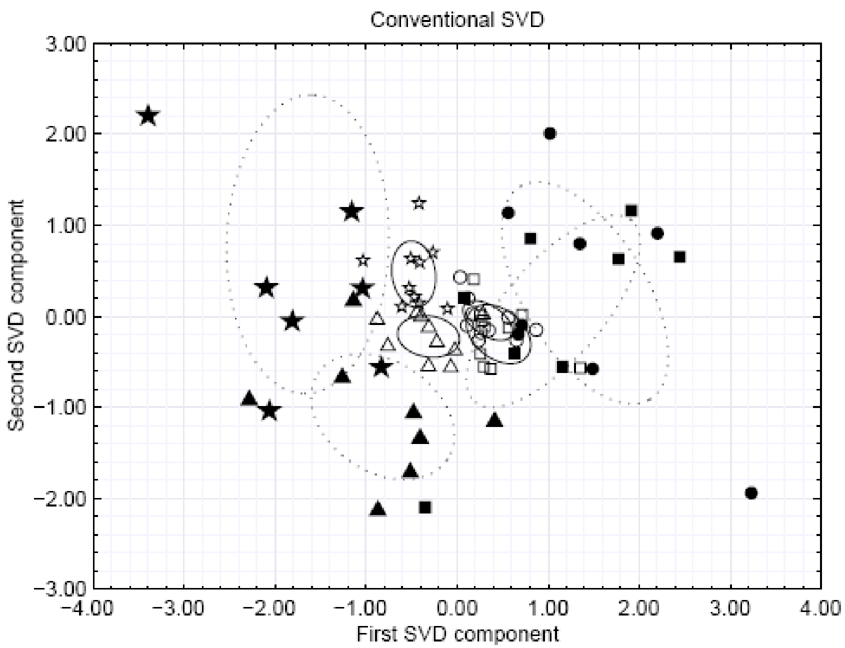
- Now what happens if you are on the slope of generalization, i.e., N/D is just beyond the transition to retarded learning ?

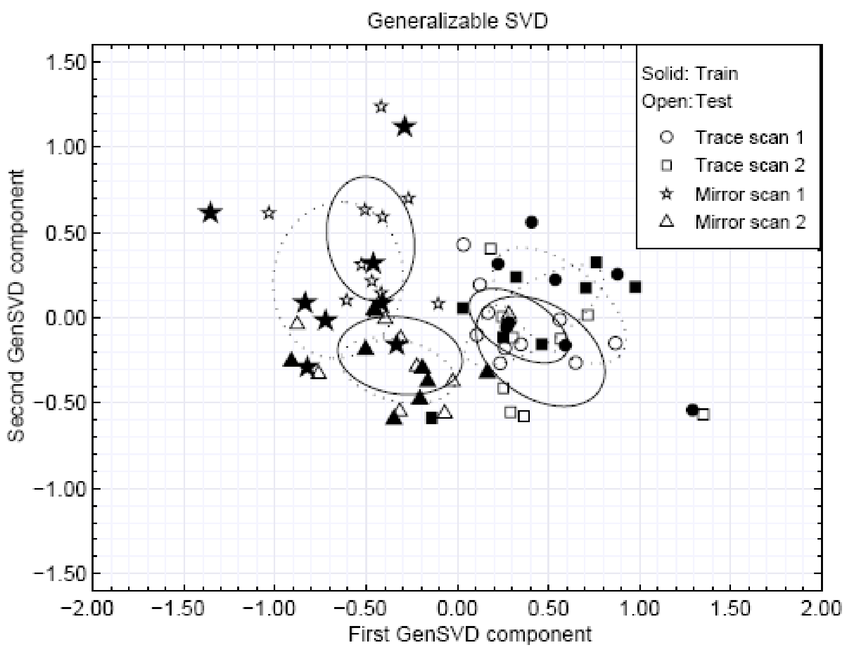


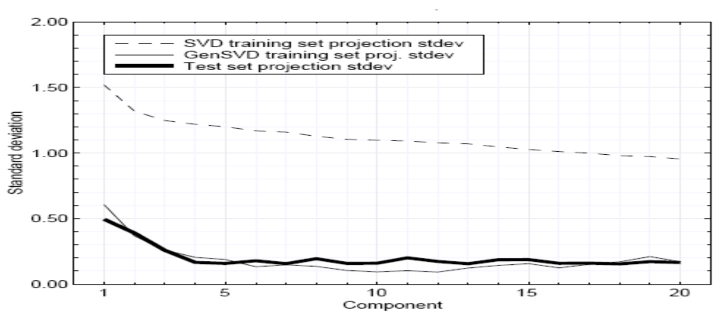


- The estimated projection is offset, hence, future projections will be too small!

