

Machine learning strategies for fMRI analysis

Lars Kai Hansen

DTU Informatics
Technical University of Denmark

Co-workers:

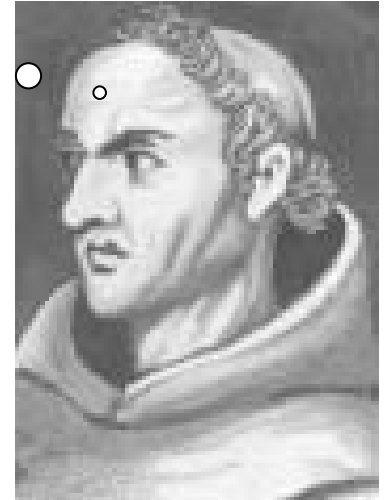
Morten Mørup, Kristoffer Madsen, Peter Mondrup, Daniel Jacobsen, Stephen Strother,



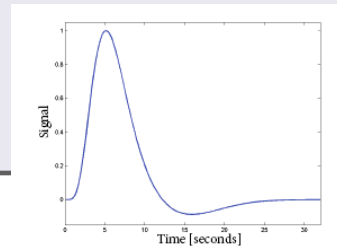
*Do not multiply
causes!*

OUTLINE

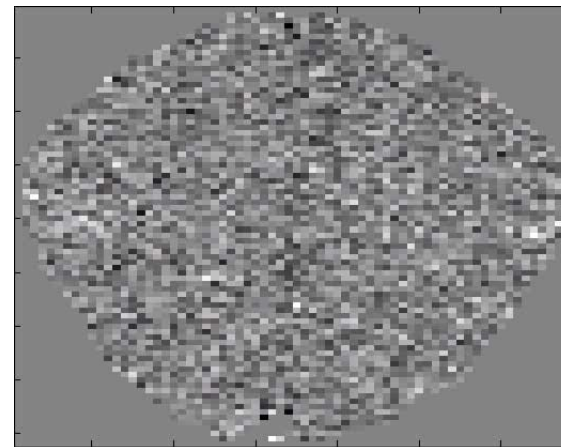
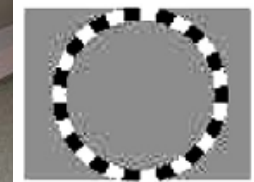
- Machine learning approach with a twist
 - Aim I: To abstract generalizable relations from data
 - Aim II: Robust visualization
- Unsupervised (explorative)
 - Factor models - Linear hidden variable representations
 - Independent component analysis (ICA)
 - Generalizations: Convolutional mixing, shift
 - Multiway modeling
- Supervised models (retrieval)
 - Visualization of non-linear kernel machines



fMRI



- Indirect measure of neural activity - hemodynamics
- A cloudy window to the human brain
- Challenges:
 - Signals are multi-dimensional mixtures
 - No simple relation between measures and brain state - "what is signal and what is noise"?



TR = 333 ms

BOLD fMRI: Is hemodynamic de-convolution feasible?

LETTER Communicated by Karl Friston

Bayesian Model Comparison in Nonlinear BOLD fMRI Hemodynamics

Daniel J. Jacobsen

dj@decision3.com

Lars Kai Hansen

lkh@imm.dtu.dk

Intelligent Signal Processing, Informatics and Mathematical Modeling,

Technical University of Denmark, Lyngby, N/A 2800, Denmark

Kristoffer Hougaard Madsen

khm@imm.dtu.dk

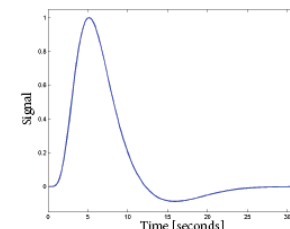
Intelligent Signal Processing, Informatics and Mathematical Modeling,

Technical University of Denmark, Lyngby, N/A 2800, Denmark

Neural Computation **20**, 738–755 (2008)

Bayesian averaging with split-1/2 resampling loop
to establish generalizability and reproducibility

$$R(M) = -\frac{1}{K} \sum_{i=1}^K \int p(\theta | D_i^1, M) \log \frac{p(\theta | D_i^1, M)}{p(\theta | D_i^2, M)} d\theta,$$



Balloon model: Non-linear relations between stimulus and physiology – described by four non-linear differential eqs.

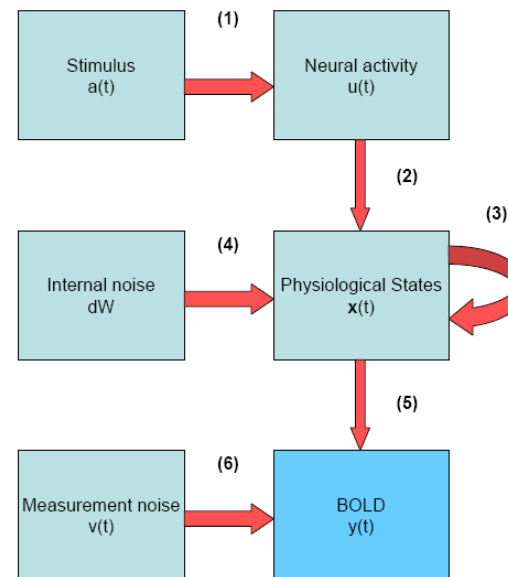


Figure 3.1: Overview diagram of hemodynamic models.

BOLD hemodynamics R-Bayes model selection

Model A: constant input vs Model B: Fading input

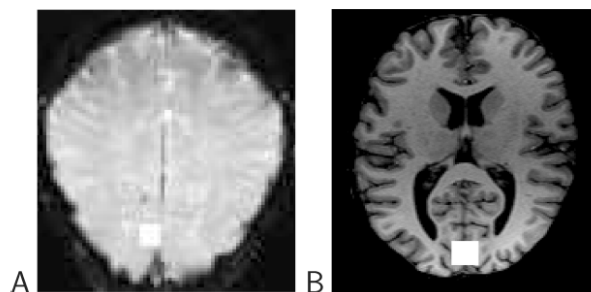


Figure 3: Regions of interest, marked with white squares. (A) Data set 1; T2* weighted image slice parallel to the calcarine sulcus. (B) Data set 2; MPRAGE (magnetization prepared rapid gradient echo) horizontal slice.

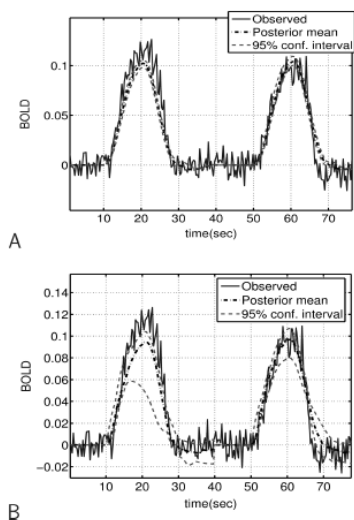


Figure 4: Prediction of data set 1. (A) Model A. (B) Model B. Note that the confidence interval is an empirical confidence interval for the mean prediction, based on the MCMC samples.

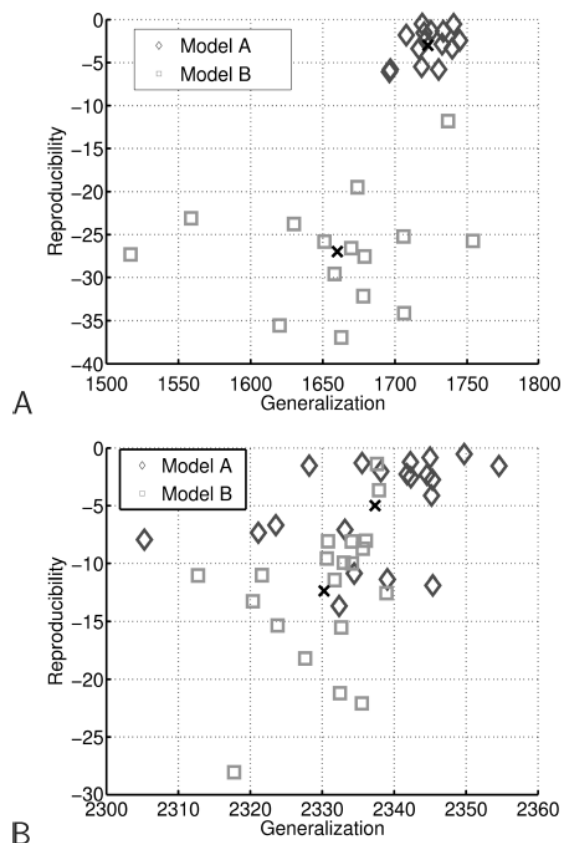


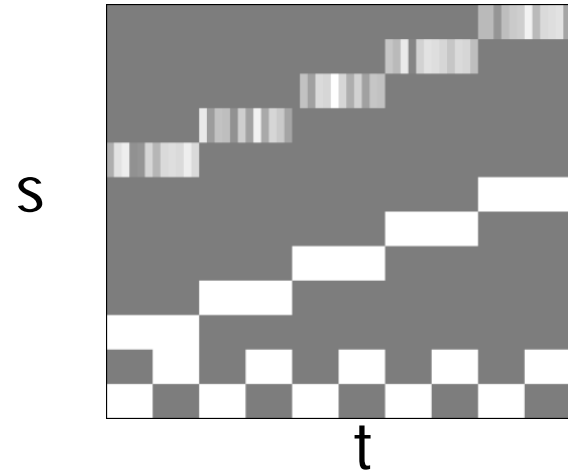
Figure 6: Generalization and reproducibility for real data; crosses mark the mean. (A) Data set 1. (B) Data set 2.

