Resampling based methods for design and evaluation of neuroimaging models

Lars Kai Hansen
DTU Informatics
Technical University of Denmark

Co-workers:

Morten Mørup, Kristoffer Madsen, Peter Mondrup, Trine Abrahamsen, ... Stephen Strother (Toronto)





Cimbi
Center for integrated molecular brain imaging

DTU

Do not multiply causes!

OUTLINE

Machine learning – with two agenda

- Aim I: To abstract generalizable relations from data
- Aim II: Robust interpretation / visualization
- The PR-plot for optimization

Unsupervised (explorative)

- Factor models - Linear hidden variable representations
- Generalization in unsupervised models
- Visualization
- Non-linear models, KPCA, PR-plots for tuning
- Network modeling

Supervised models (retrieval)

- Visualization of non-linear kernel machines
- PR-plotting supervised models
- Pattern recognition in network models

Outlook

Recent reviews

NeuroImage 45 (2009) S199–S209

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Machine learning classifiers and fMRI: A tutorial overview

Francisco Pereira [a,*], Tom Mitchell [b], Matthew Botvinick [a]

[a] Princeton Neuroscience Institute/Psychology Department, Princeton University, Princeton, NJ 08540, USA
[b] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

YNIMG-07534; No. of pages: 11; 4C: 3, 4, 7, 9

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

ARTICLE

NeuroImage xx›

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

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.c

NeuroImage 56 (2011) 387–399

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Decoding fMRI brain states in real-time

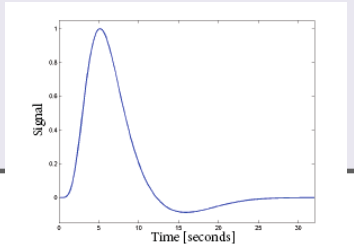
Stephen M. LaConte

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

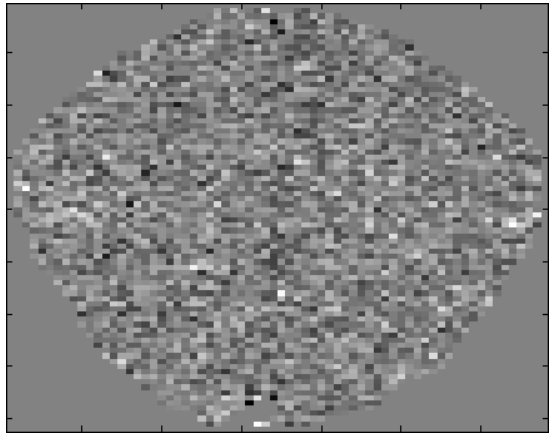
Introduction to machine learning for brain imaging

Steven Lemm [a,c,*], Benjamin Blankertz [a,b,c], Thorsten Dickhaus [a], Klaus-Robert Müller [a,c]

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

Main message: Models should be predictive and informative

BOLD fMRI:Is hemodynamic de-convolution feasible?

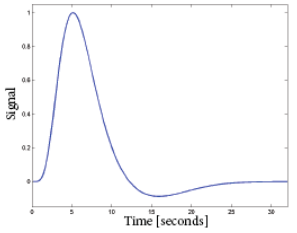
LETTER Communicated by Karl Friston

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)



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

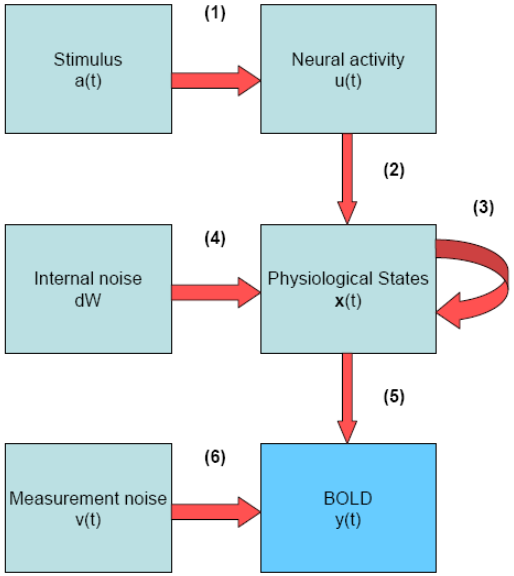


Figure 3.1: Overview diagram of hemodynamic models.

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

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

BOLD hemodynamics R-Bayes model selection

Model A: Sustained neural input vs Model B: Fading input

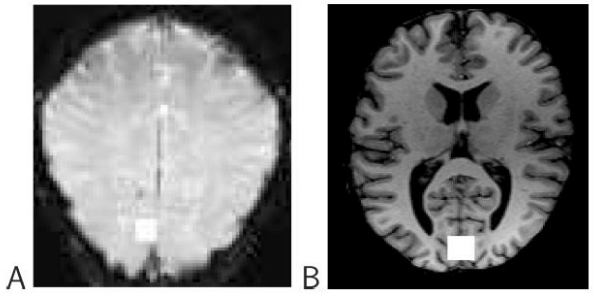


Figure 3: Regions of interest, marked with white squares. (A) Data set 1; T2* weighted image slice parallel to the calcarine sulcus. (B) Data set 2; MPRAGE (magnetization prepared rapid gradient echo) horizontal slice.

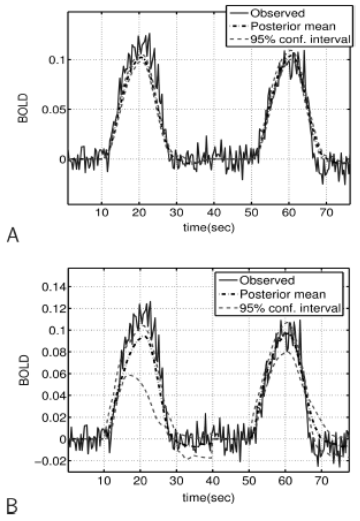


Figure 4: Prediction of data set 1. (A) Model A. (B) Model B. Note that the confidence interval is an empirical confidence interval for the mean prediction, based on the MCMC samples.

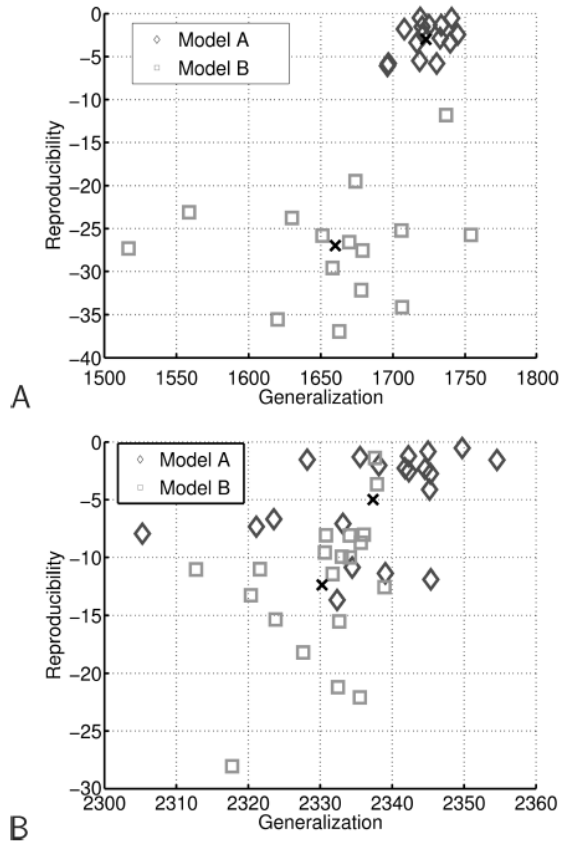


Figure 6: Generalization and reproducibility for real data; crosses mark the mean. (A) Data set 1. (B) Data set 2.