- …problem if discriminant is optimized for unbalanced classes in the training data!





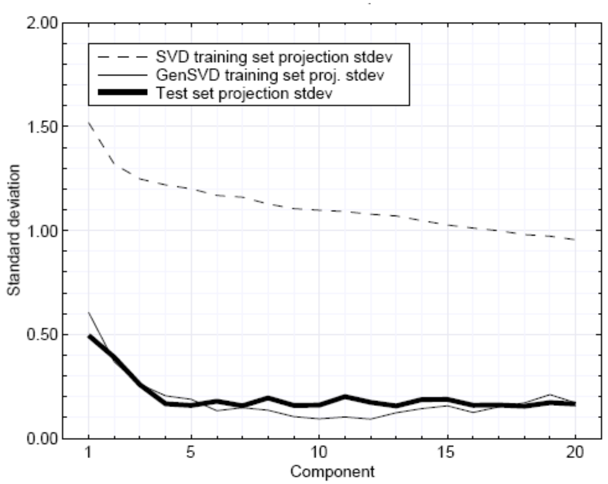


N=72, D=2.5 10^4

Kjems, Hansen, Strother: ”Generalizable SVD for Ill-posed data sets” NIPS (2001)

Re-scaling the component variances

- Possible to compute the new scales by leave-one-out doing N SVD's of size N << D



$$\text{Compute } \boldsymbol{U}_0 \boldsymbol{\Lambda}_0 \boldsymbol{V}_0^\top = \text{svd}(X) \text{ and } \boldsymbol{Q}_0 = [\, \boldsymbol{q}_j \,] = \boldsymbol{\Lambda}_0 \boldsymbol{V}_0^\top$$

$$\text{foreach } j = 1 \ldots N$$

$$\bar{\boldsymbol{q}}_{-j} = \frac{1}{N-1} \sum_{j' \neq j} \boldsymbol{q}_{j'}$$

$$\text{Compute } \boldsymbol{B}_{-j} \boldsymbol{\Lambda}_{-j} \boldsymbol{V}_{-j}^\top = \text{svd}(\boldsymbol{Q}_{-j} - \bar{\boldsymbol{Q}}_{-j})$$

$$\boldsymbol{z}_j = \boldsymbol{B}_{-j} \boldsymbol{B}_{-j}^\top (\boldsymbol{q}_j - \bar{\boldsymbol{q}}_{-j})$$

$$\hat{\lambda}_i^2 = \frac{1}{N-1} \sum_j z_{ij}^2$$

Kjems, Hansen, Strother: NIPS (2001)

Identifiability problem in SVD / PCA

Linear model + normality = failure!

$$x(j,t) = \sum_k A(j,k)\ s(k,t)$$

$$R = \frac{1}{T}\sum_t x(j,t)x(j',t) = \langle xx' \rangle = A\langle ss' \rangle A' = AA'$$

$$= AUU'A' = AU(AU)' = BB'$$

If columns of A are in general position,
A can not be recovered by PCA
Call ICA: the rotation buster

Factor analysis with S ”independent” - ICA issues in fMRI analysis

- Basic assumptions:
  - Spatial or temporal independence?
  - Which ICA model, higher order stat or temporal corr to bust rotation?
  - Are confounds independent of design variables?

- Statistical issues
  - How many components to use?
  - Systematic evaluation in terms of generalization

Consensus and model averaging

(1000+ papers/abstracts on ICA+fMRI)

Independent Component Analyses (select. refs.)

