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,

.

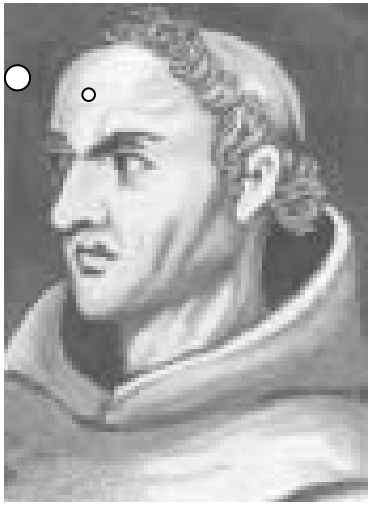




Cimbi
Center for integrated molecular brain imaging

DTU

Do not multiply causes!



OUTLINE

- Machine learning –the doubble agenda
  - Aim I: To abstract generalizable relations from data
  - Aim II: Robust interpretation / visualization
- Unsupervised (explorative)
  - Factor models - Linear hidden variable representations
  - Independent component analysis (ICA)
  - Generalizations: Convolutive mixing, shift
  - Multiway modeling
- Supervised models (retrieval)
  - Visualization of non-linear kernel machines

Recent reviews

ARTICLE IN PRESS
YNIMG-07534; No. of pag
NeuroImage xxx (2010) xxx–xxx
Contents lists available at ScienceDirect
NeuroImage
journal homepage: www.elsevier.com/locate/ynimg

Review

Encoding and decoding in fMRI

Thomas Naselaris [a], Kendrick N. Kay [b], Shinji Nishimoto [a], Jack L. Gallant [a,b,*]

[a] Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA
[b] Department of Psychology, University of California, Berkeley, CA 94720, USA

NeuroImage xxx (2010) xxx–xxx
Contents lists available at ScienceDirect
NeuroImage
journal homepage: www.elsevier.com/locate/ynimg

Decoding fMRI brain states in real-time

Stephen M. LaConte

Department of Neuroscience, Baylor College of Medicine, One Baylor Plaza, T-115, Houston TX 77030, United States

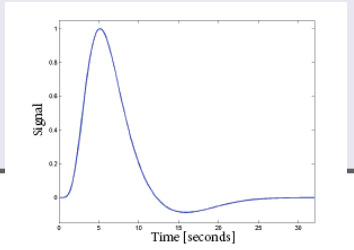
NEUROIMAGING

Decoding mental states from brain activity in humans

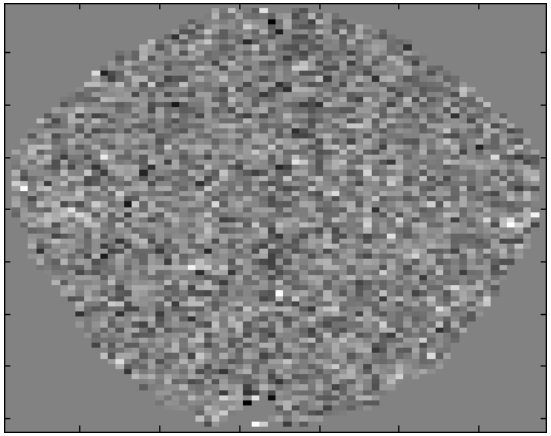
John-Dylan Haynes [*‡§] and Geraint Rees* [‡§]

Abstract | Recent advances in human neuroimaging have shown that it is possible to accurately decode a person's conscious experience based only on non-invasive measurements of their brain activity. Such 'brain reading' has mostly been studied in the domain of visual perception, where it helps reveal the way in which individual experiences are encoded in the human brain. The same approach can also be extended to other types of mental state, such as covert attitudes and lie detection. Such applications raise important ethical issues concerning the privacy of personal thought.

Functional MRI



- 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”?



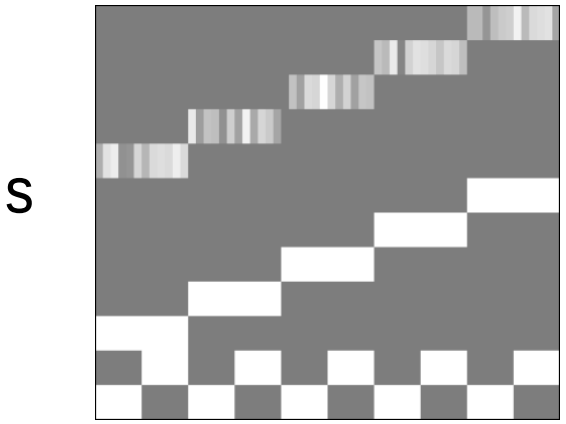
TR = 333 ms

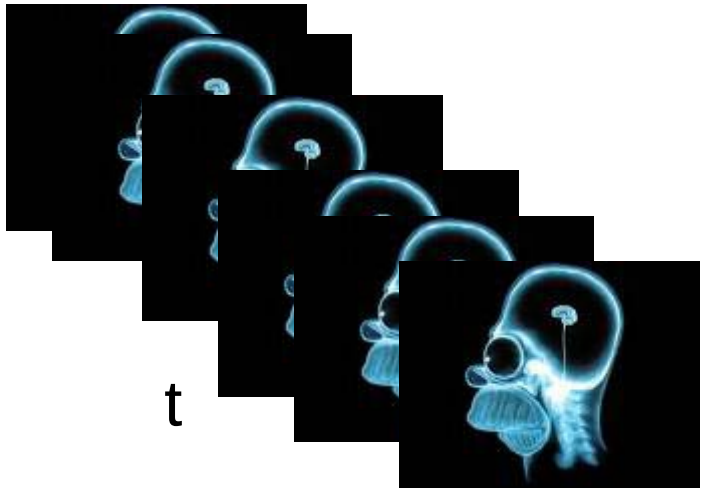
Multivariate neuroimaging models

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.

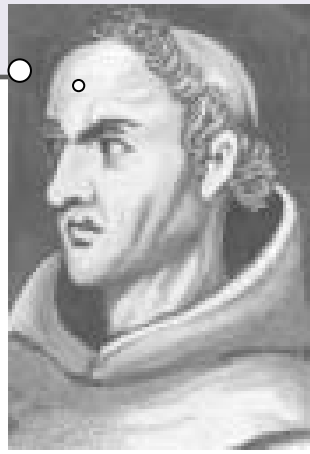
$$p(x,s) = p(x \mid s)p(s) = \prod_j p(x_j \mid s) \cdot p(s)$$

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

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

- Modeling from data with parameterized function families – rather than testing silly null hypotheses

AIM I: Generalizability



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

- Typical loss functions:

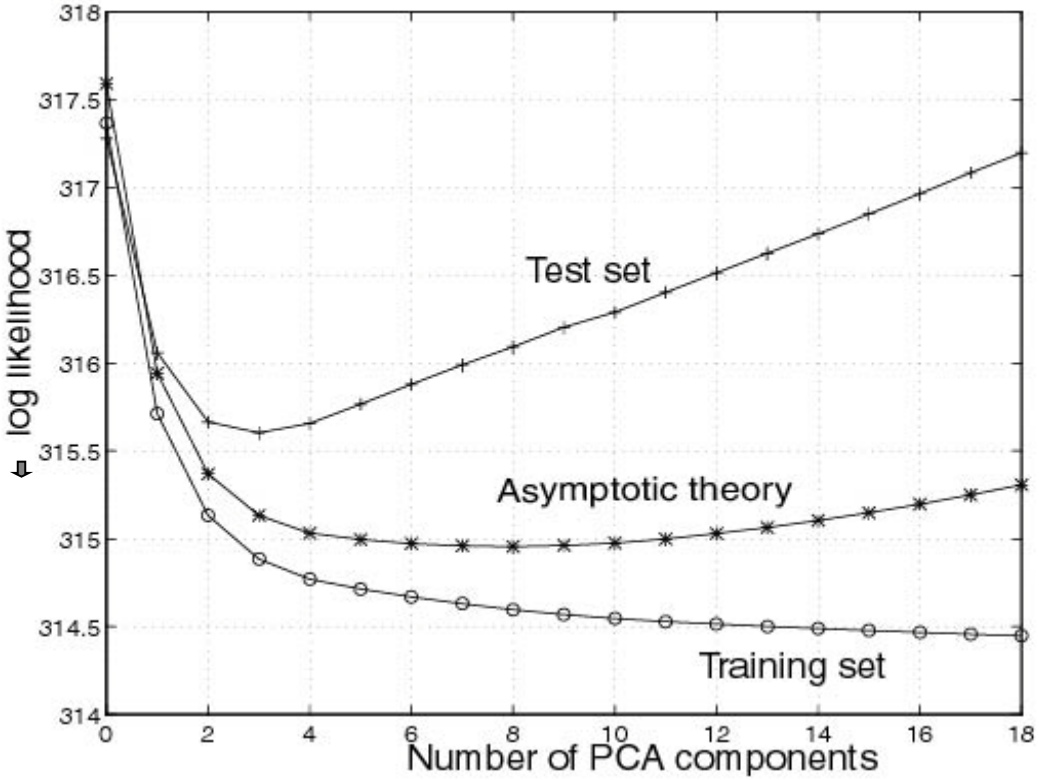
$$\left\langle -\log p(\mathbf{s}\,|\,\mathbf{x}, D)\right\rangle, \qquad \left\langle -\log p(\mathbf{x}\,|\,D)\right\rangle,$$

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

- Note: No problem to estimate generalization in hidden variable models!