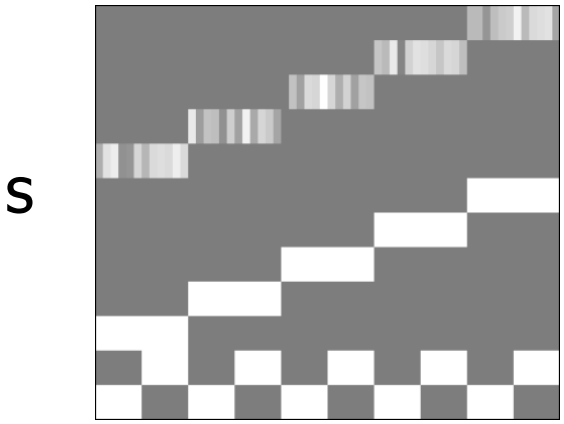
Lars Kai Hansen
IMM, Technical University of Denmark

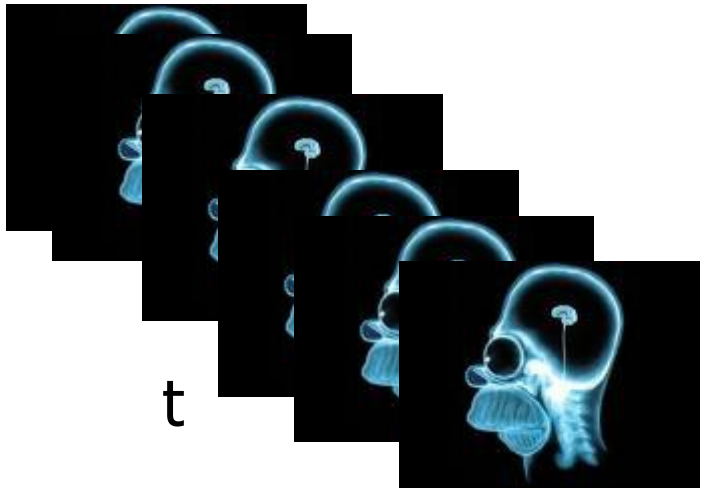
Machine learning for neuroimage data

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:

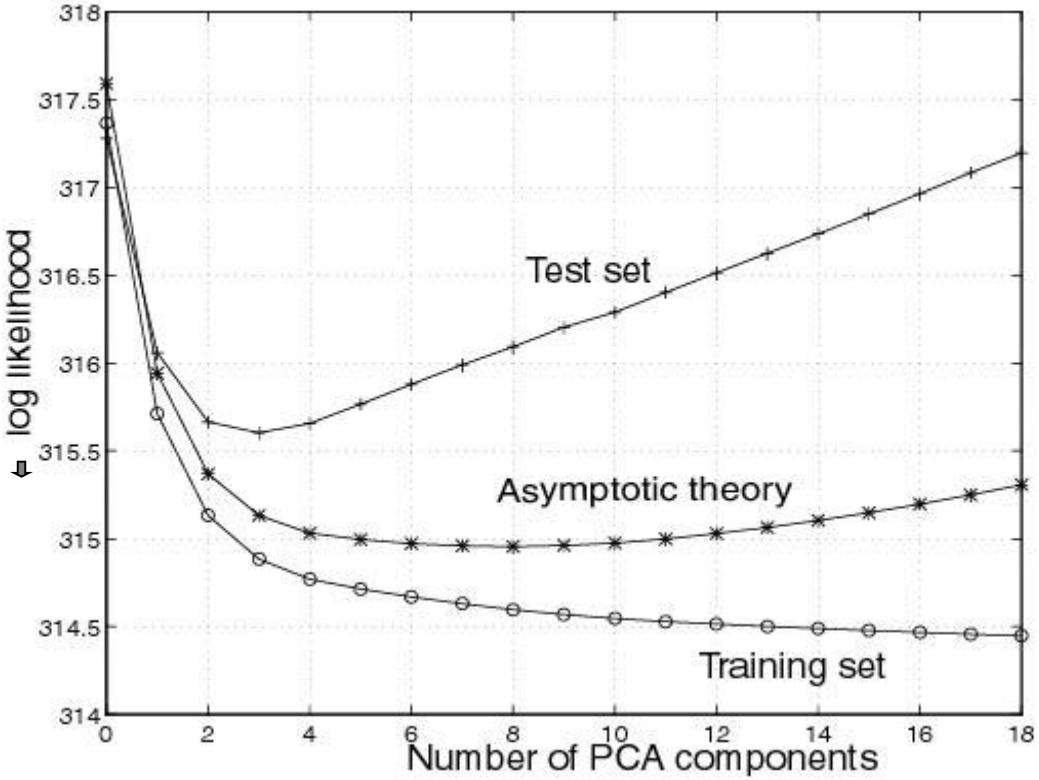
$$\left\langle -\log p(\mathbf{s}\,|\,\mathbf{x}, D)\right\rangle, \qquad \left\langle -\log p(\mathbf{x}\,|\,D)\right\rangle,$$

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

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

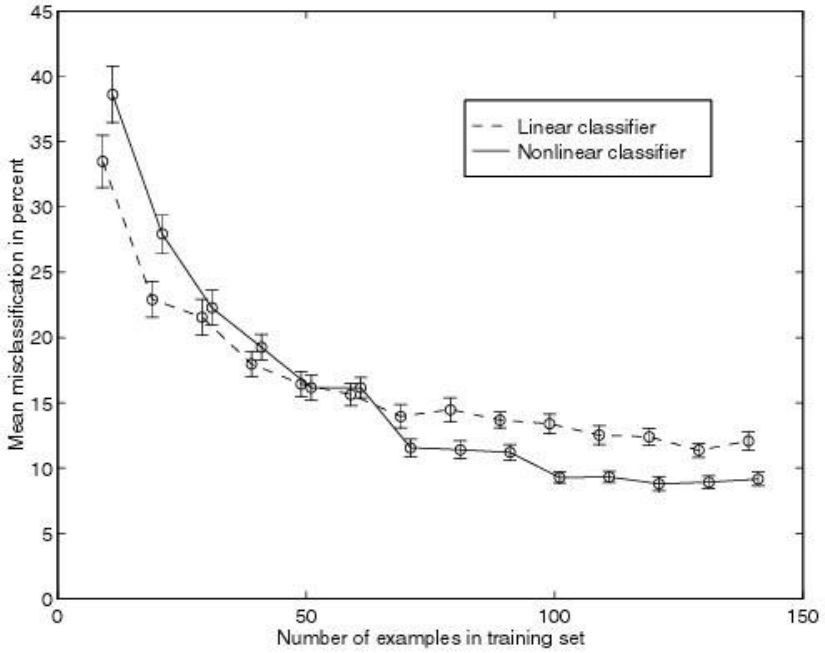
- Results can be presented as ”bias-variance trade-off curves” or ”learning curves”