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".....



Multivariate neuroimaging models

- Univariate models -SPM, fMRI time series models etc.

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



- Multivariate models -PCA, ICA, SVM, ANN (Lautrup et al., 1994, Mørch et al. 1997)

$$p(x, s) = p(s | x)p(x)$$

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

$$p(s | x) \square p(s | x, D) \square p(s | x, \hat{\theta}), \quad \hat{\theta} = \hat{\theta}(D)$$

$$p(x) \square p(x | D) \square \int p(x | s, \hat{\theta}) p(s | \hat{\theta}) ds,$$

Generalizability

Do not multiply causes!

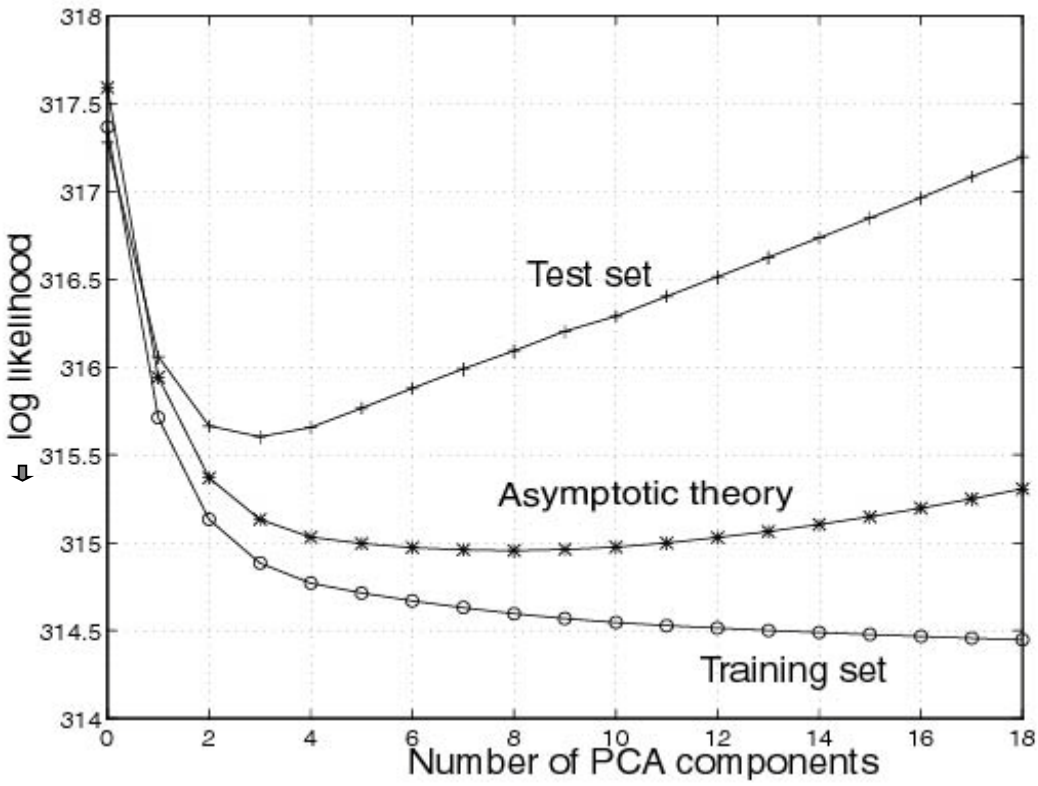


- Generalizability is defined as *the expected performance on a random new sample*
 - A model's mean performance on a "fresh" data set is an unbiased estimate of generalization
- Typical loss functions:

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

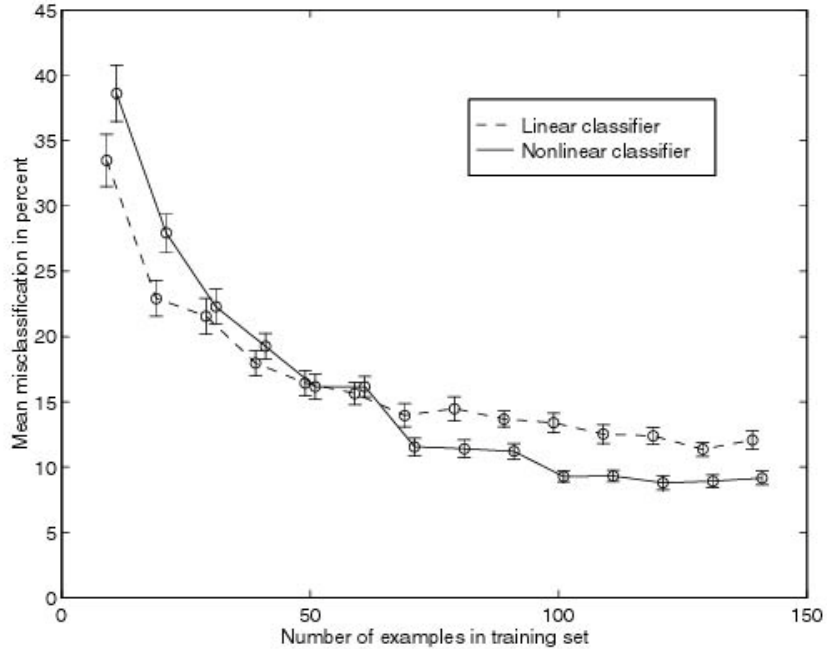
- Note: No problem to estimate generalization in hidden variable models!
- Results can be presented as "bias-variance trade-off curves" or "learning curves"

Bias-variance trade-off as function of PCA dimension

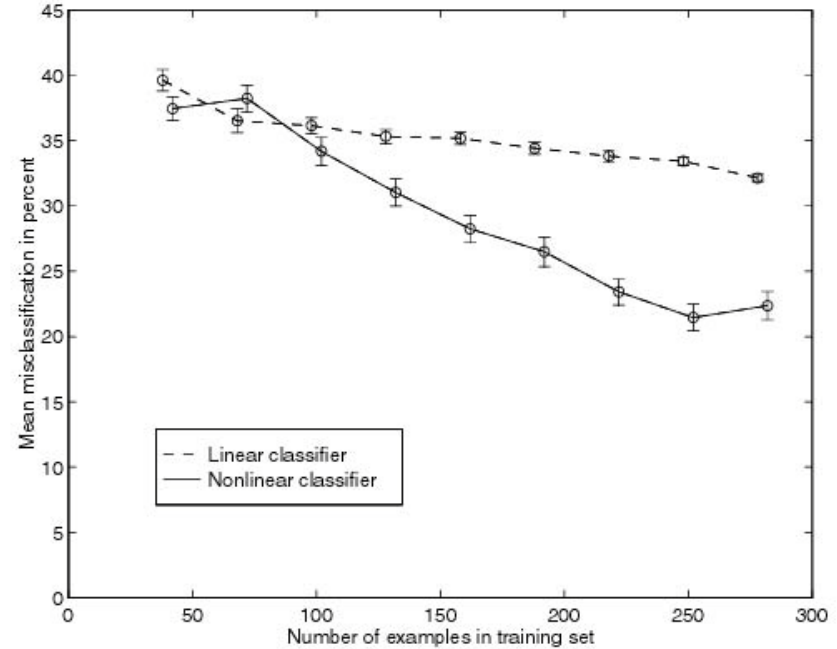


Hansen et al. *NeuroImage* (1999)