- First non-linear approach (Herault&Jutten, 1985)
- Blind signal separation (Cardoso, Comon, 1987-)

- Temporal decorrelation (Molgedey&Schuster, 1994)
- Infomax: High impact work by (Bell&Sejnowski, 1995)
- Infomax applications in EEG: Scott Makeig et al. (1996) ”EEG-Lab”

- Application for fMRI expl. analysis (McKeown et al, 1998)
- High-dimensional mixtures (Hansen&Larsen, 1998)
- Noisy image mixtures => estimated sources
  are non-linear in the measurement! (Hansen, 2000)
- Bayesian ICA (Højen-Sørensen, Hansen &Winther 2001,2002a,2002b)

- Reviews of ICA and fMRI (e.g. McKeown, Hansen, Sejnowski, 2003)
- Special issue of IEEE Trans Bio. Engineering (Calhoun & Adali, 2006)

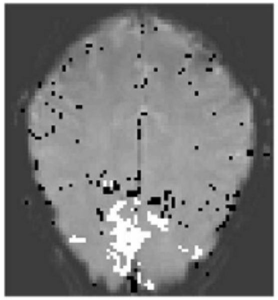
DTU:ICA toolbox (www.imm.dtu.dk/cisp)

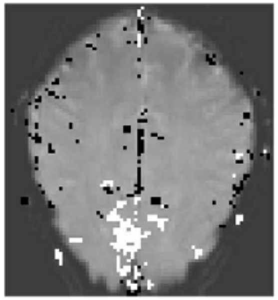
- **Infomax/Maximum likelihood**
  Bell & Sejnowski (1995), McKeown et al (1998)
- **Dynamic Components**
  Molgedey-Schuster (1994), Petersen et al (2001)
- **Mean Field ICA**
  Højen-Sørensen et al. (2001,2002)

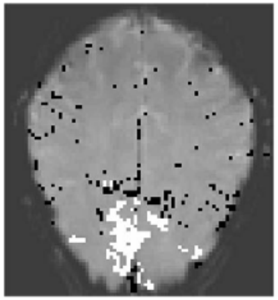
- **Features:**

v Number of components (BIC)
v Parameter tuning
v Binary and mixing contraints (A)
v Demo scripts incl. fMRI data

Spatial mode ICA

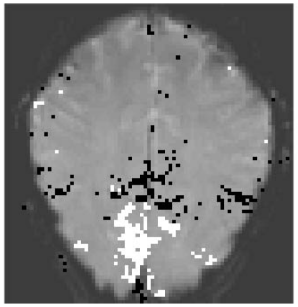


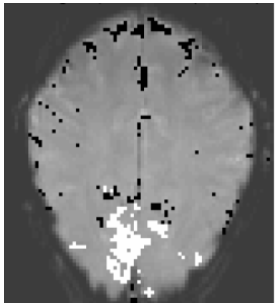


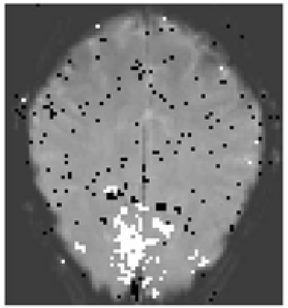


Attias Decorr Max Lik.