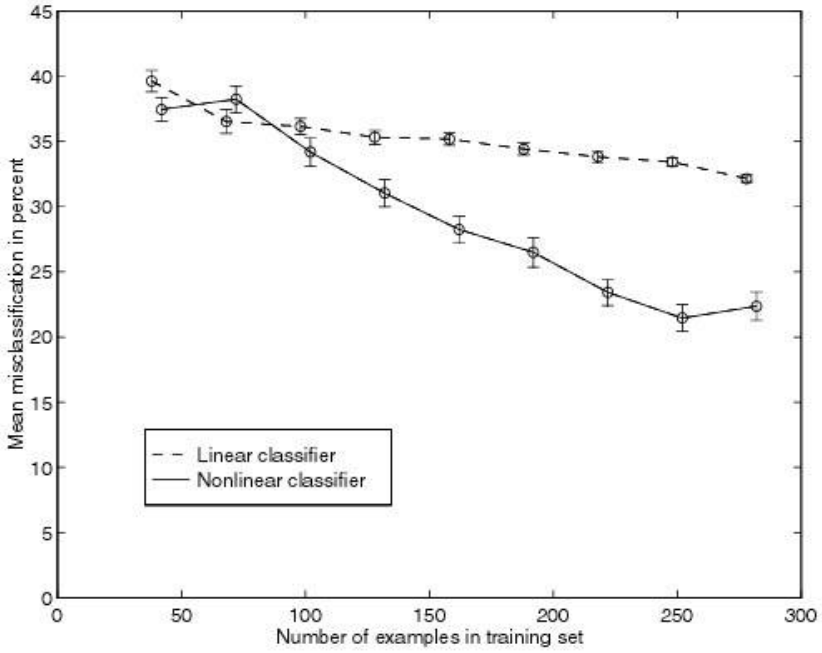
Bias-variance trade-off as function of PCA dimension



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)…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…

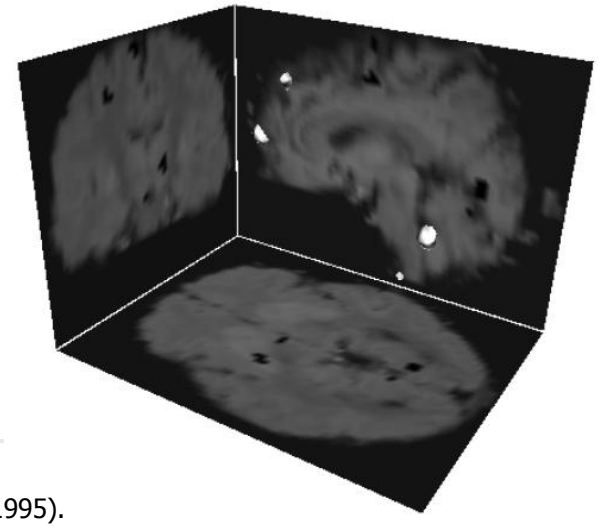
Statistical Parametric Maps

Weight maps in linear models

The saliency map (1995)

The sensitivity map (2000)

Consensus maps



N. Mørch et al.: Visualization of Neural Networks Using Saliency Maps.
in Proc 1995 IEEE Int Conference on Neural Networks, vol. 2, pp.2085-2090 (1995).

Reproducibility of parameters/visualization?
...hints from asymptotic theory

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

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 available also 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 & the PR plot

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

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

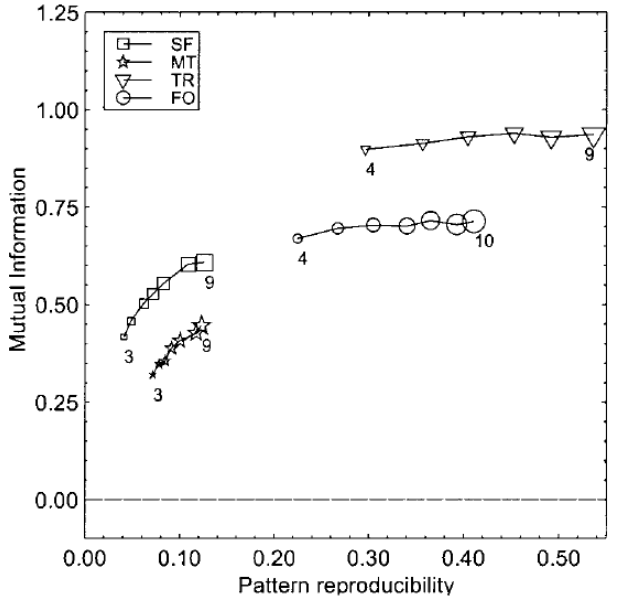
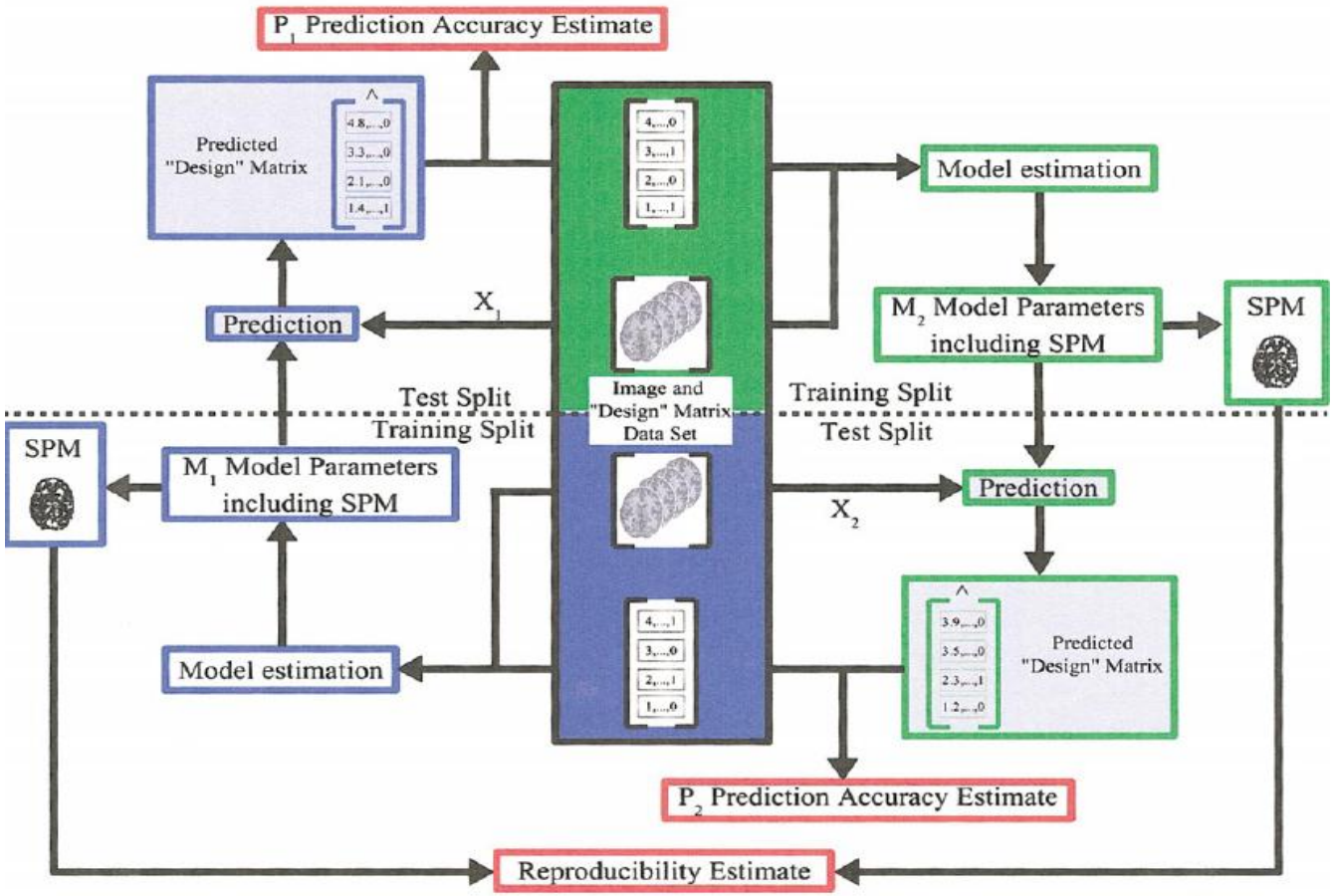