Learning curves for multivariate brain state decoding



PET



fMRI

Finger tapping, analysed by PCA dimensional reduction and Fisher LD / ML Perceptron. Mørch et al. *IPMI* (1997)...“first brain state decoding in fMRI”

Visualization of brain state decoder representation

- A brain map is a visualization of the information captured by the model:
 - The map should take on a high value in voxels/regions involved in the response and a low value in other regions...
- Statistical Parametric Maps
- The saliency map
- The sensitivity map
- Consensus maps

...hints from asymptotic theory

Linear unlearning for cross-validation

Lars Kai Hansen and Jan Larsen

CONNECT, Electronics Institute B349, Technical University of Denmark, DK-2800 Lyngby, Denmark
E-mail: lkhanse,jlarsen@ei.dtu.dk

- Asymptotic theory investigates the sampling fluctuations in the limit $N \rightarrow \infty$
- Cross-validation good news: The ensemble average predictor is equivalent to training on all data (Hansen & Larsen, 1996)
- Simple asymptotics for parametric and semi-parametric models
- Some results for non-parametric e.g. kernel machines
- In general: Asymptotic predictive performance has bias and variance components, there is proportionality between parameter fluctuation and the variance component...

The sensitivity map

NeuroImage 15, 772-786 (2002)
doi:10.1006/nimg.2001.1033, available online at <http://www.idealibrary.com> on IDEAL®

The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves

U. Kjems,*¹ L. K. Hansen,* J. Anderson,^{†‡} S. Frutiger,^{‡§} S. Muley,[§]
J. Sidtis,[§] D. Rottenberg,^{†‡§} and S. C. Strother^{†‡§¶}

*Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark; [†]Radiology Department,
[§]Neurology Department, and [¶]Biomedical Engineering, University of Minnesota, Minneapolis, Minnesota 55455;
and [‡]PET Imaging Center, VA Medical Center, Minneapolis, Minnesota 55417

$$m_j = \left\langle \left(\frac{\partial \log p(s|x)}{\partial x_j} \right)^2 \right\rangle$$

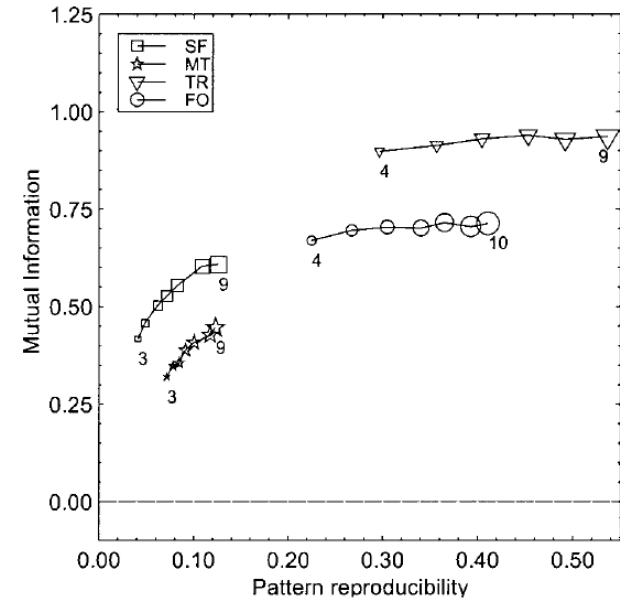
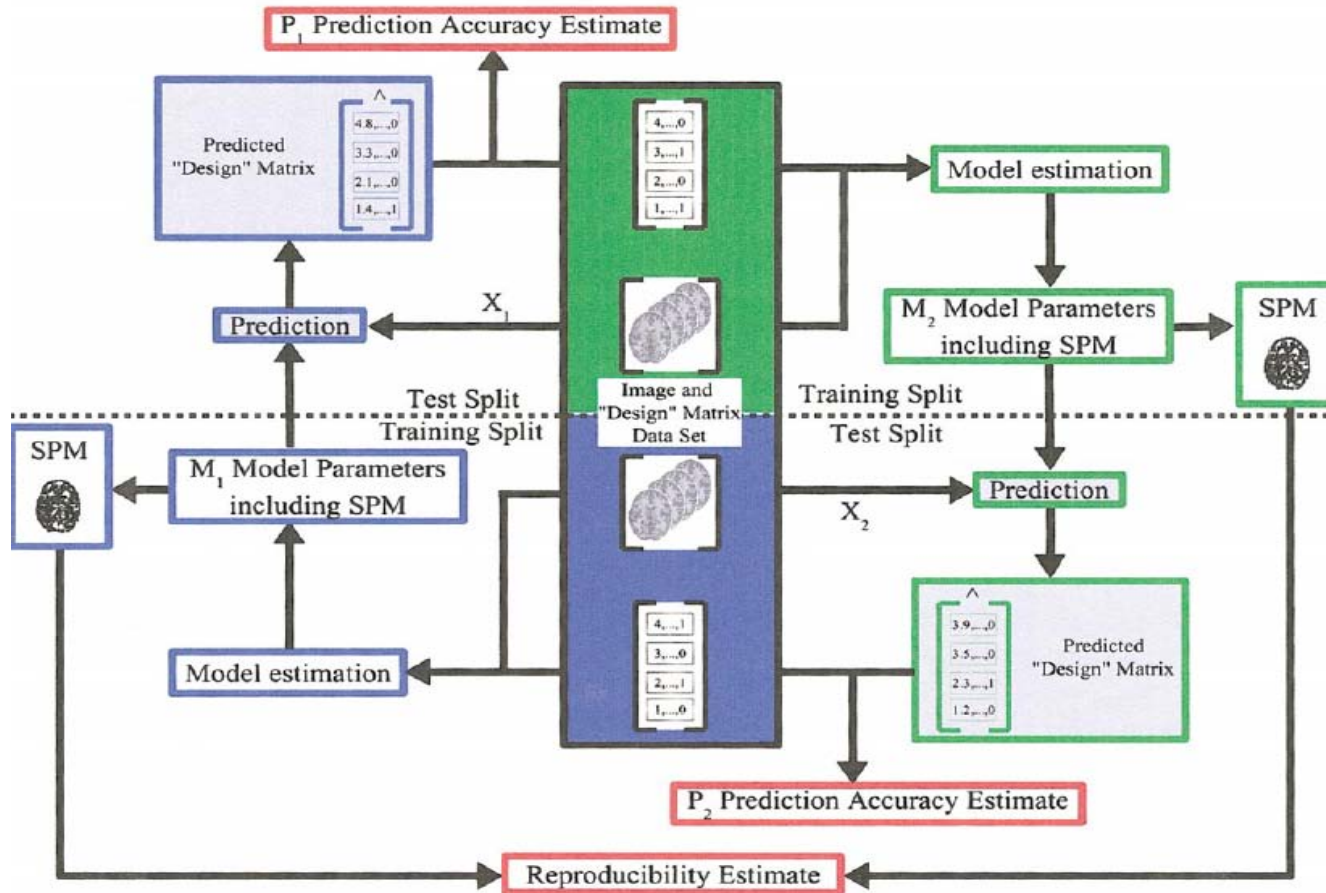


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both measures indicate improved performance of the model as the number of subjects increases.