Temporal mode ICA

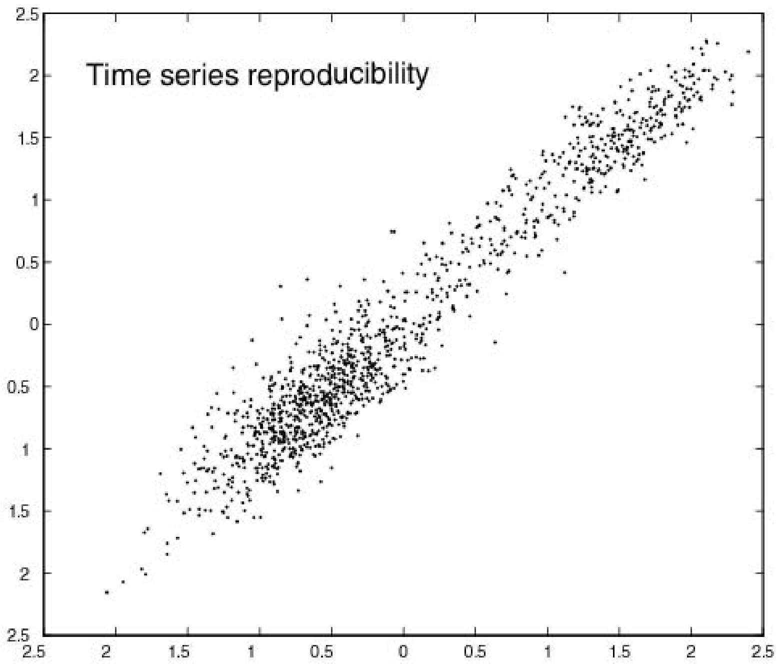




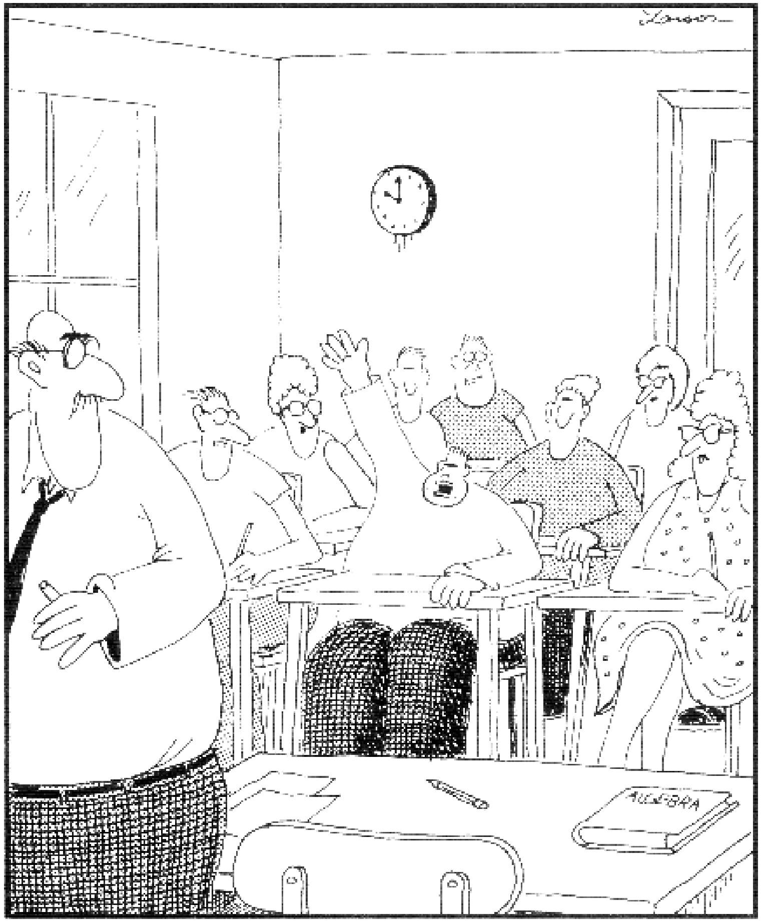


Attias Decorr Max Lik.

Reproducibility of the time series recovered from spatial vs temporal ICA



Petersen et al. ICA2000



"Mr. Osborne, may I be excused? My brain is full."

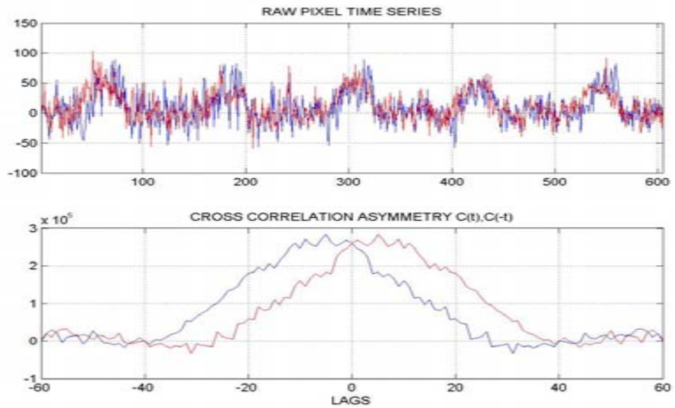
Challenges for the linear factor model

- Temporal structure in networks -> Convolutive ICA
- Group study, repeat trials -> Multiway methods
- Causal models –sparse ICA