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both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

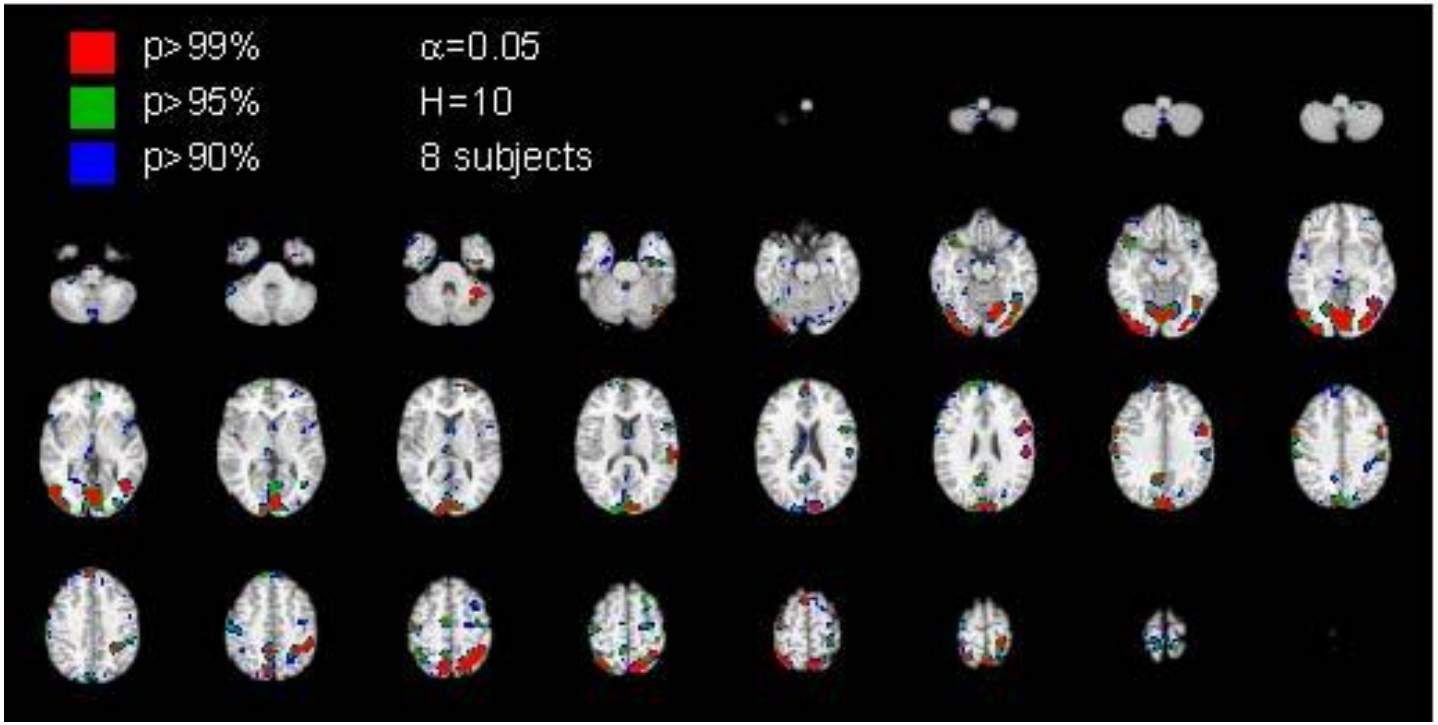
NPAIRS: Reproducibility of parameters



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), Mondrup et al (2011)
Brain and Language: Hansen (2007)

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),

Unsupervised learning

Explorative modeling
Learning stable structures in data p(x,s)

Unsupervised learning:
Factor analysis generative model

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

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

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

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

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

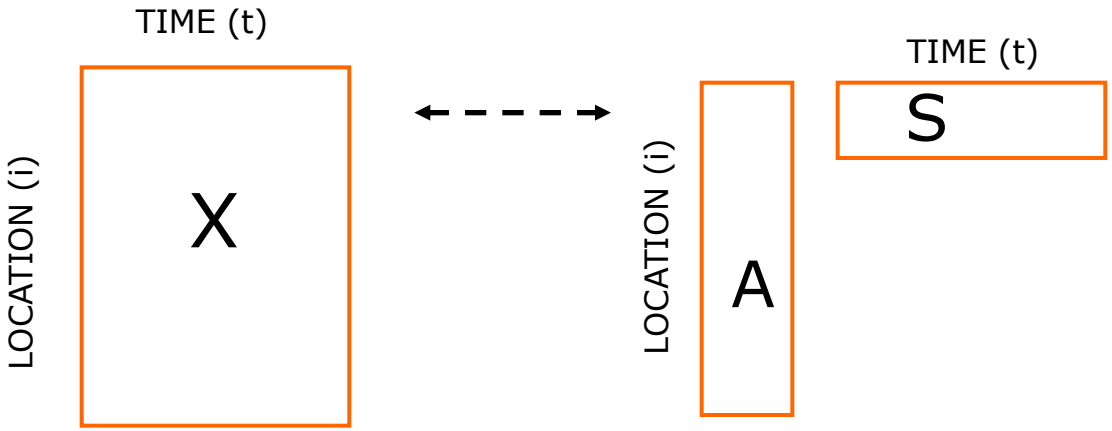
FA: $\boldsymbol{\Sigma} = \mathbf{D}$

S known: GLM
$(1\text{-}A)^{-1}$ 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