- The sensitivity map measures the impact of a specific feature/location on the predictive distribution

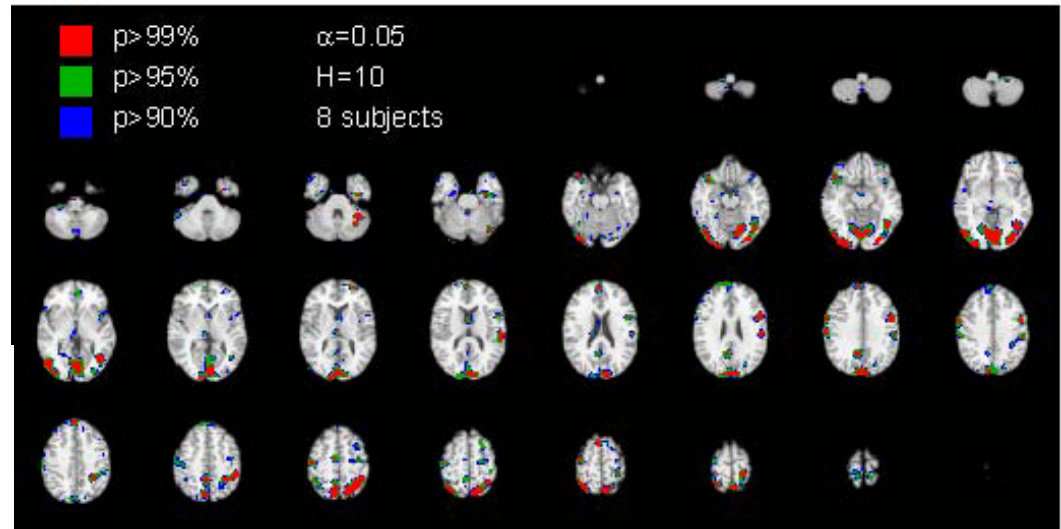
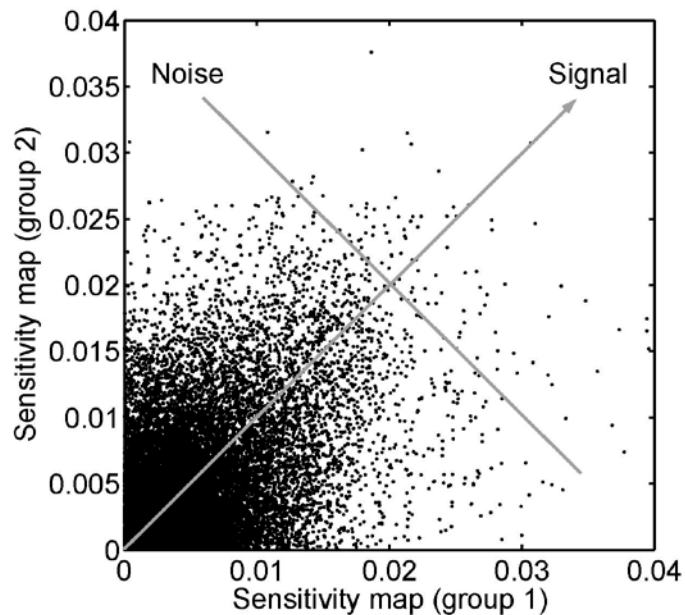
NPAIRS: Reproducibility of parameters



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

Reproducibility of internal representations

Predicting applied static force
with visual feed-back



Split-half resampling provides unbiased
estimate of reproducibility of SPMs

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),
Mondrup Rasmussen et al., submitted, 2009))

Unsupervised learning:

Factor analysis generative model

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

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

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

Source distribution:

PCA: ... normal

ICA: ... other

IFA: ... Gauss. Mixt.

kMeans: .. binary

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

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

S known: GLM

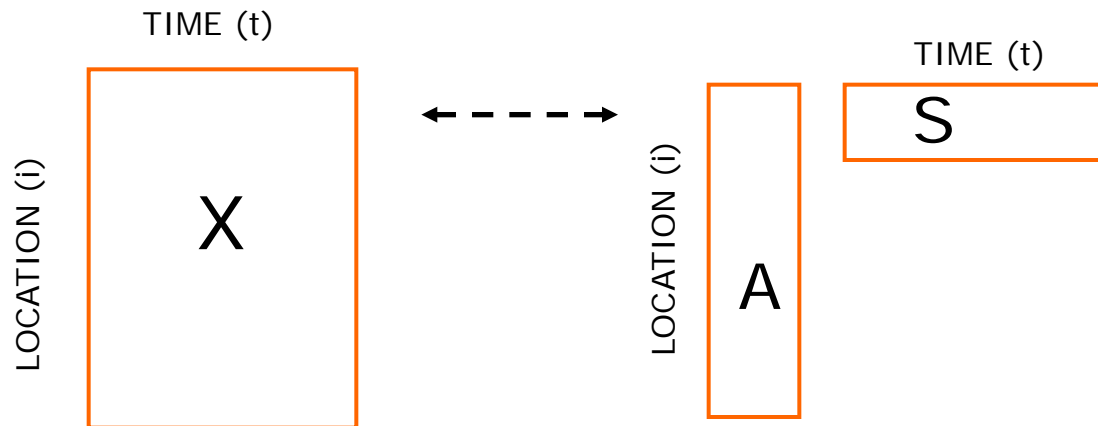
(1-A)⁻¹ sparse: SEM

S, A positive: NMF

Højten-Sørensen, Winther, Hansen,
Neural Computation (2002), Neurocomputing (2002)

Factor models

- Represent a datamatrix by a low-dimensional approximation
- Identify spatio-temporal networks of activation



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

Matrix factorization: SVD/PCA, NMF, Clustering

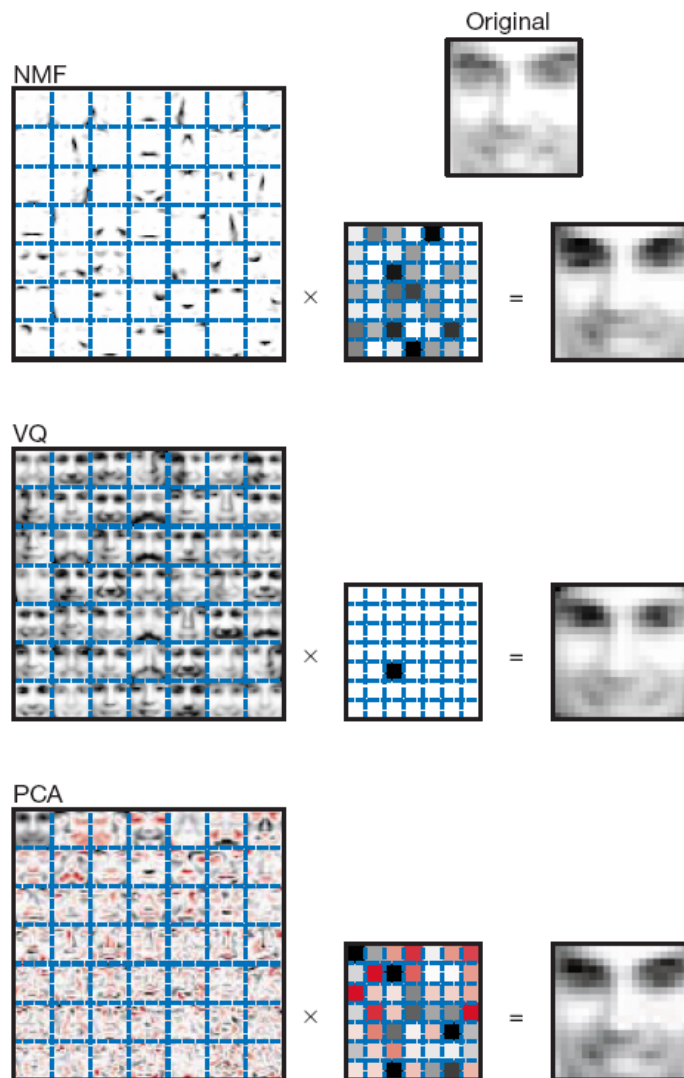


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2,429$ facial images, each consisting of $n = 19 \times 19$ 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