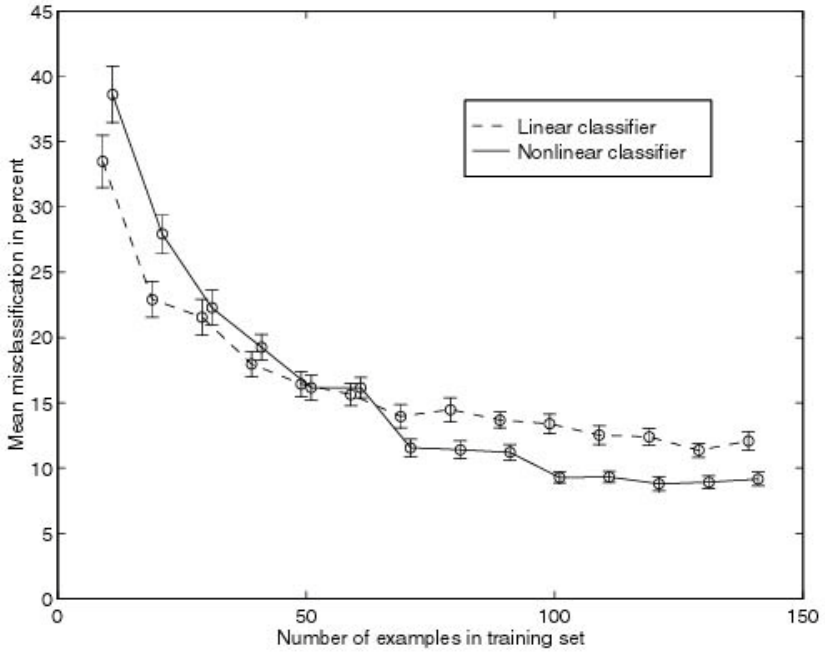
- Results can be presented as ”bias-variance trade-off curves” or ”learning curves”

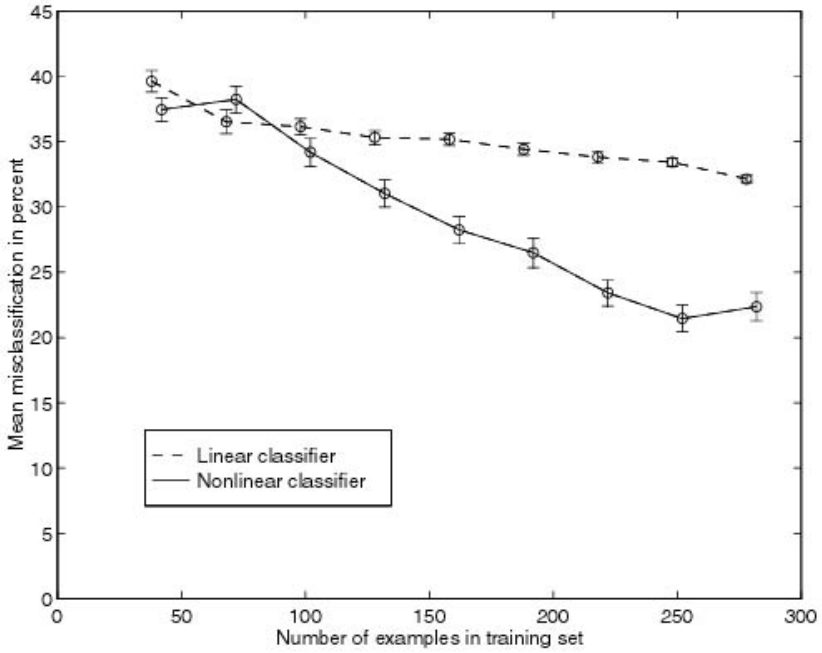
Bias-variance trade-off as function of PCA dimension



Hansen et al. *NeuroImage* (1999)

Learning curves for multivariate brain state decoding





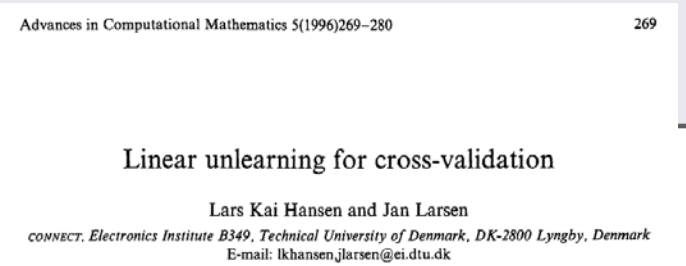
PET

fMRI

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

AIM II Interpretation: Visualization of networks

- 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
- Weight maps in linear models
- The saliency map
- The sentivity map
- Consensus maps



Advances in Computational Mathematics 5(1996)269–280 269

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: lkhansen,jlarsen@ei.dtu.dk

...hints from asymptotic theory

- Asymptotic theory investigates the sampling fluctuations in the limit N -> ∞
- 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

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

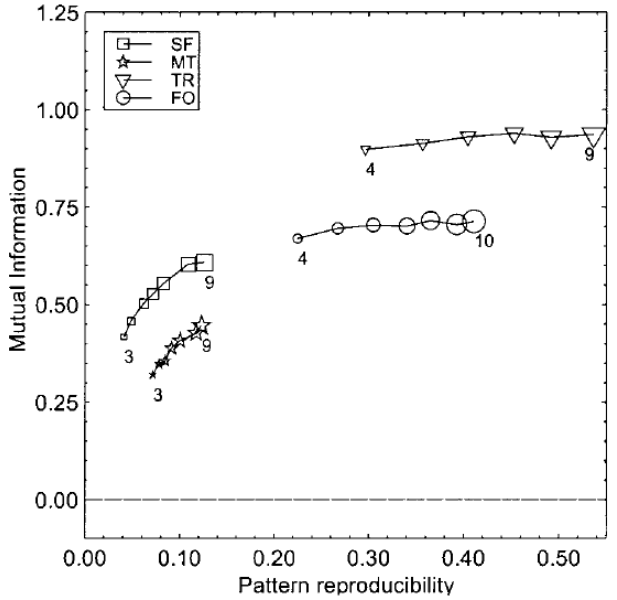
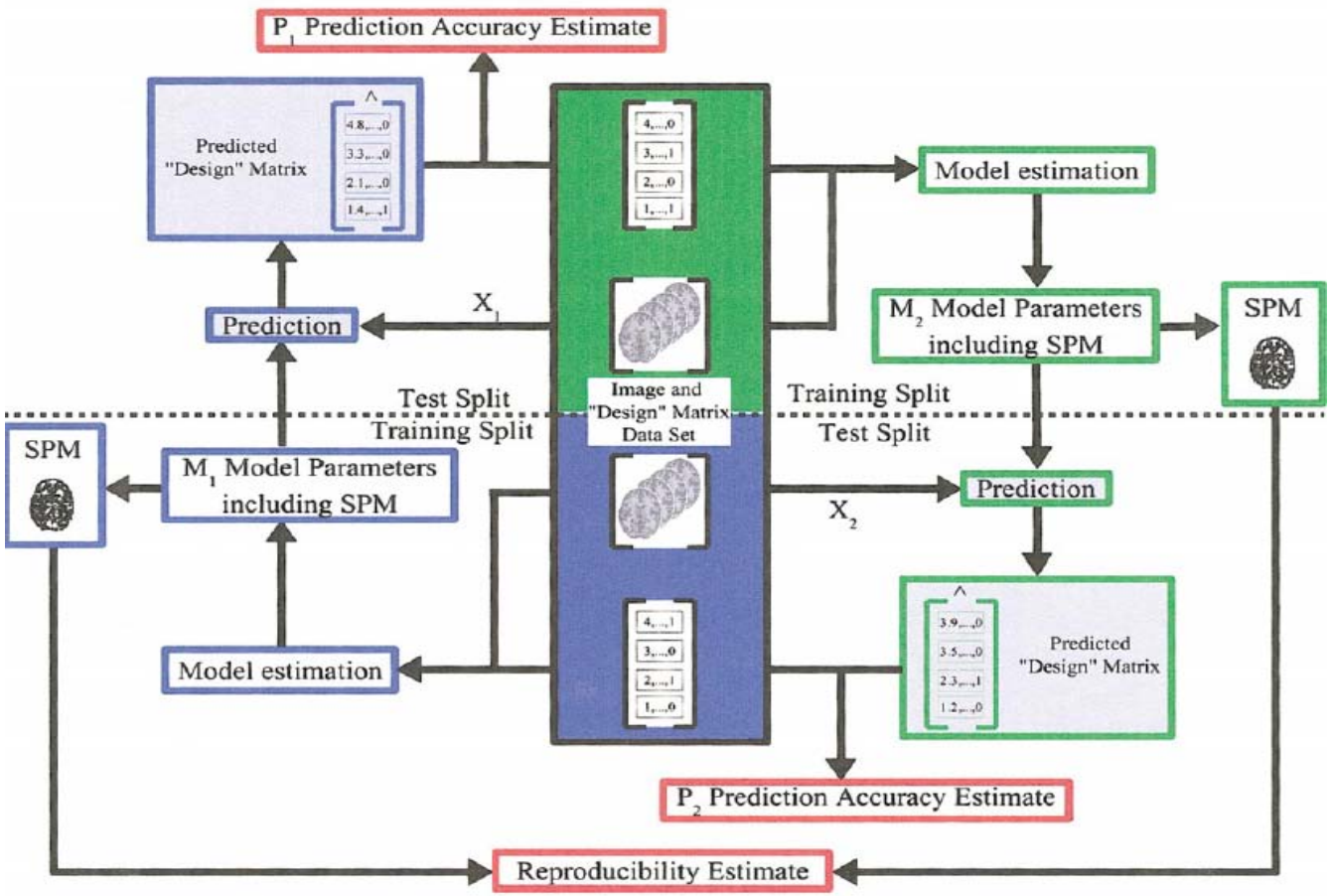


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both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