Lars Kai Hansen
IMM, Technical University of Denmark

fMRI: Delayed activation in visual cortex

Two voxel temporal cross-correlation



$$\langle x(j,t)x(j',t+\tau)\rangle = \sum\nolimits_{k,k'} A(j,k)A(j',k')\langle s(k,t)s(k',t+\tau)\rangle$$

$$\langle x(j,t)x(j',t+\tau)\rangle = \sum\nolimits_{k} A(j,k)A(j',k)C_k(\tau)$$

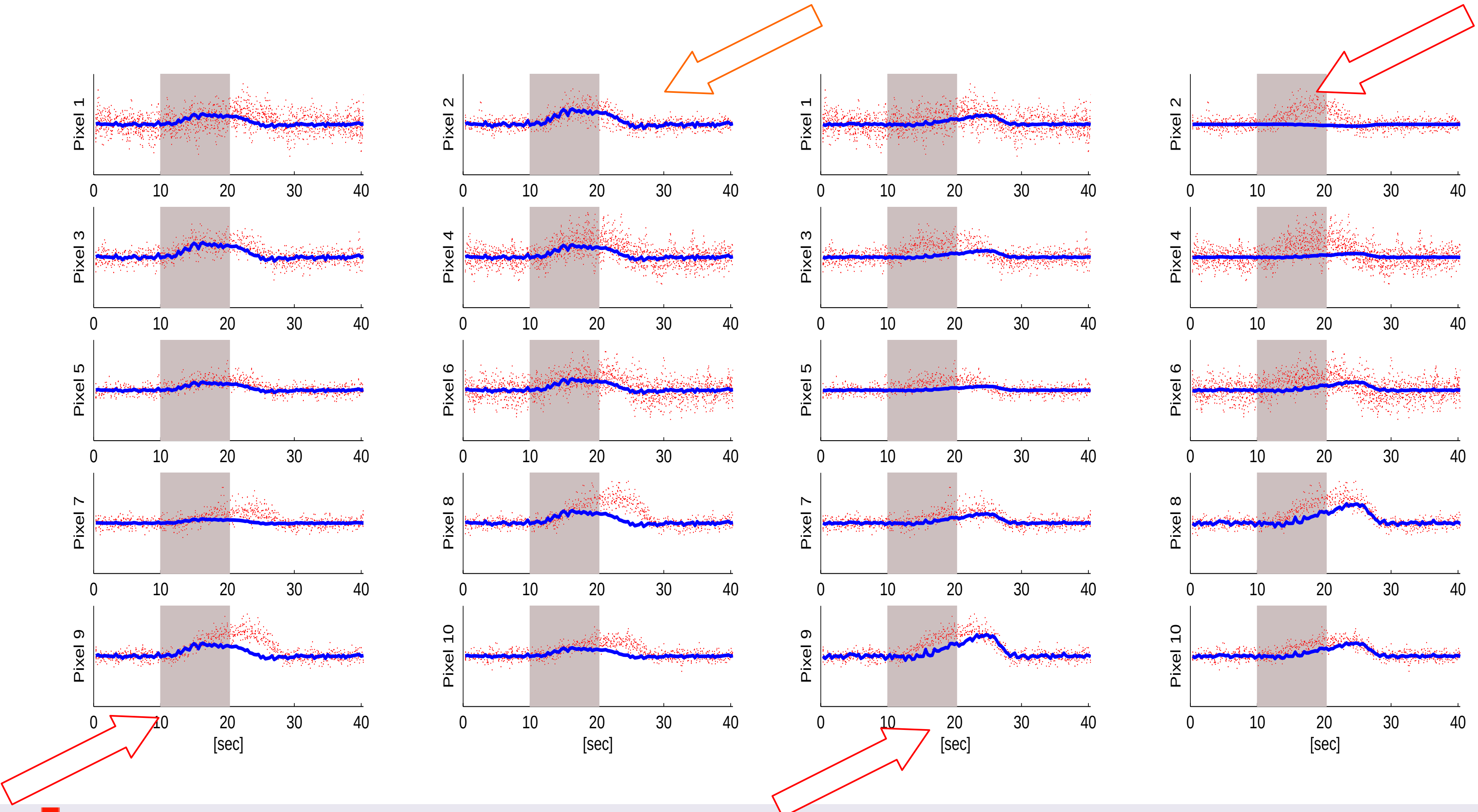
$$\Sigma^{xx}_{j,j'}(\tau) = \Sigma^{xx}_{j,j'}(-\tau)$$