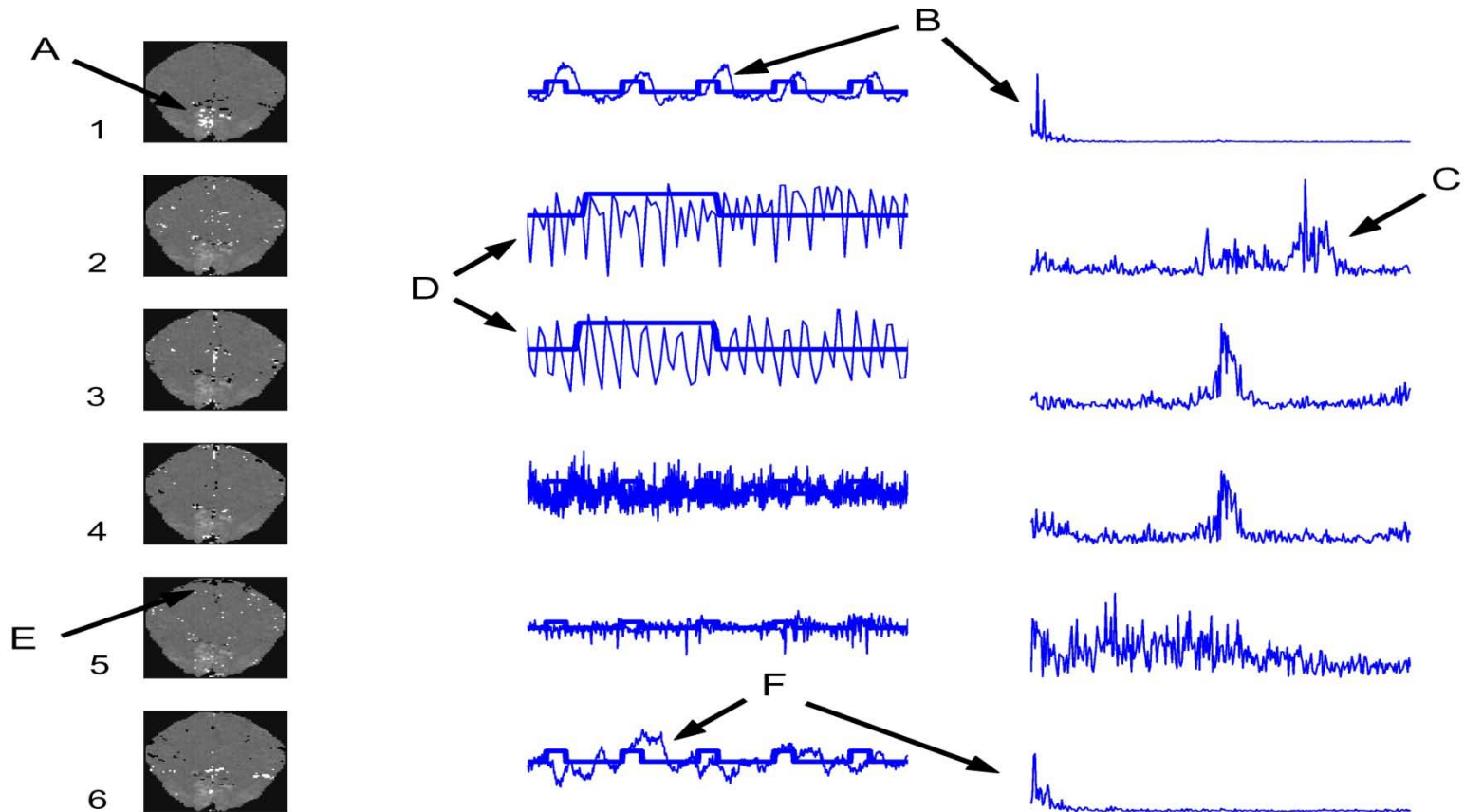
Daniel D. Lee* & H. Sebastian Seung*†

* 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 statistically independent



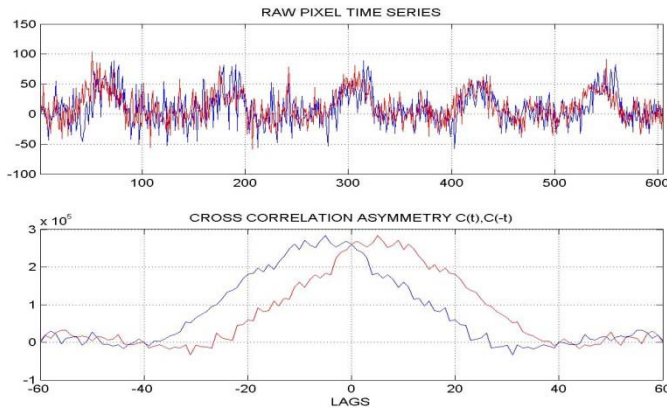
(McKeown, Hansen, Sejnowski, Curr. Op. in Neurobiology (2003))

Challenges for the linear factor model

- Too simple?
 - Temporal structure in networks -> Convolutional ICA
- Too rich and over-parametrized?
 - Multi-dimensional macro and micro variables (space/time/frequency, group study, repeat trials)
 - > Multiway methods

fMRI: Delayed activation in visual cortex

Two voxel temporal cross-correlation



$$\langle x(j,t)x(j',t+\tau) \rangle = \sum_{k,k'} A(j,k)A(j',k') \langle s(k,t)s(k',t+\tau) \rangle$$
$$\langle x(j,t)x(j',t+\tau) \rangle = \sum_k A(j,k)A(j',k)C_k(\tau)$$
$$\Sigma_{j,j'}^{xx}(\tau) = \Sigma_{j,j'}^{xx}(-\tau)$$

Recovery of source signals from an unknown, linear convolutive mixture

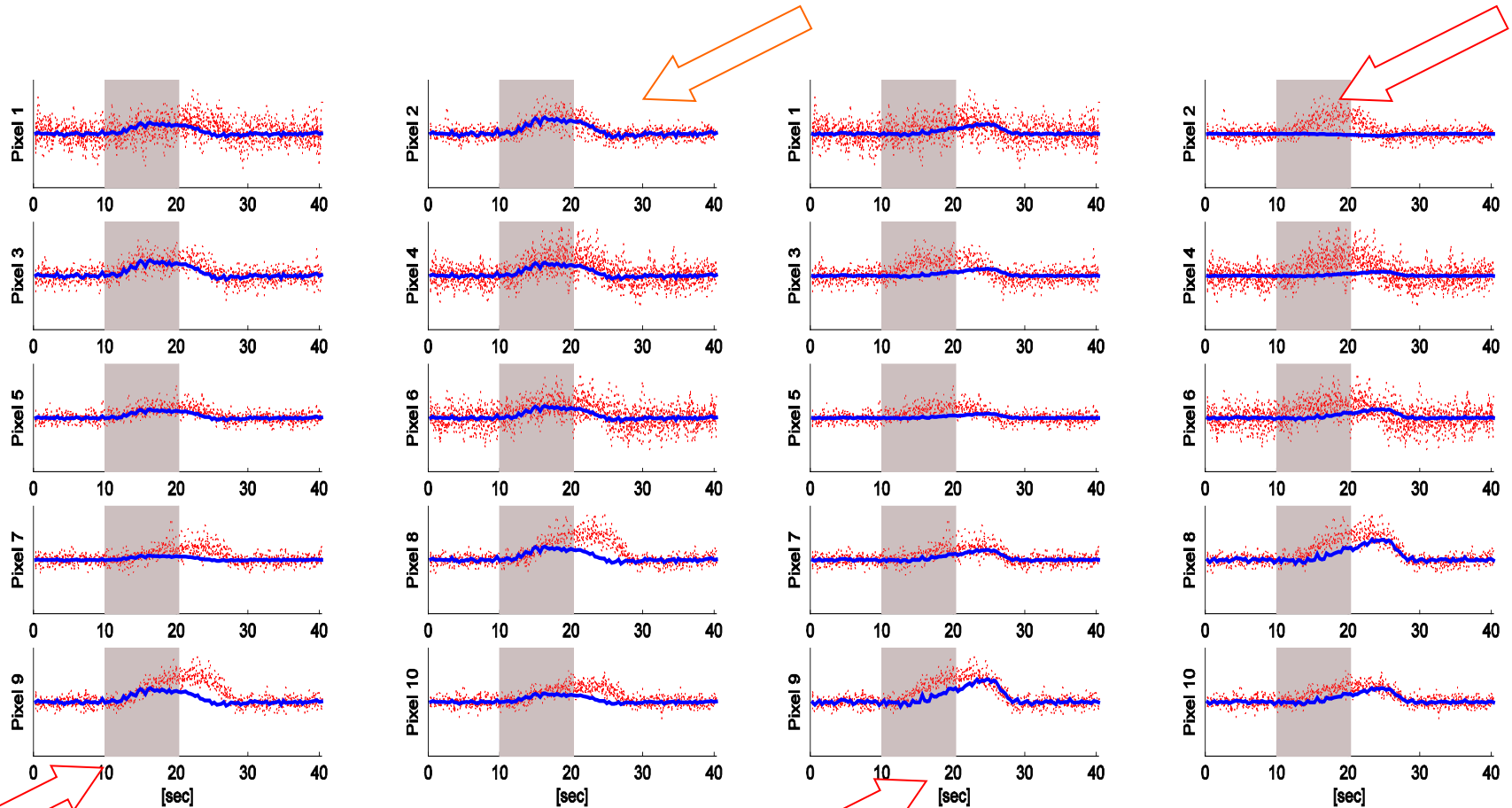
$$x(j,t) = \sum_{k,\tau} A(j,k,\tau) s(k,t-\tau)$$

Hansen & Dyrholm: Prediction matrix (Proc MLSP, 2003)