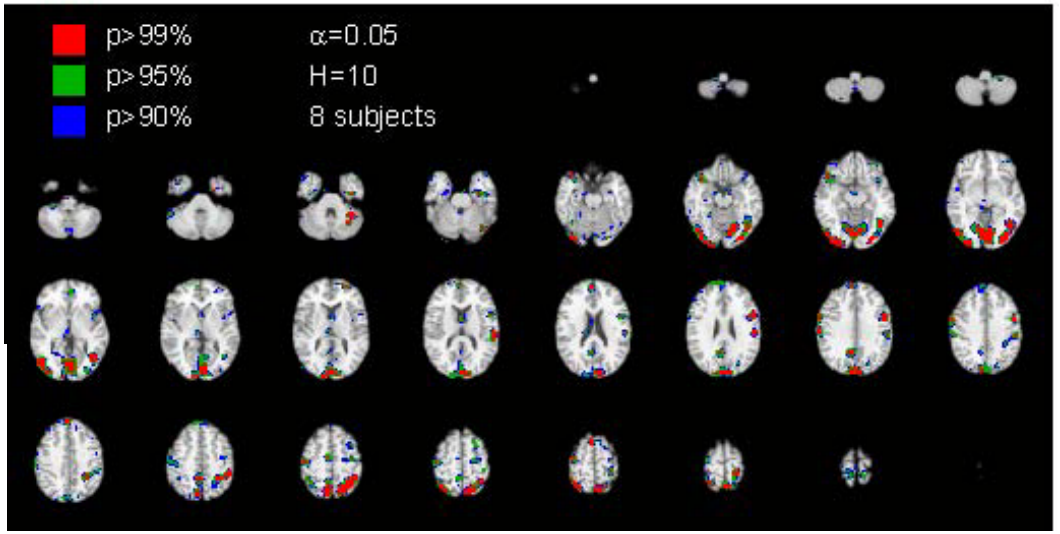
NPAIRS: Reproducibility of parameters

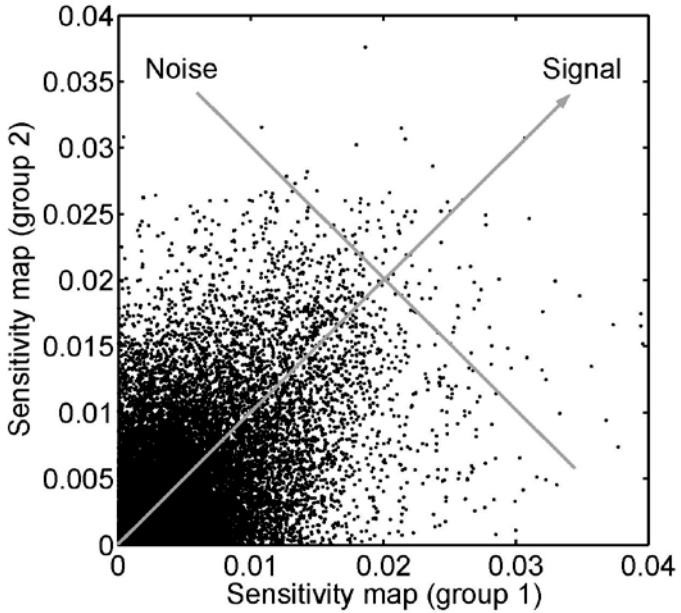


NeuroImage: Hansen et al (1999), Lange et al. (1999)
Hansen et al (2000), Strother et al (2002),
Kjems et al. (2002), LaConte et al (2003), Strother et al (2004)

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

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

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

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

Source distribution:
PCA: ... normal
ICA: ... other
IFA: ... Gauss. Mixt.
kMeans: .. binary

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

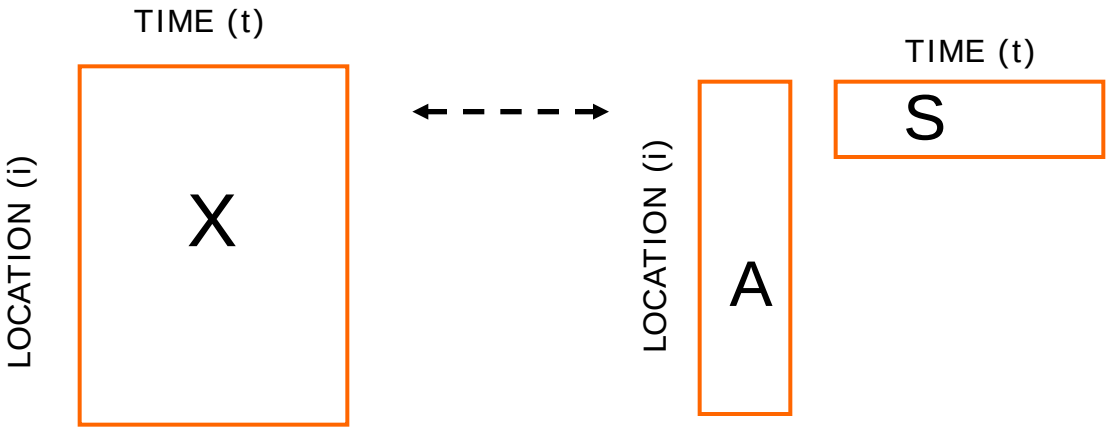
FA: $\boldsymbol{\Sigma} = \mathbf{D}$

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

Højen-Sørensen, Winther, Hansen,
Neural Compution (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

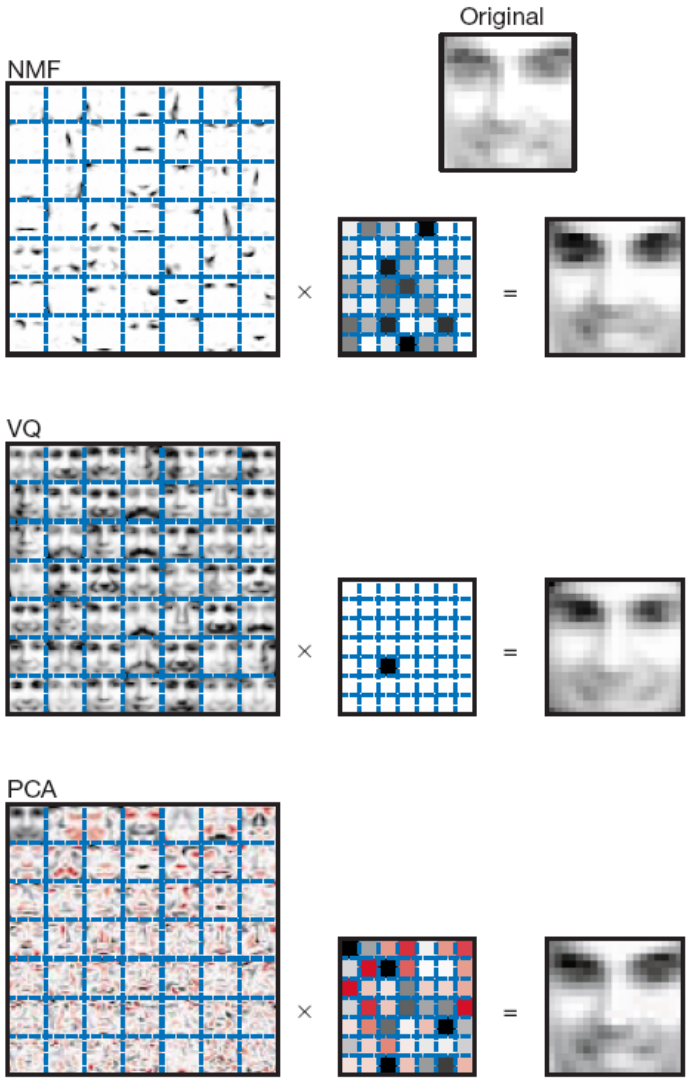


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting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

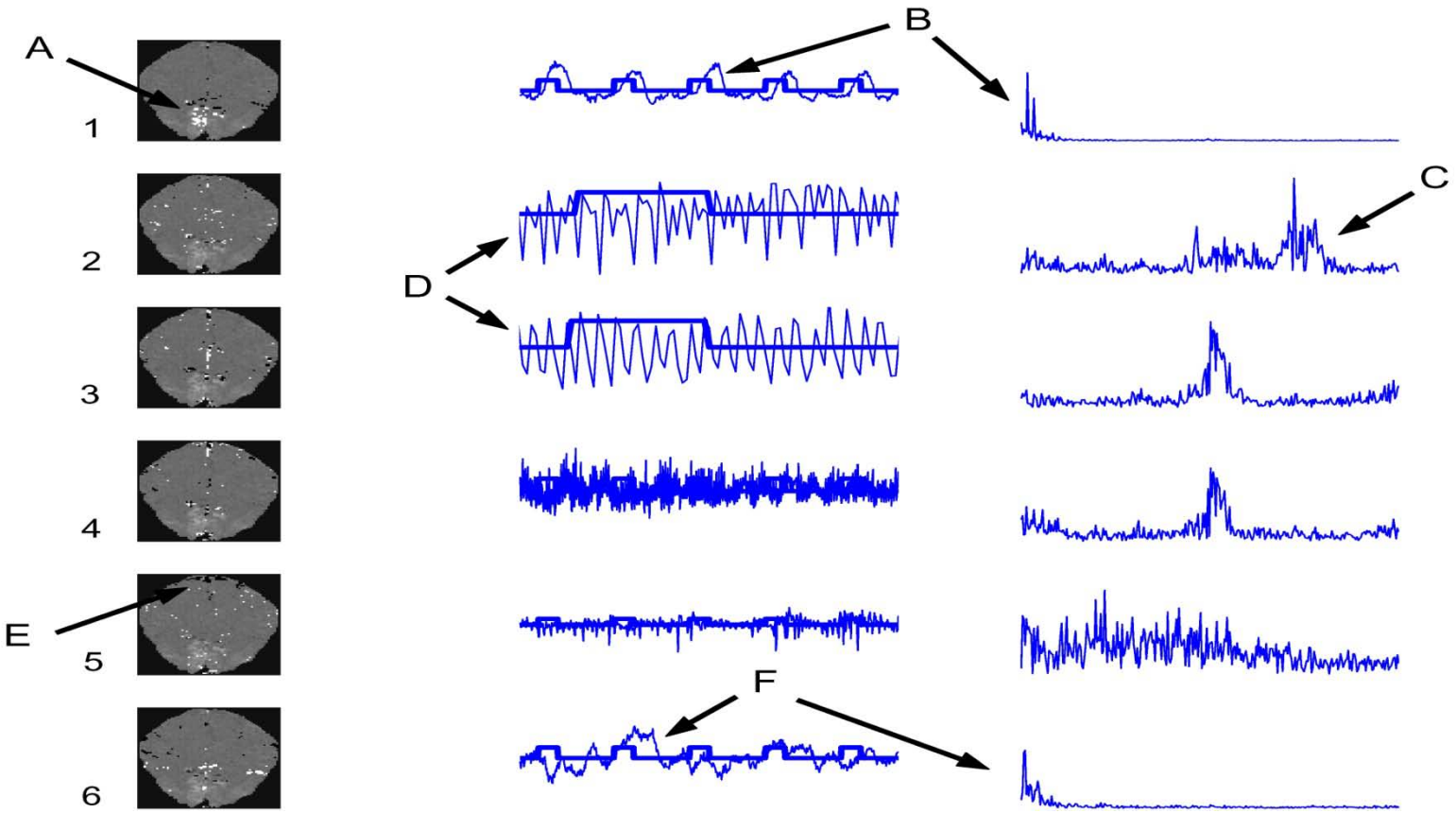
Learning the parts of objects by non-negative matrix factorization