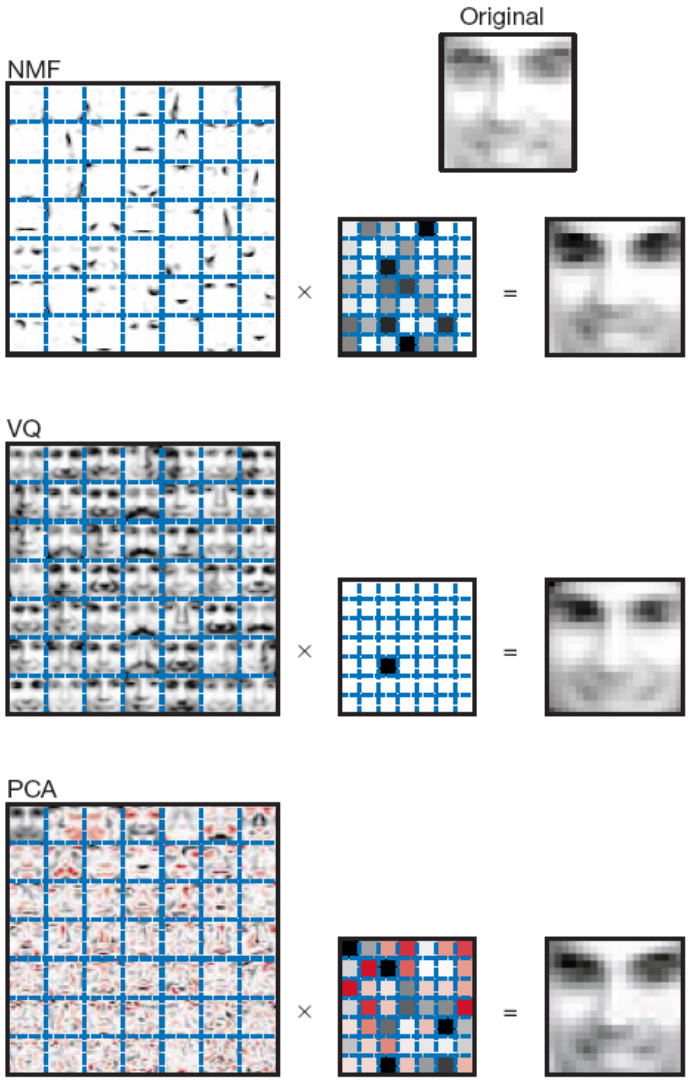


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting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

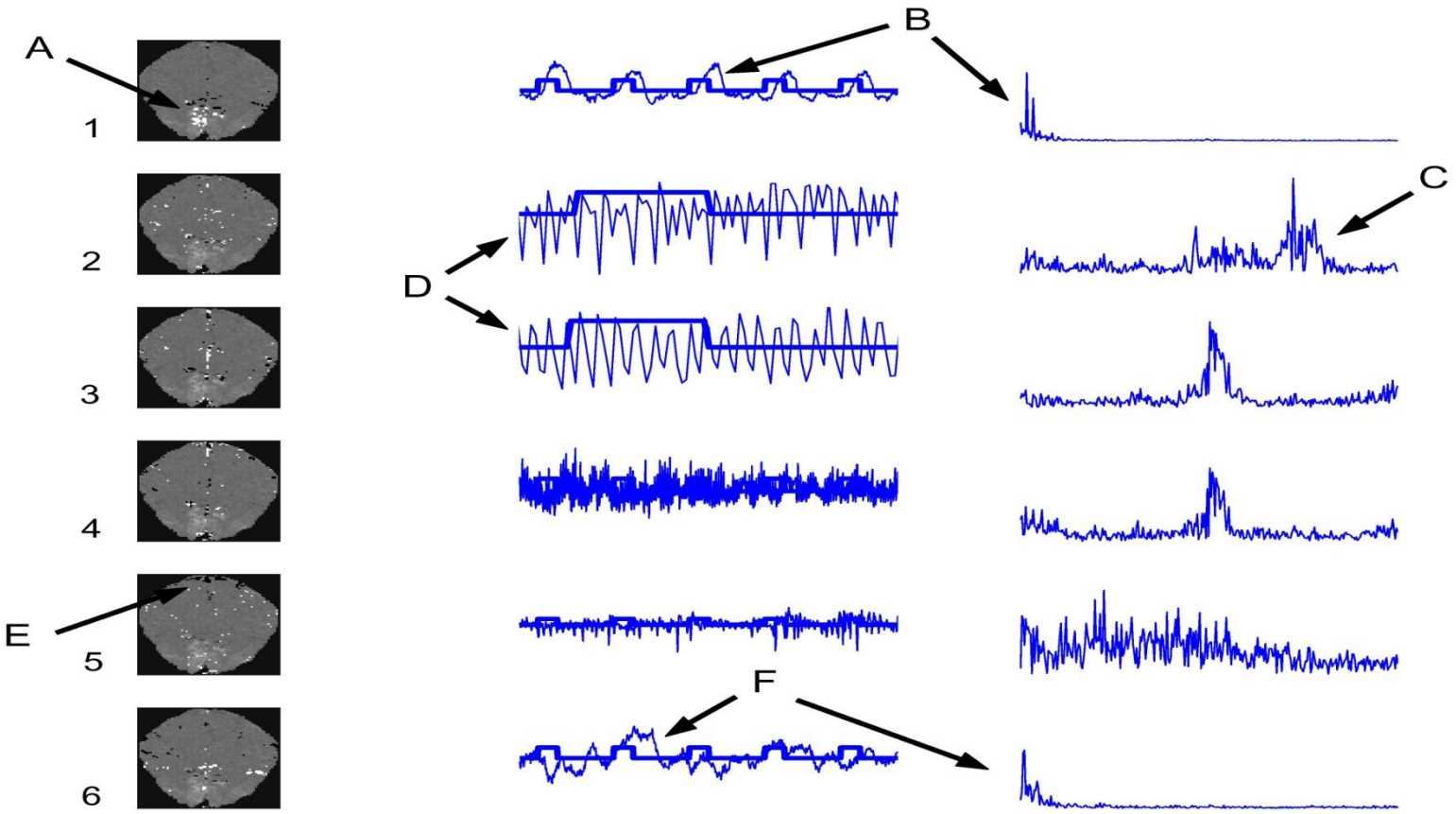
Learning the parts of objects by non-negative matrix factorization

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)