Recovery of source signals from an unknown, linear convolutive mixture

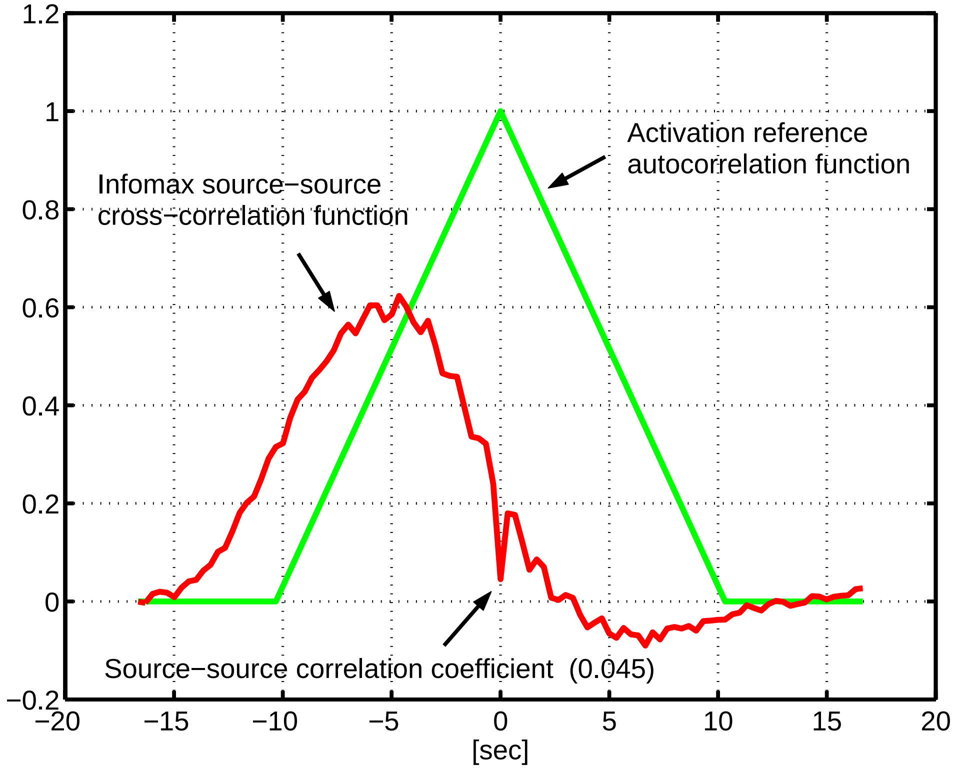
$$x(j,t) = \sum\nolimits_{k,\tau} A(j,k,\tau)\; s(k,t-\tau)$$

Hansen & Dyrholm: Prediction matrix (Proc MLSP, 2003)
Olsson & Hansen: Kalman (Proc NIPS 2004), J. Mach. Learning Res. (2006)
Dyrholm & Hansen: ”CICAAR” (Proc ICA, 2004), Neural Comp (2006)