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)

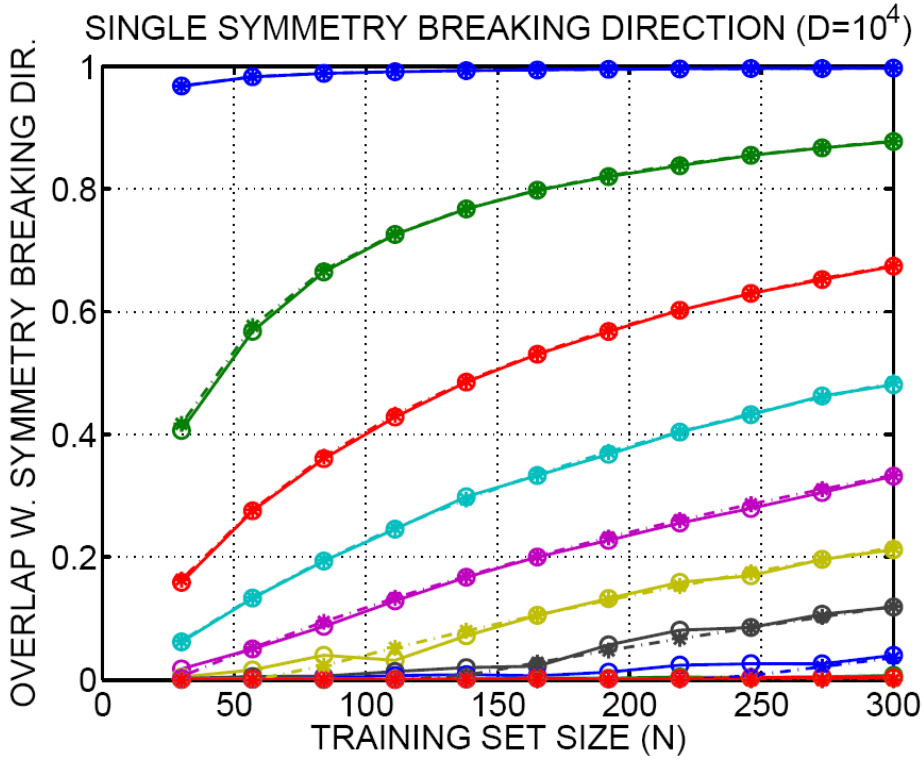
Modeling the generalizability of SVD

- 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 -> ∞

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

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

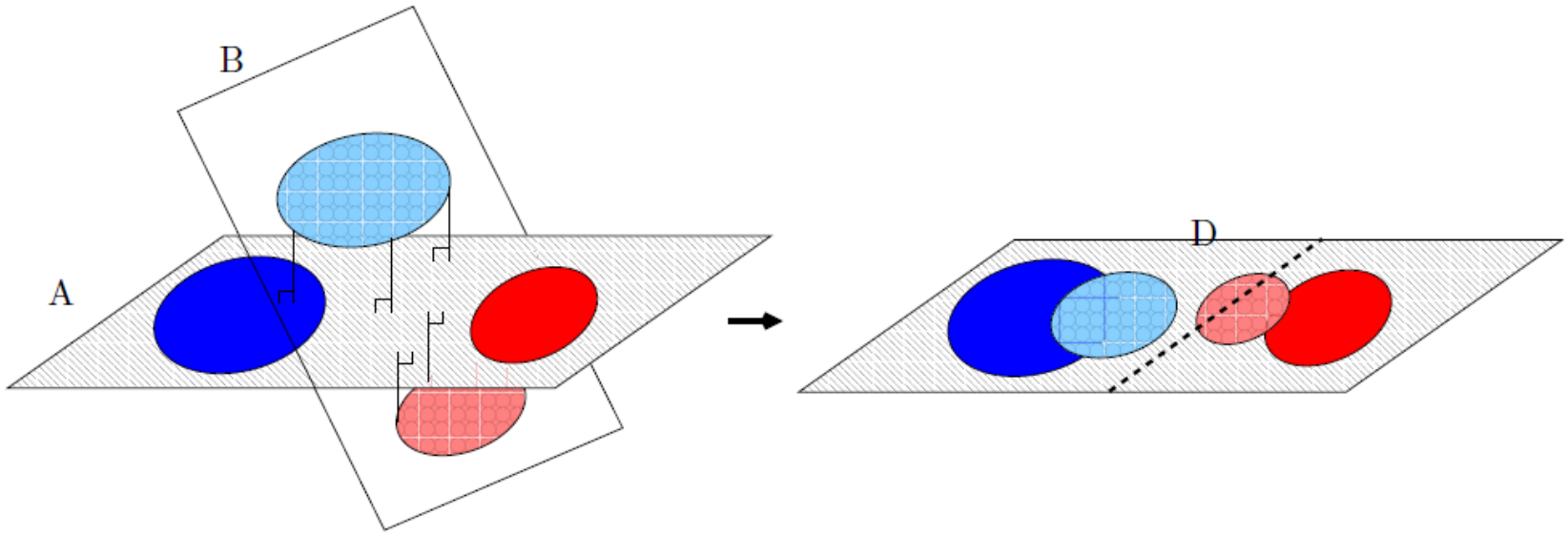
Hoyle, Rattray: Phys Rev E **75** 016101 (2007)



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

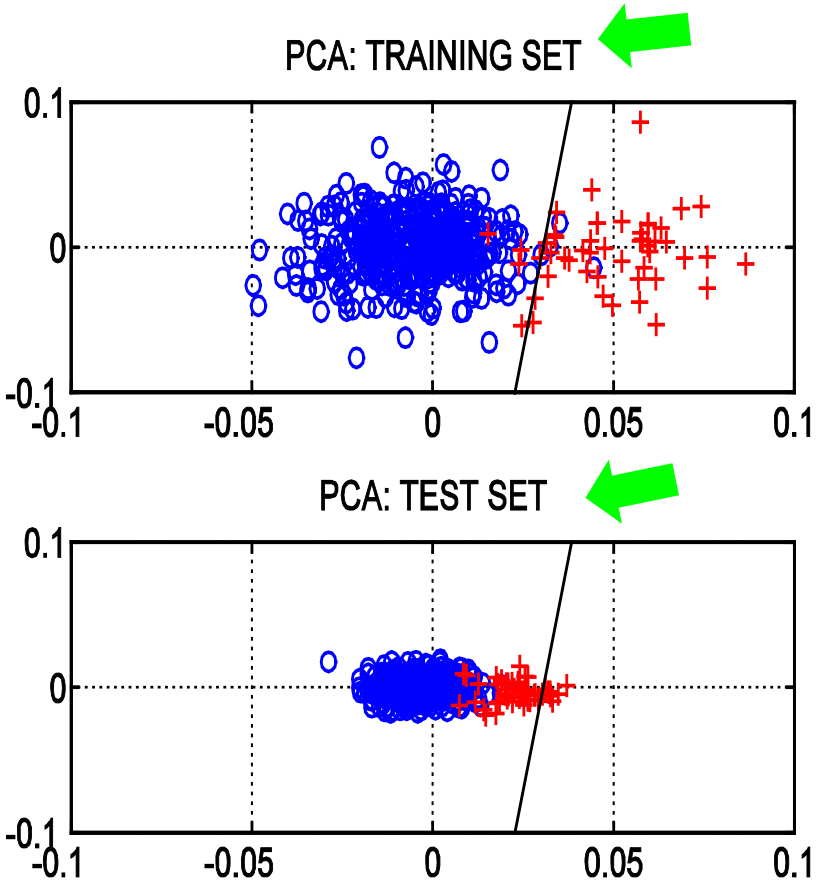
σ = (0.01, 0.06, 0.12, 0.17, 0.23, 0.28, 0.34, 0.39, 0.45, 0.5)