DTU:ICA toolbox

- 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

http://cogsys.imm.dtu.dk/toolbox/ica/

Modeling the generalizability of SVD

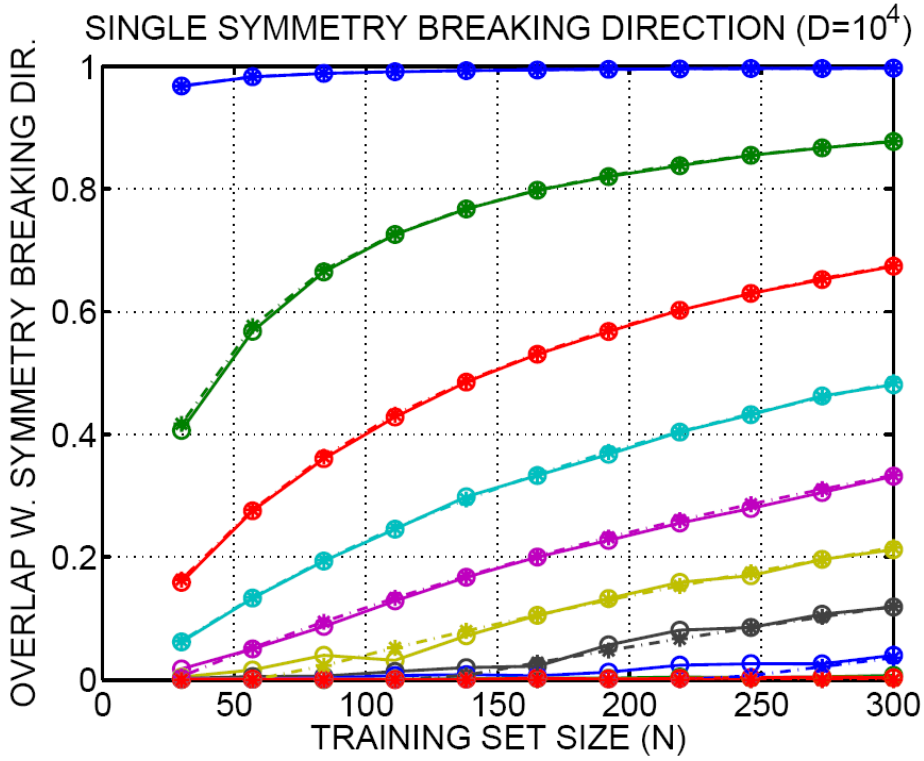
Rich physics literature on ”retarded” learning

- **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 -> ∞

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

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

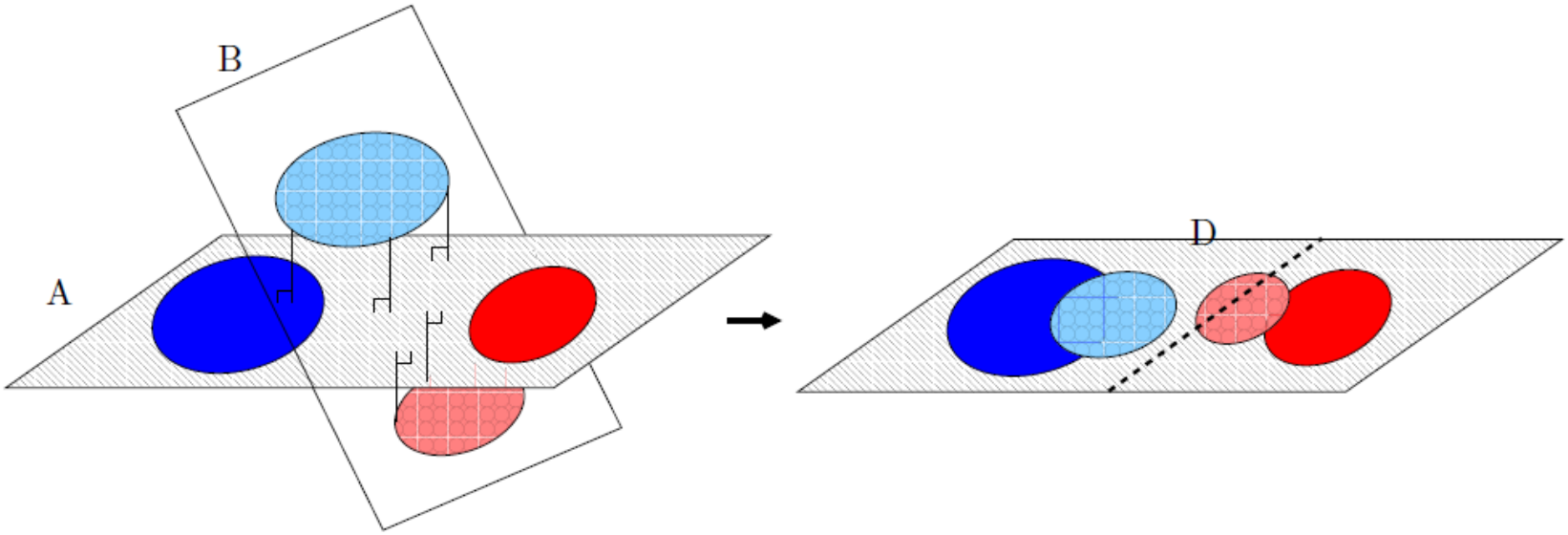
Hoyle, Rattray: Phys Rev E **75** 016101 (2007)