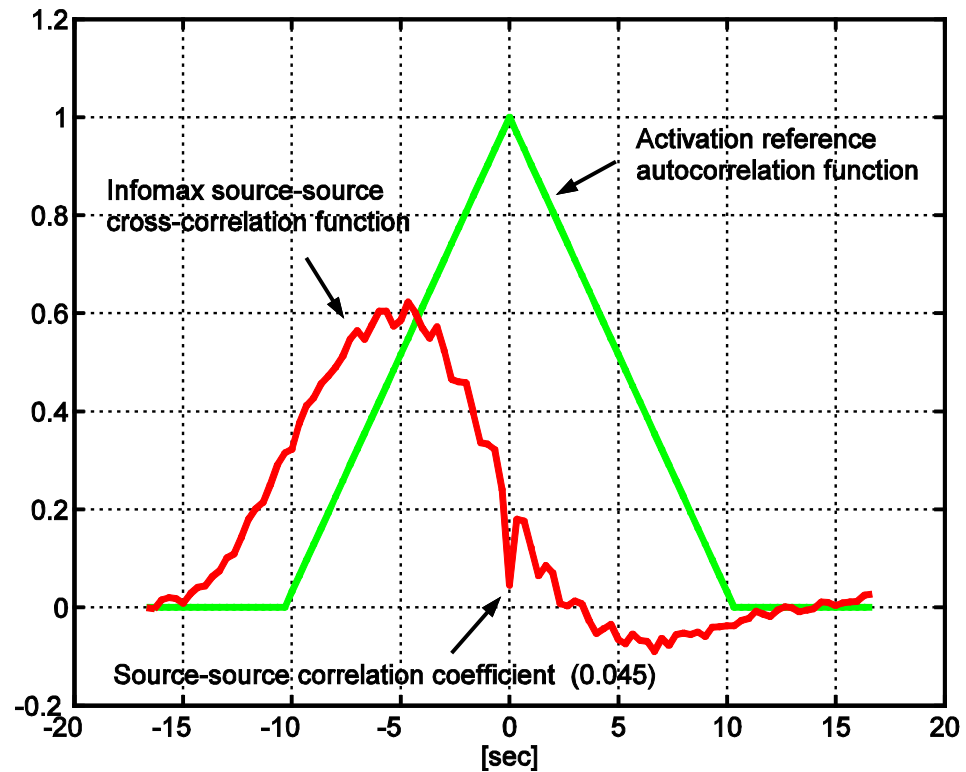
Olsson & Hansen: Kalman (Proc NIPS 2004), J. Mach. Learning Res. (2006)

Dyrholm & Hansen: "CICAAR" (Proc ICA, 2004), Neural Comp (2007)

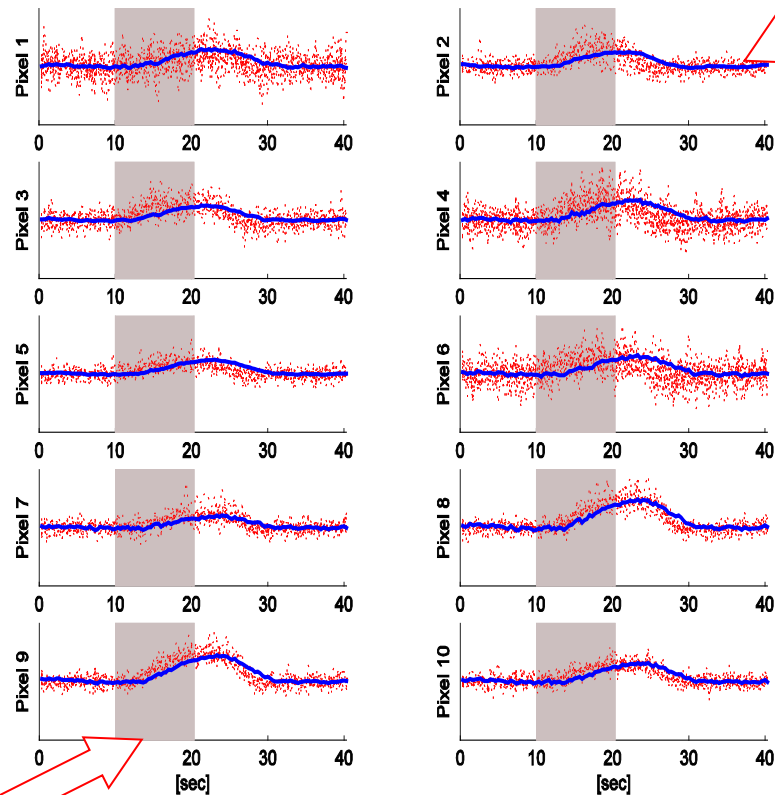
Instantaneous mixing: Components 1,2 (30%+30% of variance)



Instantaneous mixing cross correlation between components



Convulsive mixing (60% of total variance)



Convulsive ICA is a more appropriate mixing model for fMRI

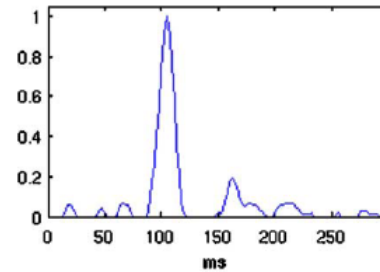
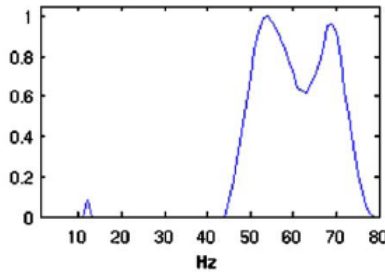
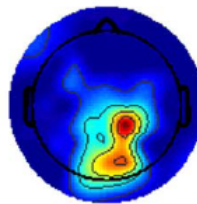
Data represented as multiway arrays

$$\begin{array}{c}
 \text{[Grey Box]} = \sum_{\lambda=1}^F \mathbf{a}_{\lambda} \mathbf{s}_{\lambda} \\
 x_{i_1 i_2} = \sum_{\lambda=1}^F a_{i_1 \lambda} s_{i_2 \lambda} + e_{i_1 i_2} \\
 \text{Factor Analysis}
 \end{array}
 \qquad
 \begin{array}{c}
 \text{[Stack of Grey Boxes]} = \sum_{\lambda=1}^F \mathbf{a}_{\lambda} \mathbf{d}_{\lambda} \mathbf{s}_{\lambda} \\
 x_{i_1 i_2 i_3} = \sum_{\lambda=1}^F a_{i_1 \lambda} d_{i_2 \lambda} s_{i_3 \lambda} + e_{i_1 i_2 i_3} \\
 \text{PARAFAC}
 \end{array}$$

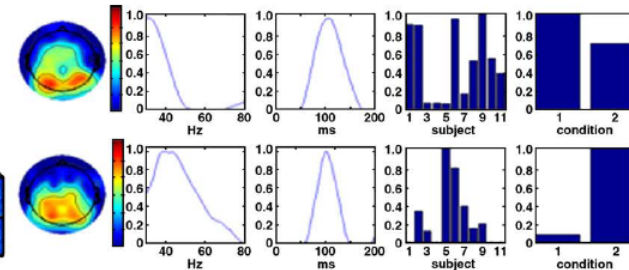
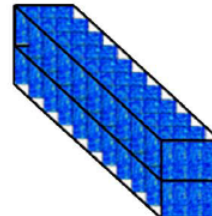
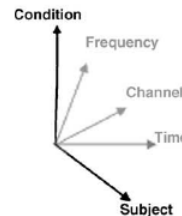
EEG visual response to meaningful vs non-meaningful drawings (N=11).

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.