Generalizability – test training misalignment

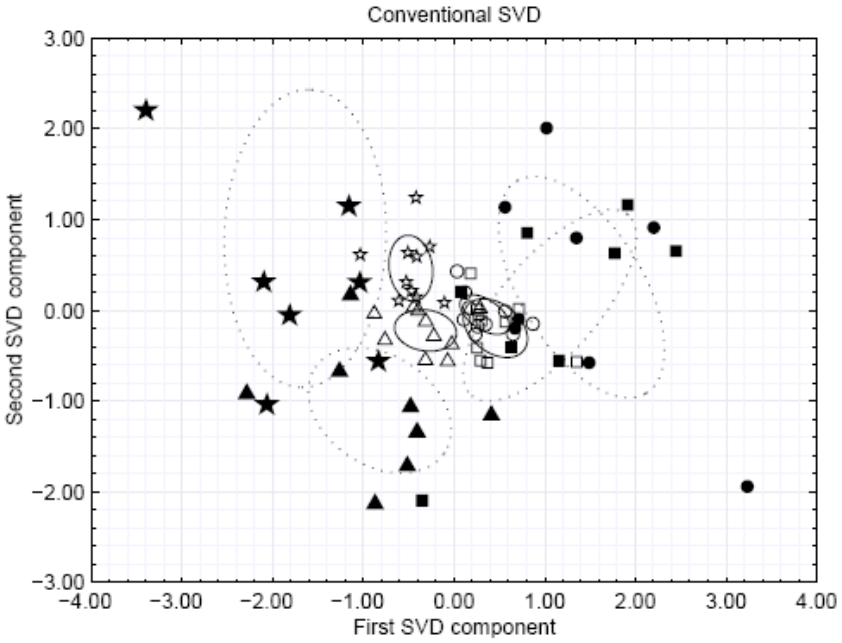


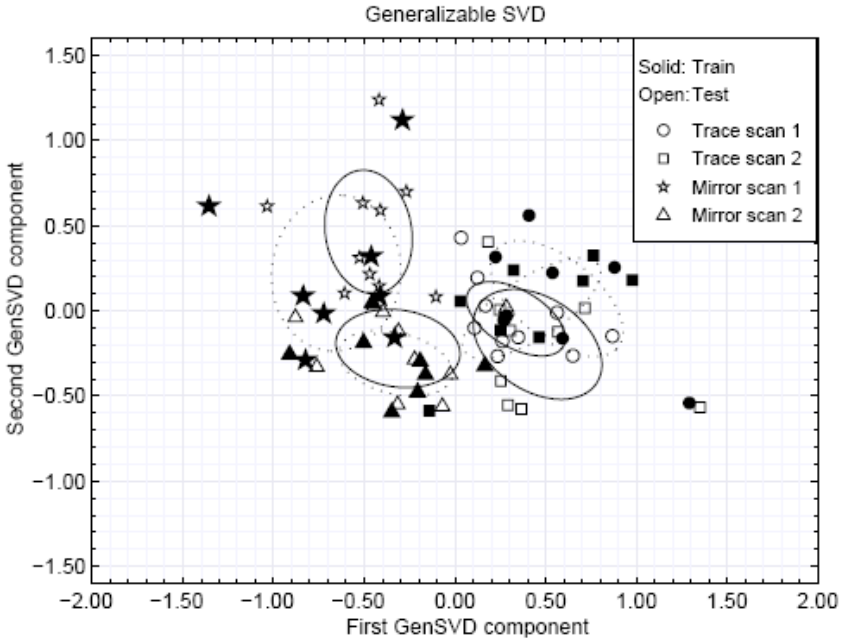
Restoring the generalizability of SVD

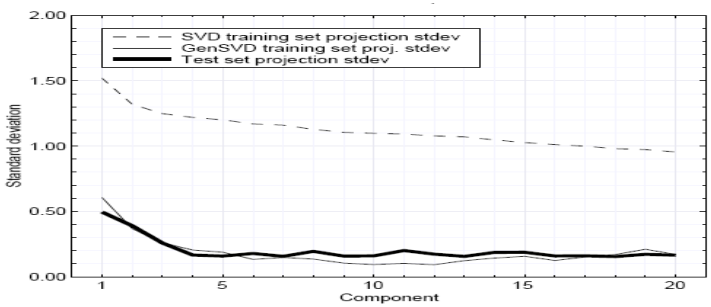
- Now what happens if you are on the slope of generalization, i.e., N/D is just beyond the transition to retarded learning ?
- The estimated projection is offset, hence, future projections will be too small!
- ...problem if discriminant is optimized for unbalanced classes in the training data!