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

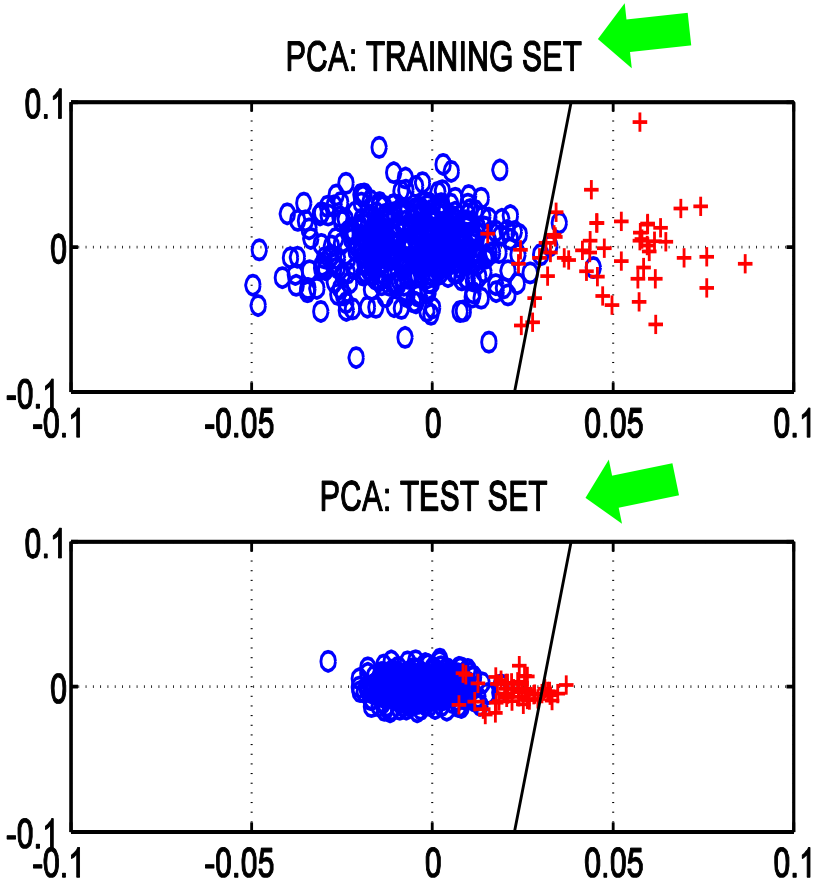


Restoring the generalizability of SVD

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

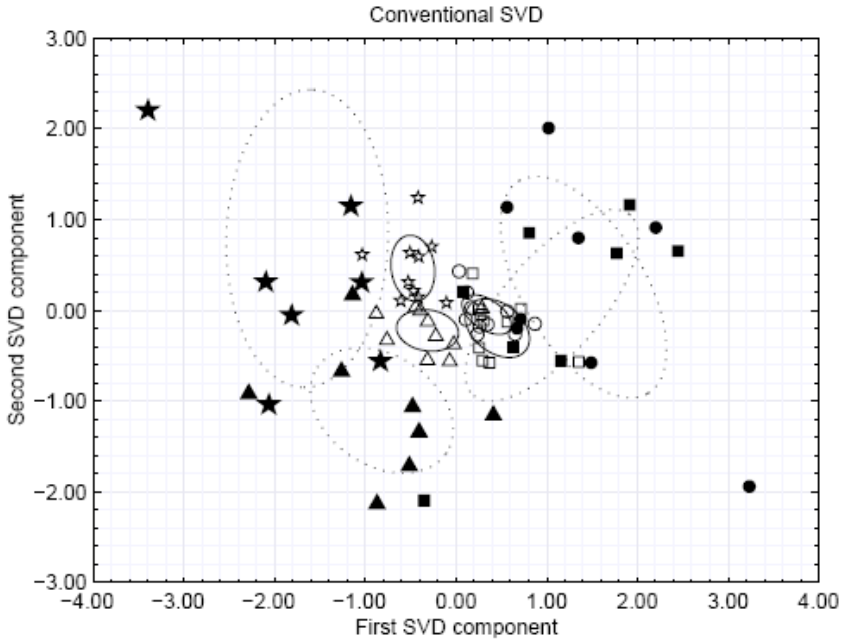
- The estimated projection is offset, hence, future projections will be too small!

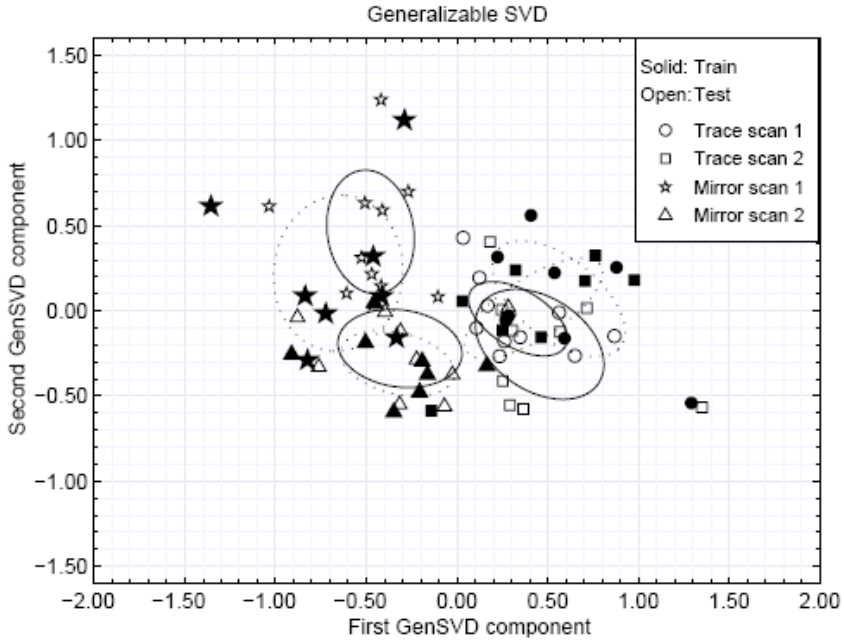
- …problem if discriminant is optimized for unbalanced classes in the training data!

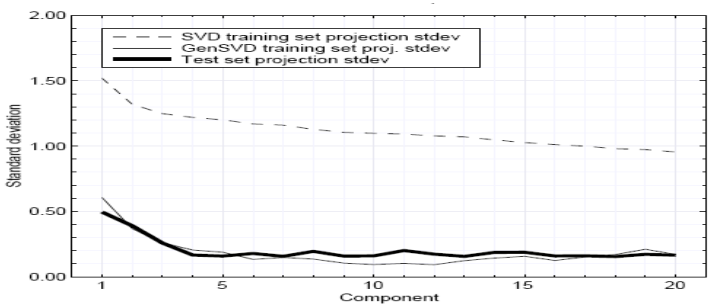


TJ. Abrahamsen, LK. Hansen. A Cure for Variance Inflation in High Dimensional Kernel Principal Component Analysis. Journal of Machine Learning Research 12:2027-2044 (2011).

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





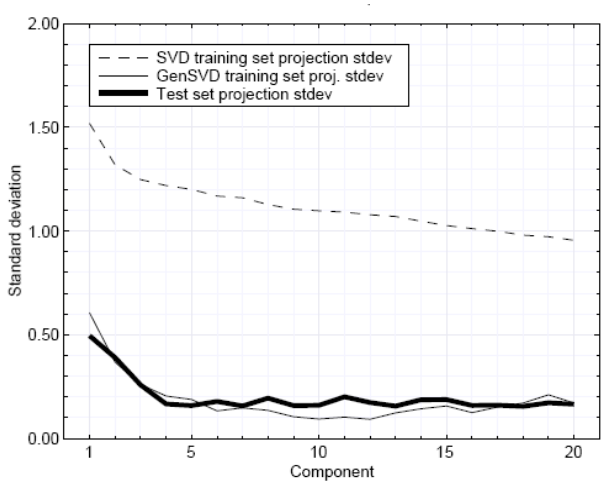


N=72, D=2.5 10^4

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

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)

More challenges for the linear factor model

- Too simple?
  - Non-linear manifolds
  - 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

Beyond the linear model: Motivation

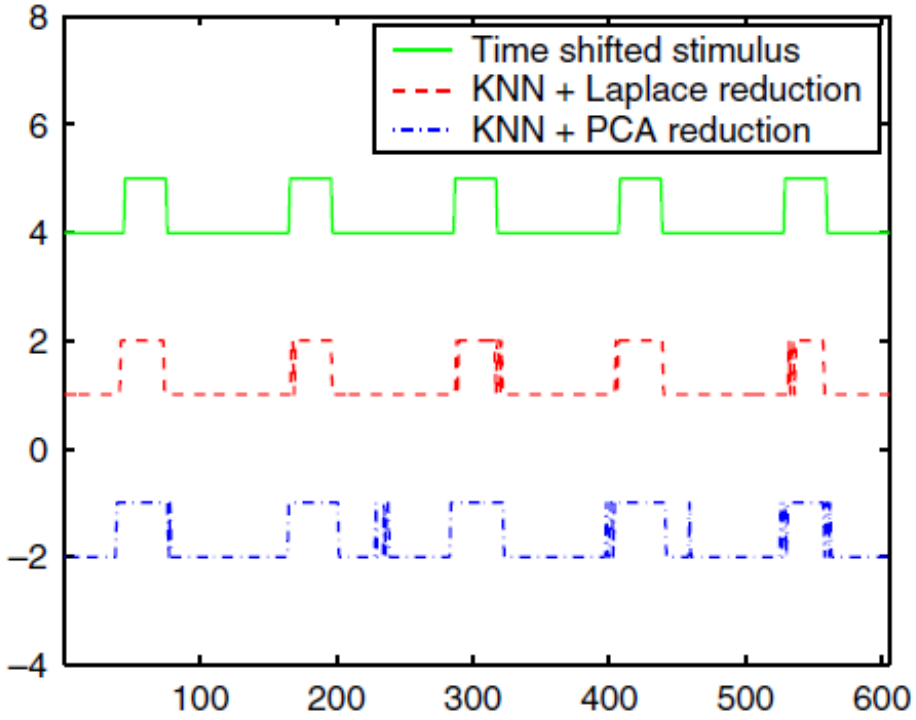


Fig. 4. Test set reference activation time course and activation time courses produced by *k*-nearest neighbor classifiers based on linear and non-linear feature spaces. The non-linear feature based classifier's errors basically occurs at the onset and at end of stimulation. The KNN model based on the linear feature representation make additional generalization errors in the baseline, where it suggest a few short burst of activation.