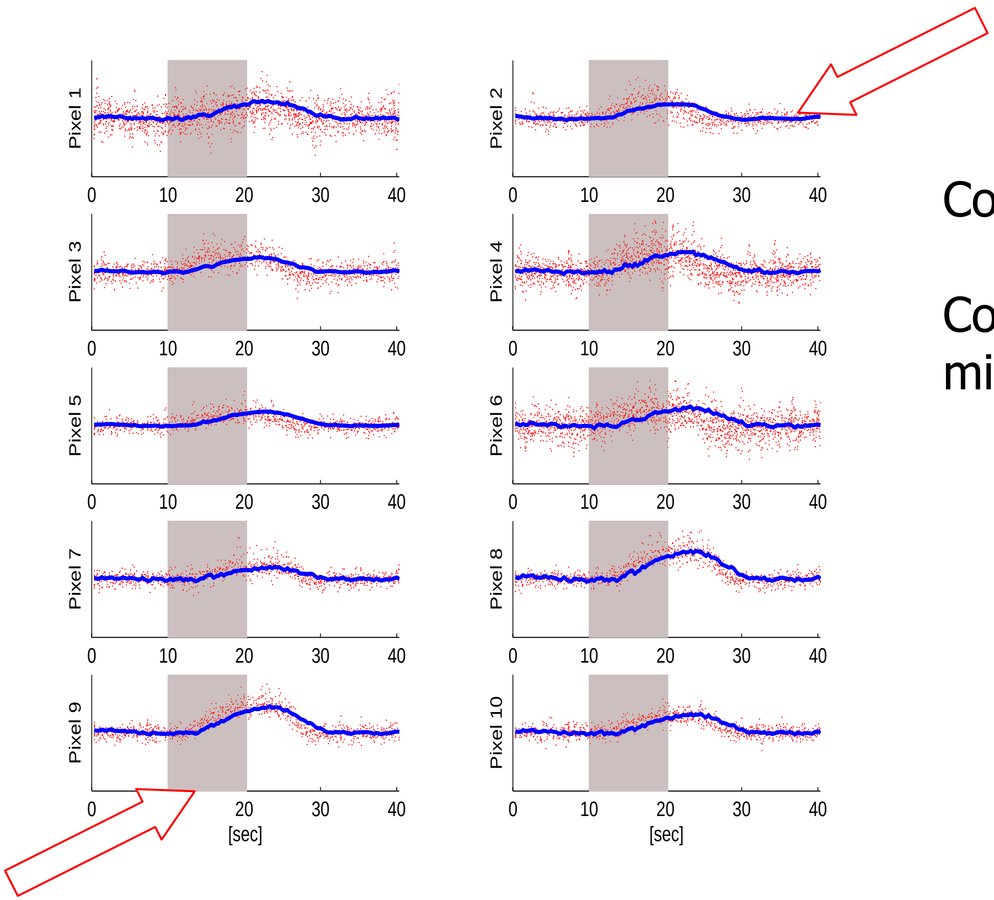
Instantaneous mixing:
Components 1,2 (30% + 30% of variance)



Instantaneous mixing
cross correlation between components



Convolutive mixing (60% of total variance)



Conclusion:

Convolutive ICA is the appropriate mixing model for fMRI

Data represented as multiway arrays

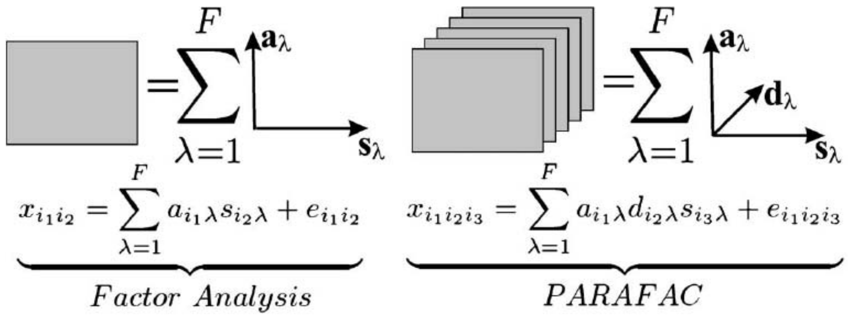
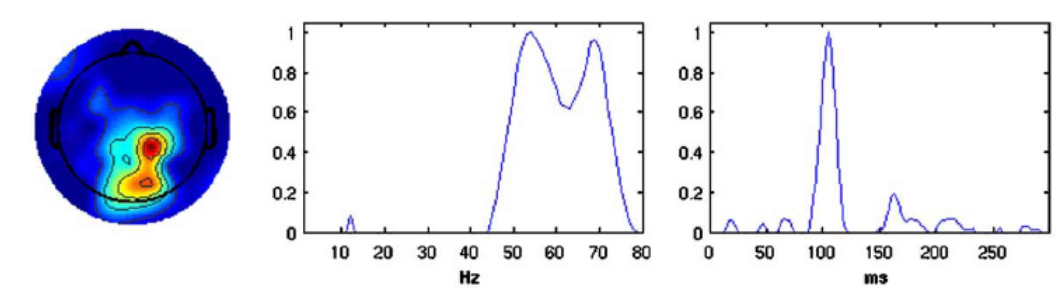