Heuristic: Leave-one-out re-scaling of SVD test projections







N=72, D=2.5 10^4

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

Cimbi Center for integrated molecular brain imaging

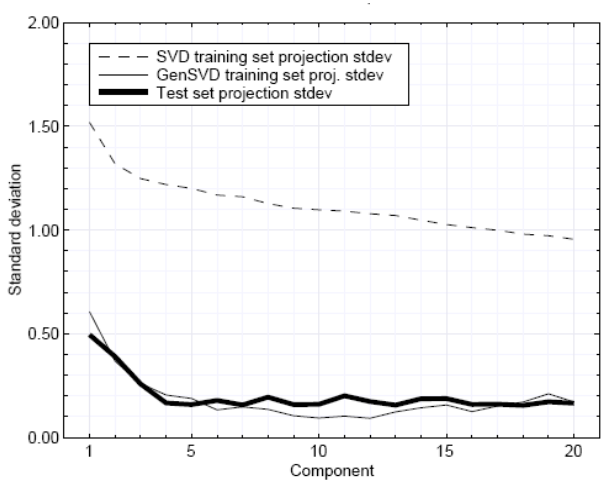
Lars Kai Hansen

IMM, Technical University of Denmark

DTU

Re-scaling the component variances

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

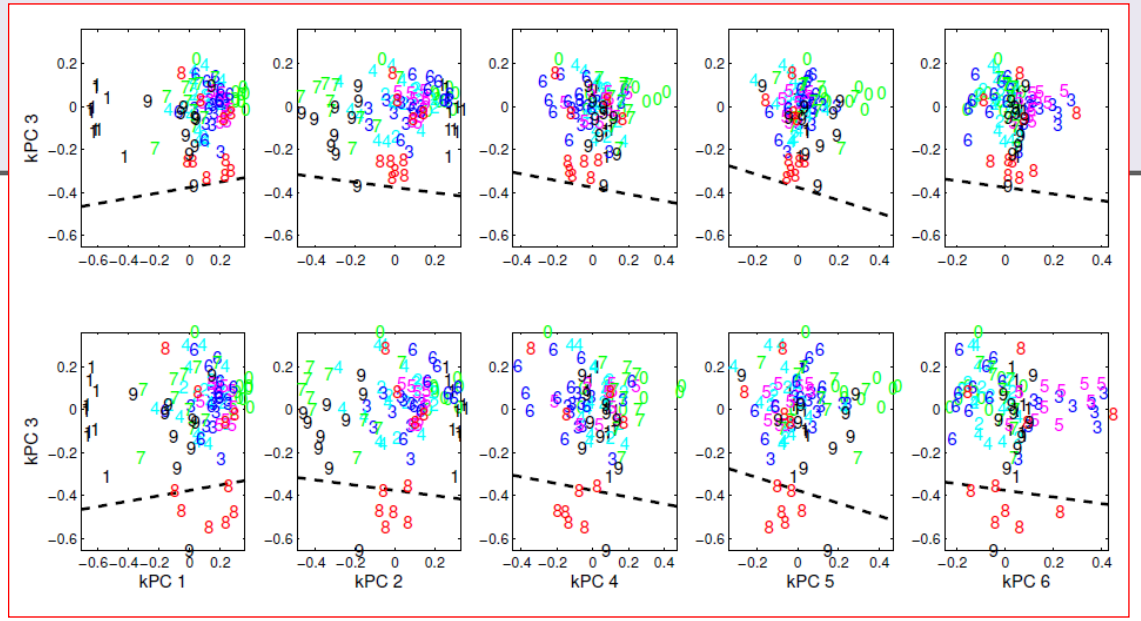


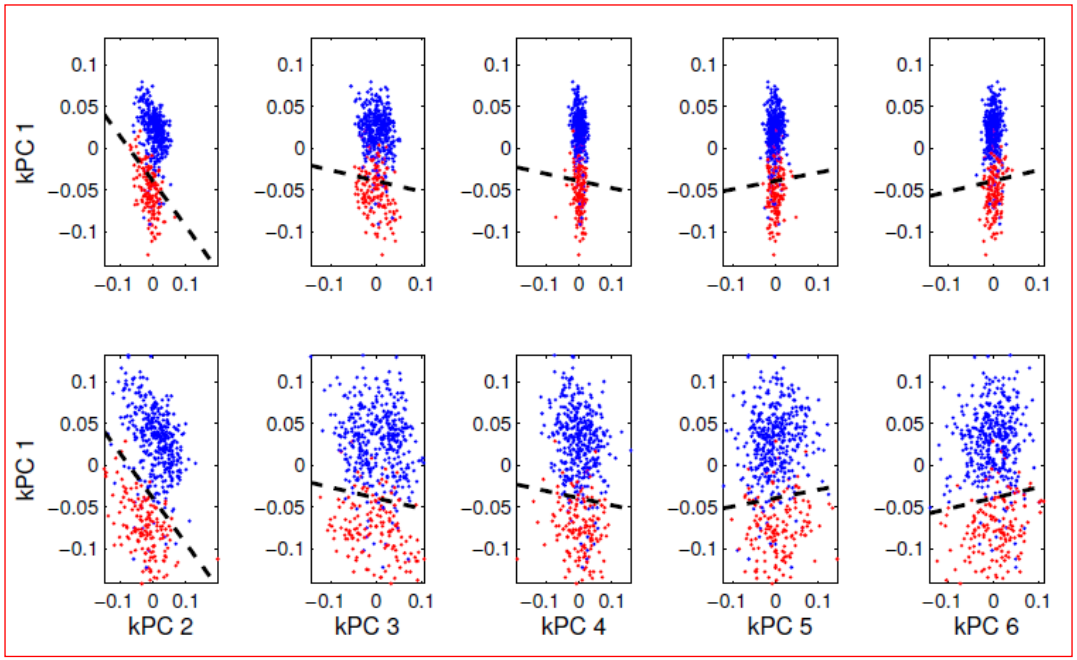
$$\text{Compute } \boldsymbol{U}_0\boldsymbol{\Lambda}_0\boldsymbol{V}_0^{\mathsf{T}} = \text{svd}(X) \text{ and } \boldsymbol{Q}_0 = [\boldsymbol{q}_j] = \boldsymbol{\Lambda}_0\boldsymbol{V}_0^{\mathsf{T}}$$
$$\text{foreach } j = 1 \ldots N$$
$$\bar{\boldsymbol{q}}_{-j} = \tfrac{1}{N-1}\textstyle\sum_{j' \neq j} \boldsymbol{q}_{j'}$$
$$\text{Compute } \boldsymbol{B}_{-j}\boldsymbol{\Lambda}_{-j}\boldsymbol{V}_{-j}^{\mathsf{T}} = \text{svd}(\boldsymbol{Q}_{-j} - \bar{\boldsymbol{Q}}_{-j})$$
$$\boldsymbol{z}_j = \boldsymbol{B}_{-j}\boldsymbol{B}_{-j}^{\mathsf{T}}(\boldsymbol{q}_j - \bar{\boldsymbol{q}}_{-j})$$
$$\hat{\lambda}_i^2 = \tfrac{1}{N-1}\textstyle\sum_j z_{ij}^2$$

Kjems, Hansen, Strother: NIPS (2001)

Variance inflation in kernel PCA

Handwritten digits:





fMRI data

single slice rest/visual stim (TR= 333 ms)

Challenges for the linear factor model

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

Data represented as multiway arrays

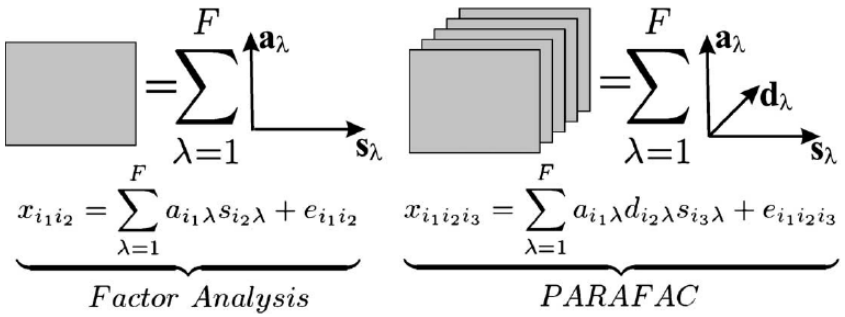
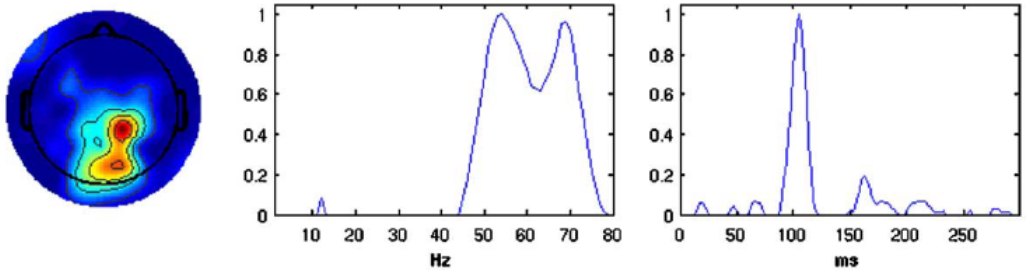


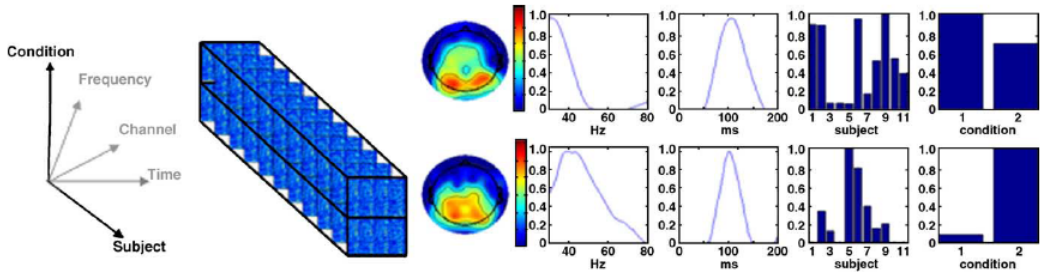
Fig. 1. Graphical representation of the factor analysis to the left and the PARAFAC decomposition of a 3-way array to the right. Like the factor analysis, PARAFAC decomposes the data into factor effects pertaining to each modality. F denotes the number of factors.

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



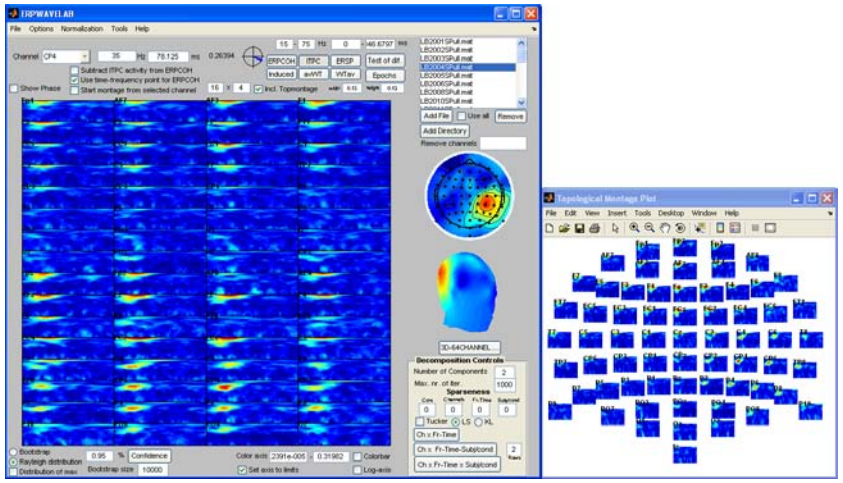
Mørup et al. NeuroImage (2005), NeuroImage (2008)

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

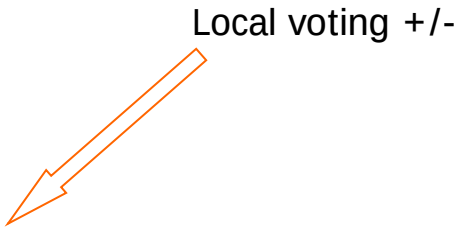


Toolbox download from www.erpwavelab.com

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

Generalizable supervised models

- Non-linear kernel machines, SVM



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

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

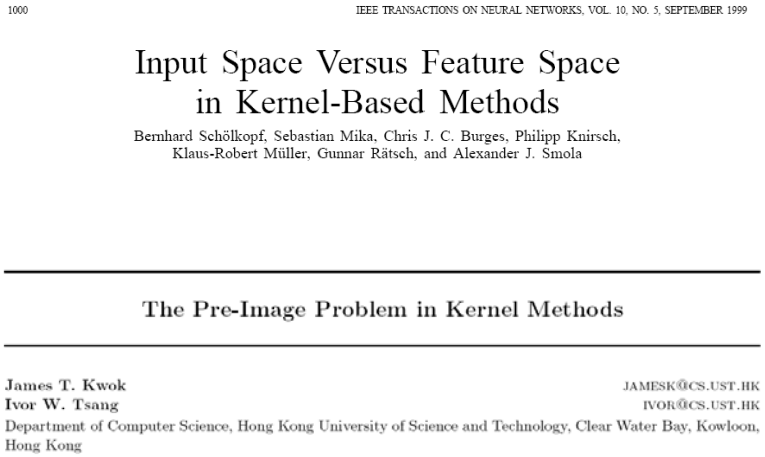
Visualization of SVM learning from fMRI

- 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



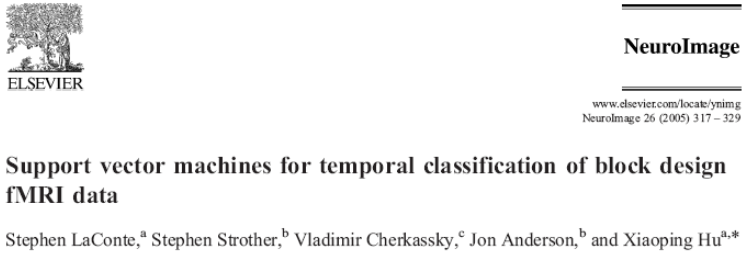
1000 IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 5, SEPTEMBER 1999

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

The Pre-Image Problem in Kernel Methods

James T. Kwok JAMESK@CS.UST.HK
Ivor W. Tsang IVOR@CS.UST.HK
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong



ELSEVIER

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,*]