3-way analysis:
Channel*freq*time



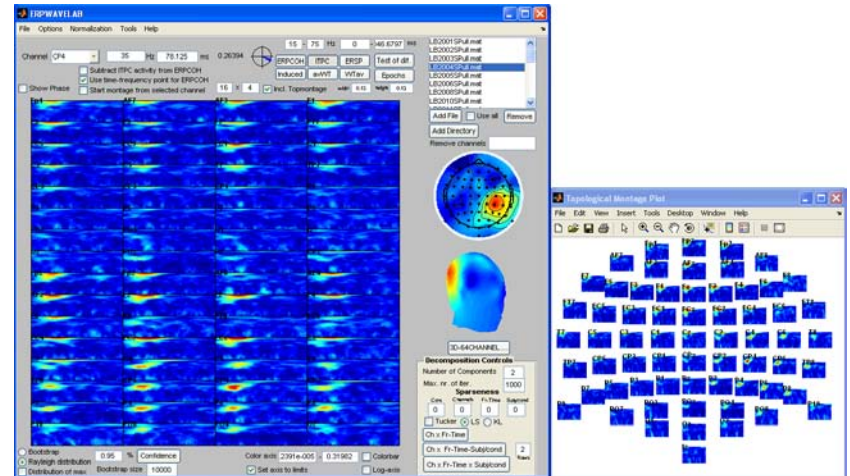
5-way analysis:
Channel*freq*time*subject*condition



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

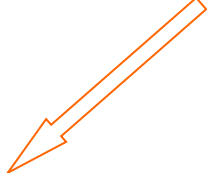
Mørup et al. J. Neuroscience Methods (2007),

Generalizable supervised models

- Non-linear kernel machines, SVM

$$s(n') \approx \sum_{n=1}^N \alpha(n) K(x_n, x_{n'})$$

Local voting +/-



$$K(x_n, x_{n'}) = \exp\left\{-\frac{\|x_n - x_{n'}\|^2}{2c}\right\}$$

Visualization of SVM learning from fMRI

- Visualization of kernel machines
 - How to create an SPM for a kernel machine
 - The sensitivity map for kernels
 - Example:

$$s(n') \approx \sum_{n=1}^N \alpha(n) K(x_n, x_{n'})$$

$$K(x_n, x_{n'}) = \exp \left\{ -\frac{\|x_n - x_{n'}\|^2}{2c} \right\}$$

Visualization of kernel machine internal representations

1000

IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 5, SEPTEMBER 1999

Input Space Versus Feature Space in Kernel-Based Methods

Bernhard Schölkopf, Sebastian Mika, Chris J. C. Burges, Philipp Knirsch,
Klaus-Robert Müller, Gunnar Rätsch, and Alexander J. Smola



NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 26 (2005) 317–329

Support vector machines for temporal classification of block design fMRI data

Stephen LaConte,^a Stephen Strother,^b Vladimir Cherkassky,^c Jon Anderson,^b and Xiaoping Hu^{a,*}

The Pre-Image Problem in Kernel Methods

James T. Kwok
Ivor W. Tsang
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon,
Hong Kong

JAMESK@CS.UST.HK
IVOR@CS.UST.HK

- Existing visualization methods
 - Pre-image (Mika et al., NIPS 1998, Schölkopf et al., 1999)
Basically an ill-defined objective, useful for denoising
 - Multi-dimensional scaling (Kwok & Tsang, ICML 2003)
Interpolates nearest neighbors, suffers in high dimensions

Problem: Existing methods provide local visualization, which point should be visualized?. Algorithms are reported unstable.

The sensitivity map

NeuroImage 15, 772-786 (2002)

doi:10.1006/nimg.2001.1033, available online at <http://www.idealibrary.com> on IDEAL®

The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves

U. Kjems,*¹ L. K. Hansen,* J. Anderson,^{†‡} S. Frutiger,^{‡§} S. Muley,[§]
J. Sidtis,[§] D. Rottenberg,^{†‡§} and S. C. Strother^{†‡§¶}

*Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark; [†]Radiology Department,

[§]Neurology Department, and [¶]Biomedical Engineering, University of Minnesota, Minneapolis, Minnesota 55455;

and [‡]PET Imaging Center, VA Medical Center, Minneapolis, Minnesota 55417

$$m_j = \left\langle \left(\frac{\partial \log p(s|x)}{\partial x_j} \right)^2 \right\rangle$$

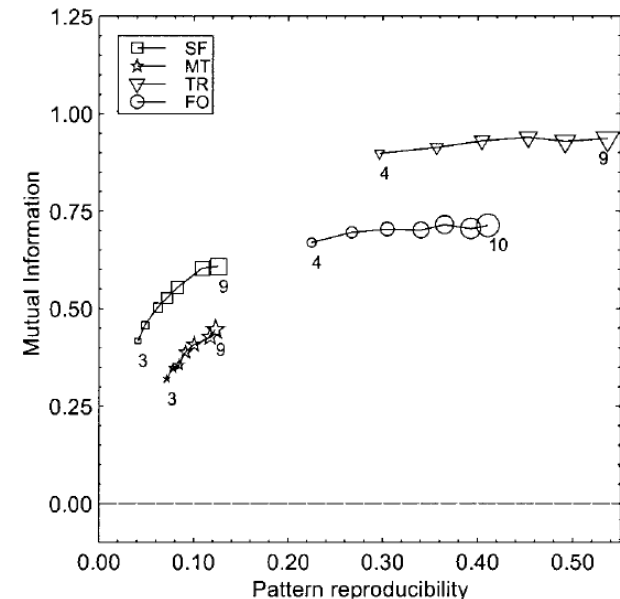
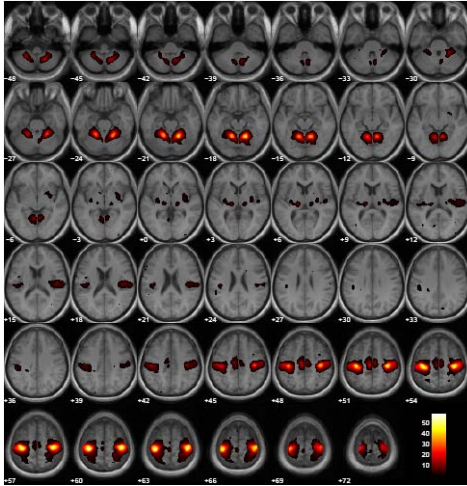


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both measures indicate improved performance of the model as the number of subjects increases.

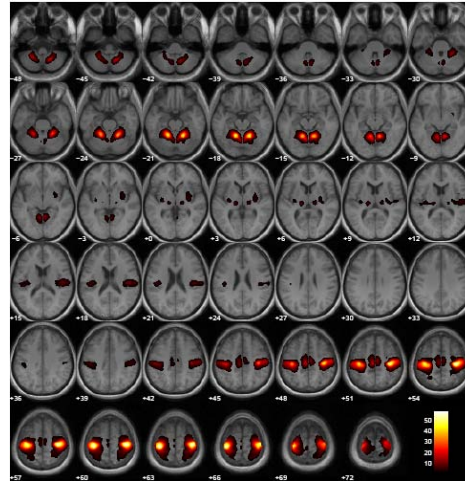
- The sensitivity map measures the impact of a specific feature/location on the predictive distribution

Consistency across models (left-right finger tapping)

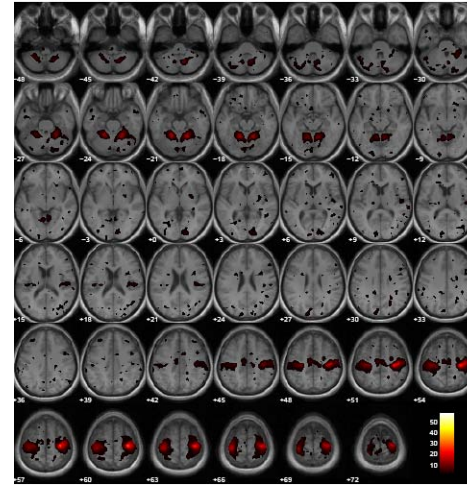
LogReg



SVM



RVM



Sparsity increasing

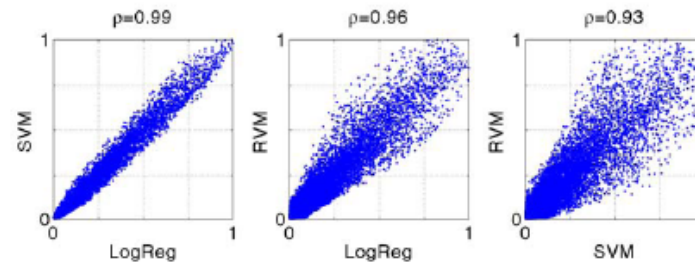


Figure 7: fMRI fingertapping experiment - consensus analysis. The plots show the extent of consensus in the average rSPI among the three models. The rSPI for LogReg was scaled by its maximum value. Hereafter the rSPIs from the SVM and RVM were transformed to match the histogram of that of LogReg. Correlation coefficients between histograms are found on top of the plots.