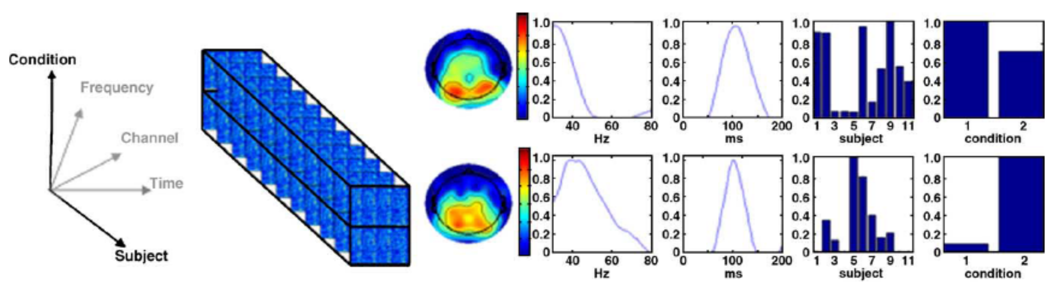
Fig. 1. Graphical representation of the factor analysis to the left and the PARAFAC decomposition of a 3-way array to the right. Like the factor analysis, PARAFAC decomposes the data into factor effects pertaining to each modality. F denotes the number of factors.

EEG visual response to meaningful vs non-meaningful drawings (N=11).

3-way analysis:
Channel*freq*time



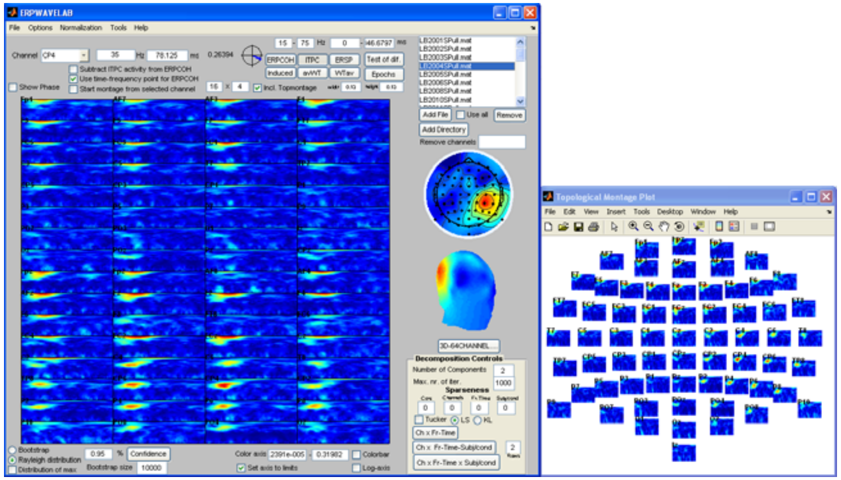
5-way analysis:
Channel*freq*time*subject*condition



Mørup et al. NeuroImage (2005), NeuroImage (2008)

ERPWAVELAB

- Interfaced with EEGLAB
- Single subject analysis
  - Artifact rejection in the time/freq domain
  - NMF decomposition
  - Cross coherence tracking
- Multi subject analysis
  - Clustering
  - Analysis of Variance (ANOVA)
  - Tensor decomposition



Toolbox download from www.erpwavelab.com

Mørup et al. J. Neuroscience Methods (2007),

Conclusions

- Explorative methods can supplement hypothesis testing.
- Multivariate models can detect brain-wide activation networks.
- Generalizability/learning curves show surprizing features: Retarded learning in large systems, crossing learning curves
- Linear hidden variable models form a flexible family including PCA, FA, ICA, Clustering
- New efficient tools for estimation and optimization of multi-linear models are emerging, applications in group analysis, repeat trial models (NeuroImage, in press)

Acknowledgments

Lundbeck Foundation (www.cimbi.org)
NIH Human Brain Project grant (P20 MH57180)
MAPAWAMO / EU Commission
Danish Research Councils

www.imm.dtu.dk/cisp
hendrix.imm.dtu.dk