- 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

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

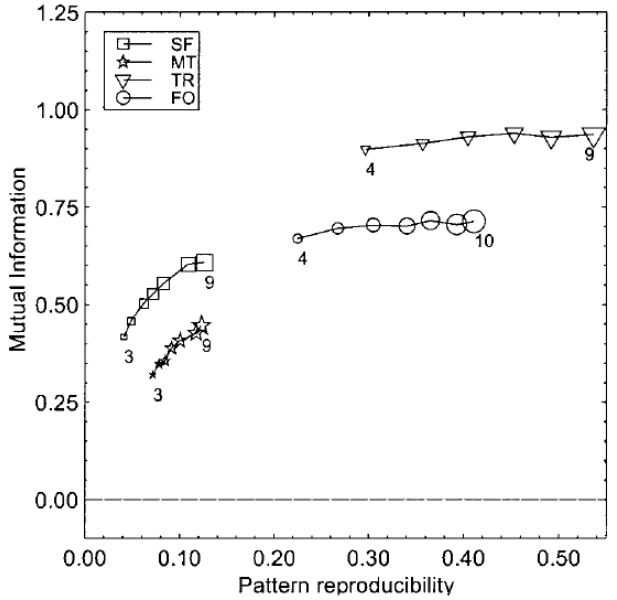
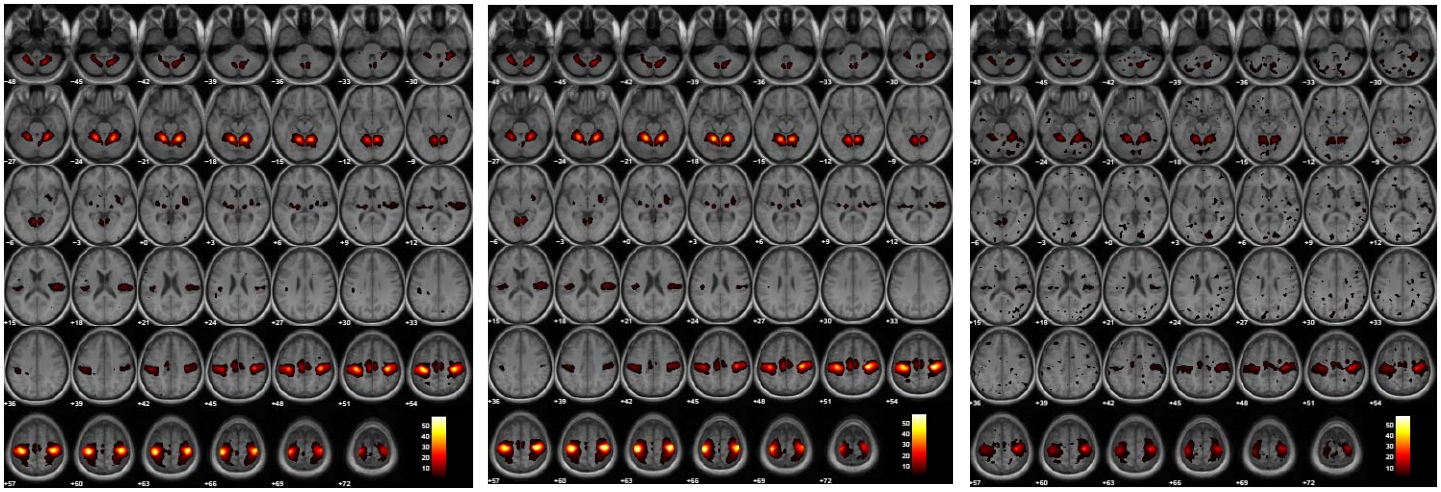


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both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

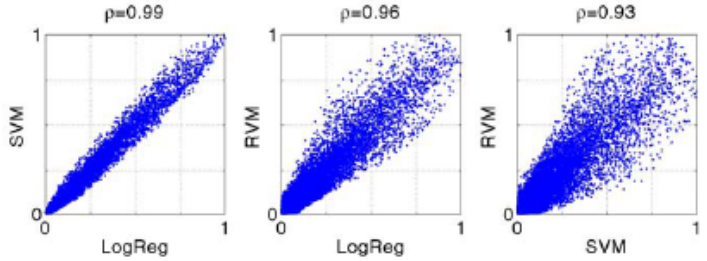


Figure 7: fMRI fingertapping experiment - consensus analysis. The plots show the extend of consensus in the average rSPI among the three models. The rSPI for LogReg was scaled by its maximum value. Hereafter the rSPIs from the SVM and RVM were transformed to match the histogram of that of LogReg. Correlation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

Sensitivity maps for non-linear kernel regression

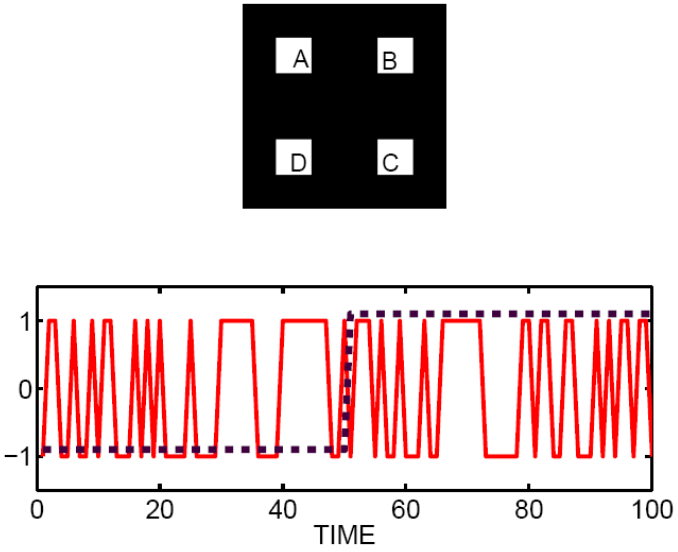


Fig. 1. XOR-image set define by four activated regions (A,B,C,D). Initially we let regions (A,B,D) be activated by random sequence taking values ± 1, as shown in example in the bottom panel (full curve). The target signal, also taking values $t_n = pm1$, and is also indicated in the bottom panel (dashed 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

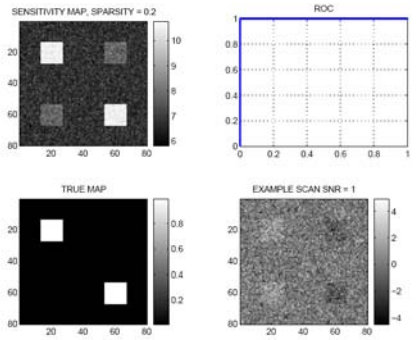


Fig. 2. XOR-image set define by four activated regions. The results of analyzing a image set with $N = 400$ examples. The image signal-to-noise ratio is $SNR = 1$, i.e., the additive noise is unit variance. The target function has in addition been contaminated by 10% random label noise. The four subplots 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.

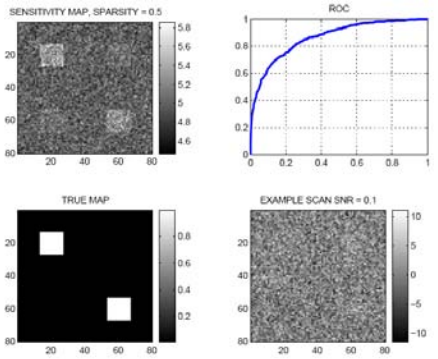
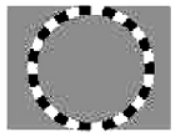


Fig. 4. XOR-image set define by four activated regions. Similar to figure 2, however the image signal-to-noise ratio is $SNR = 0.1$.

Non-linearity in fMRI?