Sensitivity maps for non-linear kernel regression

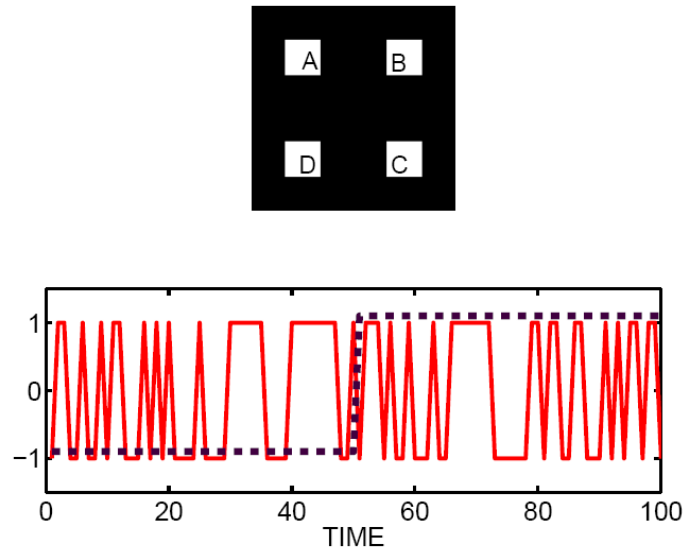


Fig. 1. XOR-image set define by four activated regions (A,B,C,D). Initially we let regions (A,B,D) be activated by random sequence taking values ± 1 , as shown in example in the bottom panel (full curve). The target signal, also taking values $t_n = pm1$, and is also indicated in the bottom panel (dashed line). The region (C) is activated with an XOR-sequence relative to (A) and t_n , so that $C_n = A_n * t_n$, hence, in the active state the two regions (A,C) are randomly, but identically activated, while in the resting condition, they are random, but opposite

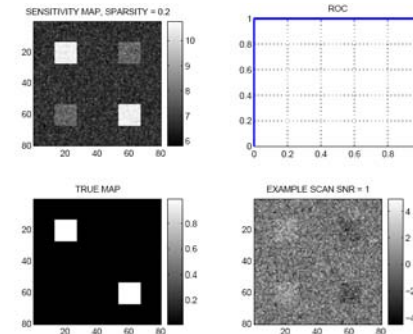


Fig. 2. XOR-image set define by four activated regions. The results of analyzing a image set with $N = 400$ examples. The image signal-to-noise ratio is $SNR = 1$, i.e., the additive noise is unit variance. The target function has in addition been contaminated by 10% random label noise. The four subplots show: The sensitivity map (upper left), the near-perfect receiver operating curve (ROC, upper right), the true activation map (lower left), and a random example of the simulated brain images. We modeled the data set using the kernel regression method. The linear model was estimated using the so-called least angle elastic net method (LARSEN) with a degree of sparsity of 0.2, i.e., using $N = 0.2 \times 400 = 80$ support vectors.

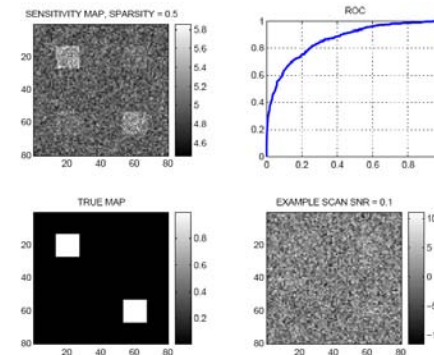
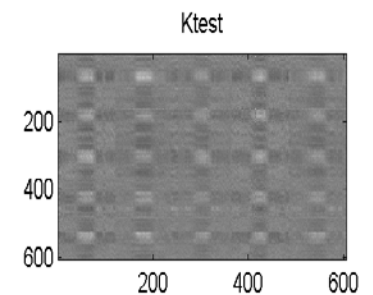
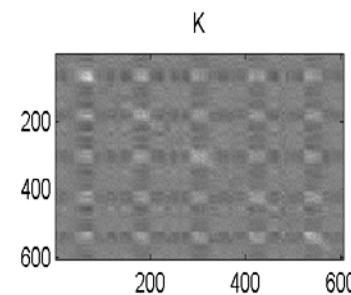
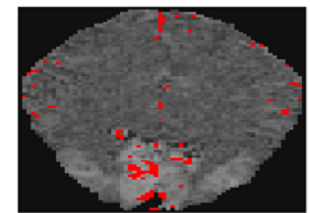
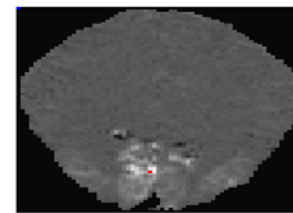
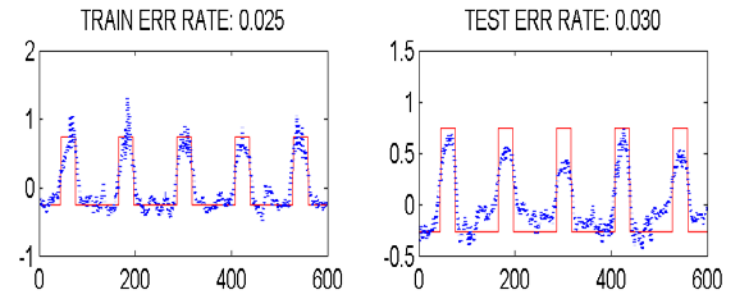


Fig. 4. XOR-image set define by four activated regions. Similar to figure 2, however the image signal-to-noise ratio is $SNR = 0.1$.

Initial dip data: Visual stimulus (TR 0.33s)

- Gaussian kernel, sparse kernel regression
- Sensitivity map computed for whole slice
- Error rates about 0.03
- How to set
 - Kernel width?
 - Sparsity?

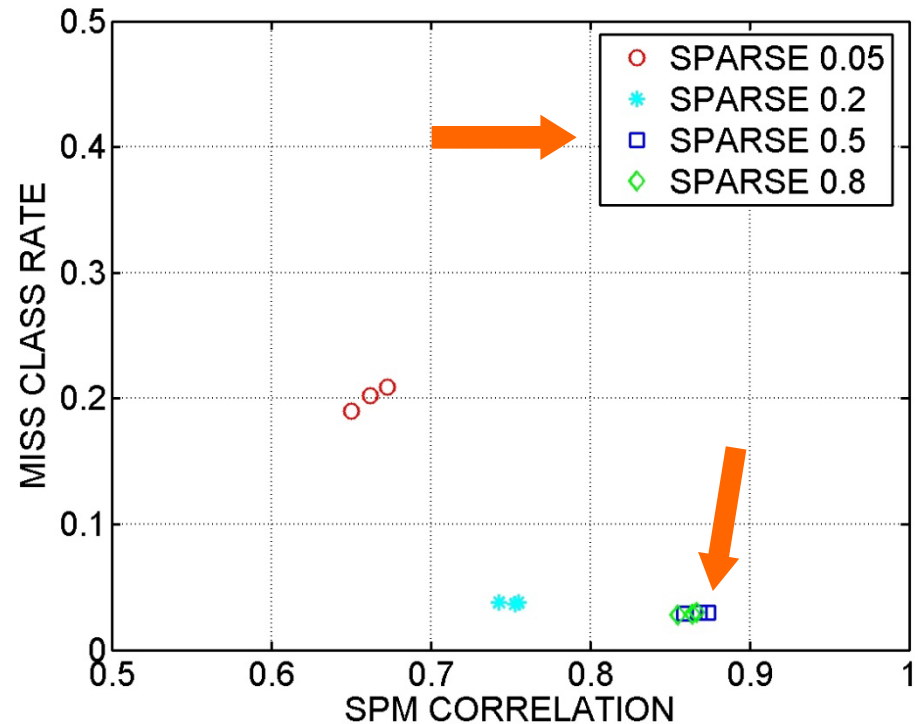


$$s(n') \approx \sum_{n=1}^N \alpha(n) K(x_n, x_{n'})$$

$$K(x_n, x_{n'}) = \exp\left\{-\frac{(x_n - x_{n'})^2}{2c}\right\}$$

Initial dip data: Visual stimulus (TR 0.33s)

- Select hyperparameters of kernel machine using NPAIRS resampling
 - Degree of sparsity
 - Kernel width, localization of map



Conclusion

- Machine learning in brain imaging has two equally important objectives
 - Generalizability
 - SPM reproducibility
- Can visualize general brain state decoders maps with perturbation based methods (saliency maps, sensitivity maps etc)
- NPAIRS split-half based framework for optimization of generalizability and robust visualizations
- Still some hope for BOLD deconvolution!



Illustration of
HypoSafe implantable device



Acknowledgments

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

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