Visual stimulus: half checker board no/left/right/both



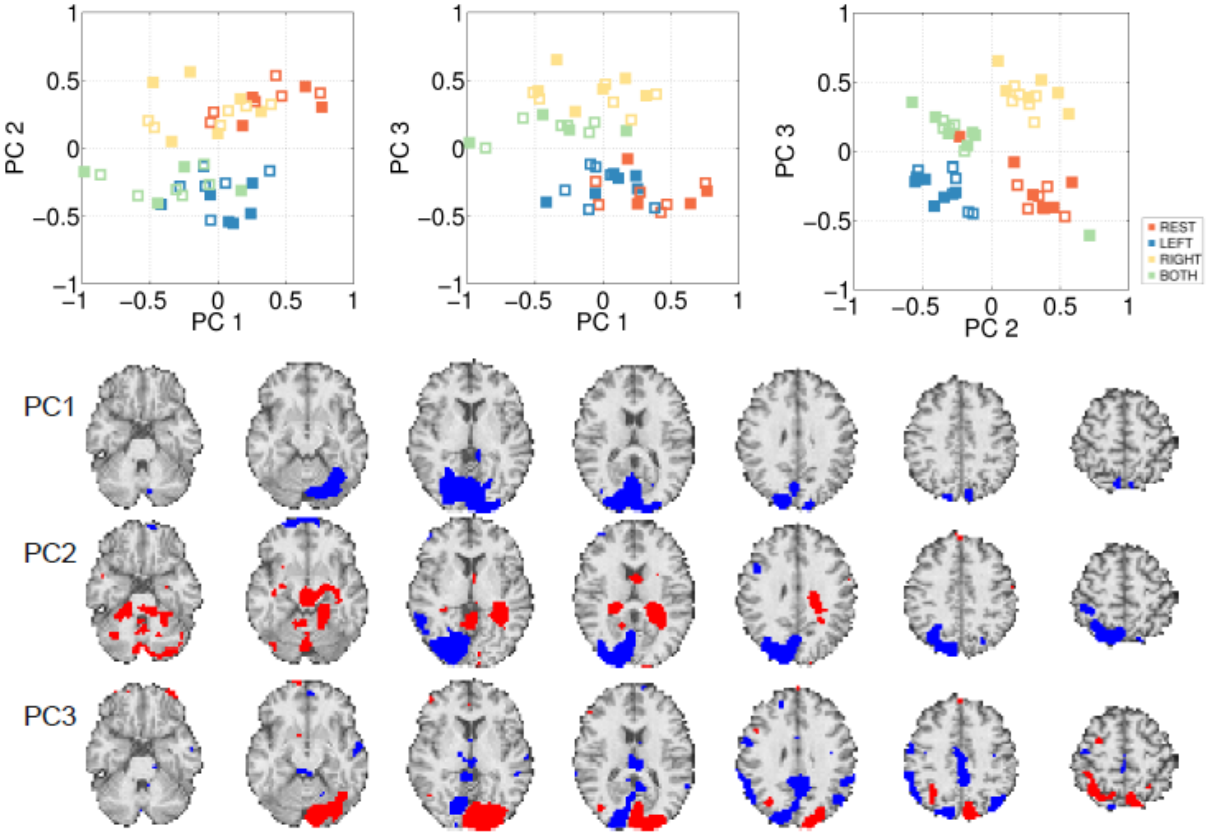
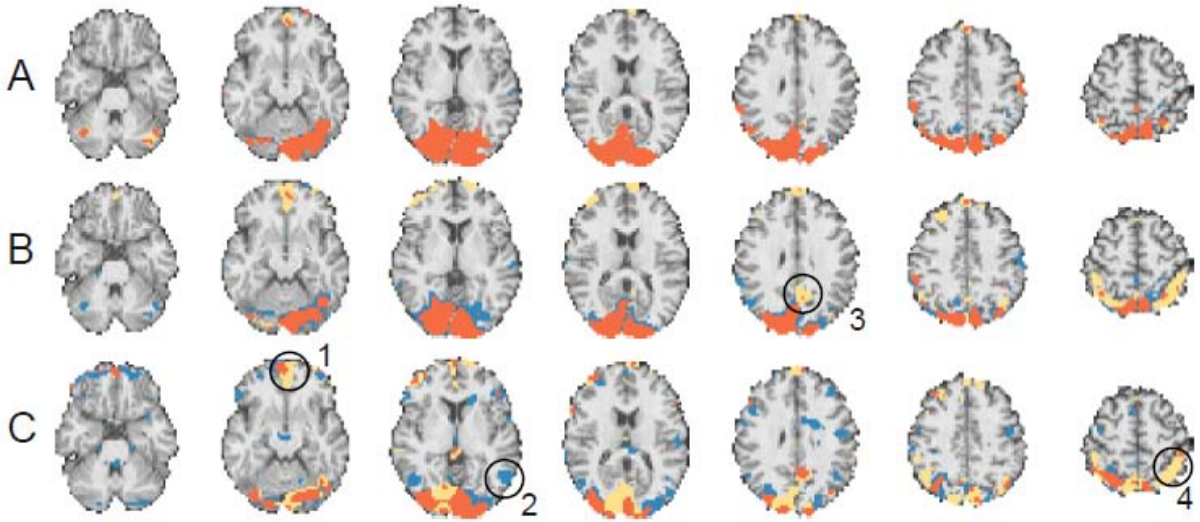
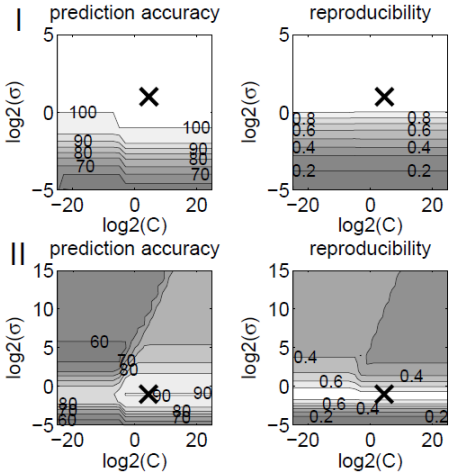


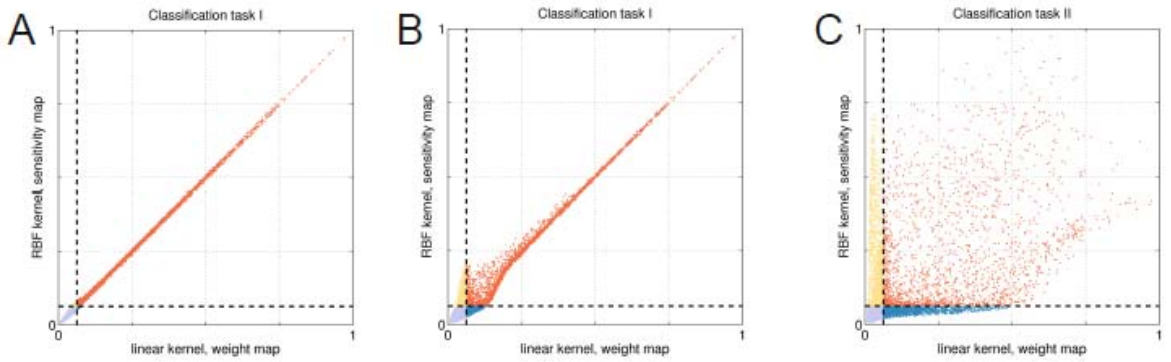
Figure 1: PCA analysis of the fMRI data set. An example of the three first PCs estimated from the training set in a NPAIRS split. The scatter plots show both training (filled markers) and test examples projected onto the PCs. The blue and red voxels on the brain slices corresponds to negative and positive PC loadings respectively. The maps are thresholded to show the 5 upper positive and negative percentiles.

Non-linearity in fMRI – detecting networks

Peter Mondrup Rasmussen et al. 2010







A: Easy problem- (Left vs Right) and RBF kernel is wide ... i.e. similar to linear kernel

B: Easy problem- Pars optimized to yield the best P-R

C : Hard XOR problem Pars optimized to yield The best P-R

Cimbi Center for integrated molecular brain imaging

Lars Kai Hansen

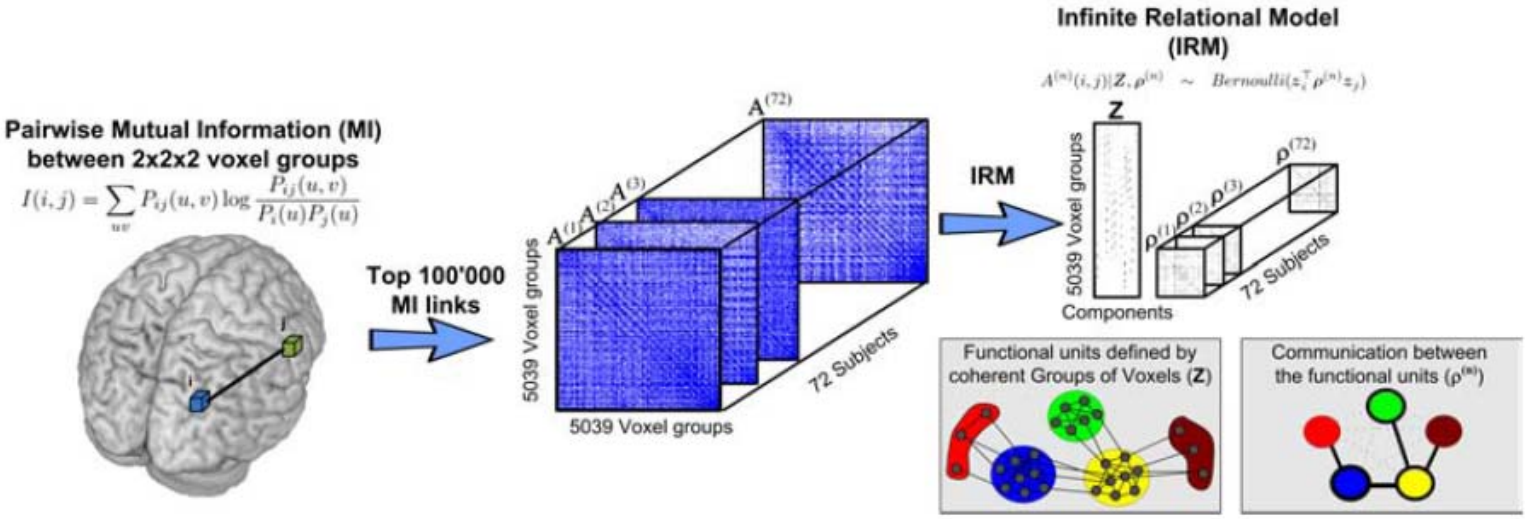
IMM, Technical University of Denmark

DTU

Detecting networks with relational models

Morten Mørup et al. NIPS 2010

Different networks in fMRI resting state fluctuations separates a group of MS patients from normal group (Ntot =72)



Basic measure: Mutual information between time series (can detect similarity by modulation)

Infinite Relational Model (IRM) is inspired by social networks:

A new clustering approach - clustering based on similar communication instead of similar time series

IRM models - Detect communities of similar communication

A is the mutual information graph, ρ the ”community” connectivity matrix, and Z is the community assignment variables

$$
\begin{aligned}
\boldsymbol{Z}|\alpha &\sim \mathrm{DP}(\alpha) \\
\boldsymbol{\rho}^{(n)}(a,b)|\boldsymbol{\beta}^{+}(a,b),\boldsymbol{\beta}^{-}(a,b) &\sim \mathrm{Beta}(\boldsymbol{\beta}^{+}(a,b),\boldsymbol{\beta}^{-}(a,b)) \\
\boldsymbol{A}^{(n)}(i,j)|\boldsymbol{Z},\boldsymbol{\rho}^{(n)} &\sim \mathrm{Bernoulli}(\boldsymbol{z}_{i_r}^{\top}\boldsymbol{\rho}^{(n)}\boldsymbol{z}_{j_r})
\end{aligned}
$$

Detecting multiple sclerosis vs normal subjects

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

Conclusion

- Machine learning in brain imaging has two equally important aims
  - Generalizability
  - Reproducible interpretation
- 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
- More complex mechanisms may be revealed with non-linear detectors

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