Beyond the linear model: De-noising by projection onto non-linear signal manifolds w/ kPCA

- Kernel PCA is based on non-linear mapping of data to

$$\mathbf{x}_n \rightarrow \varphi(\mathbf{x}_n) \equiv \varphi_n, \qquad n = 1, \ldots, N$$

Aim is to locate maximum variance directions in the feature space, i.e.

$$\mathbf{l}_1 \equiv \arg\max_{\|\mathbf{l}\|=1} \left\langle \left(\mathbf{l}^T \cdot \varphi \right)^2 \right\rangle, \qquad \varphi(\mathbf{x}_n) = \sum_k \mathbf{l}_k s_{k,n}$$

The principal direction is in the span of data:

$$\mathbf{l}_1 = \sum_{n=1}^{N} a_{1,n} \varphi_n$$

$$\mathbf{a}_1 = \arg\max_{\|\mathbf{a}\|=1} \left\langle \mathbf{a}^T \cdot \mathbf{K} \cdot \mathbf{a} \right\rangle, \qquad \mathbf{K}_{n,n'} = \varphi_n^T \cdot \varphi_{n'} = \exp\left(-\frac{\|x_n - x_{n'}\|^2}{2c} \right)$$

TJ Abrahamsen and LK Hansen. "Input Space Regularization Stabilizes Pre-image for Kernel PCA De-noising". Proc. of Int. Workshop on Machine Learning for Signal Processing, Grenoble, France (2009).

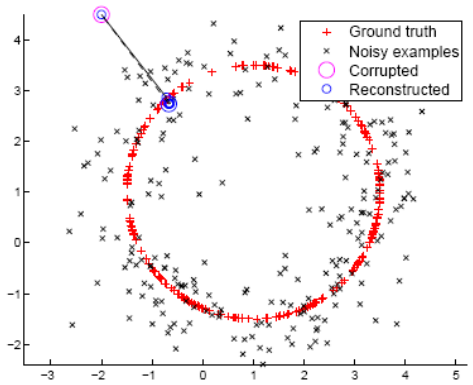
Manifold de-noising: The pre-image problem

Now, assume that we have a point of interest in feature space, e.g. a certain projection on to a principal direction "Φ", can we find its position "**z**" in measurement space?

$$\mathbf{z} = \varphi^{-1}(\phi)$$

Problems: (i) Such a point need not exist, and (ii) if it does there is no reason that it should be unique!

Mika et al. (1999): Find the closest match.



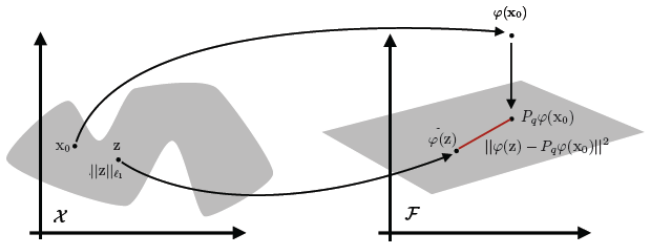


Figure 1: The pre-image problem in kernel PCA denoising concerns estimating $\mathbf{z}$ from $\mathbf{x}_0$, through the projection of the image onto the principal subspace in feature space, $\mathcal{F}$.

INPUT SPACE REGULARIZATION STABILIZES PRE-IMAGES FOR KERNEL PCA DE-NOISING

Trine Julie Abrahamsen *Lars Kai Hansen*

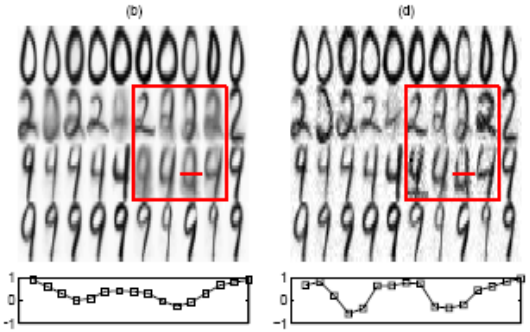


Fig. 4. Top: Example of de-noised digits using a very non-linear kernel ($c = 50$) and 100 principal components. (b) Mika et al and (d) our approach, note the visual improvement of the recovered pre-images in the red box. The colormap has been adjusted for better visualization. Bottom: The image intensity along the red line indicated above. Note the improvemed SNR in the result of the new method.

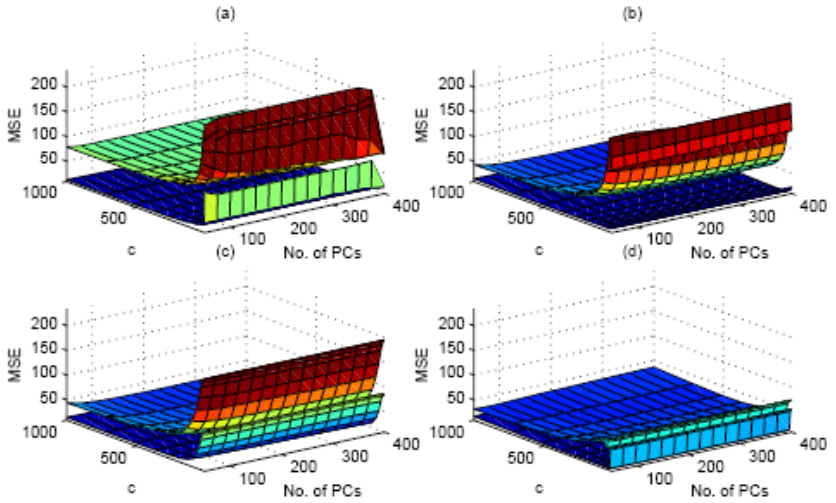


Fig. 2. Experiment to illustrate the stability of pre-image based de-noising of USPS digits. A training set of 400 digits (100@0, 2, 4, 9) is used to define the signal manifold. We show the confidence intervals (5th and the 95th percentile) for the mean square error (MSE) in different combinations of kPCA subspace dimension and non-linearity. MSE computed for 400 de-noised test samples for (a) Kwok-Tsang, (b) Mika et al., (c) Dambreville et al., and (d) the new input space distance regularization approach. The previous schemes are seen to deteriorate in the non-linear regime (small c).

TJ Abrahamsen and LK Hansen. Proc. of Int. Workshop on Machine Learning for Signal Processing, Grenoble, France (2009).

Regularization mechanisms for pre-image estimation in fMRI denoising

L2 regularization on denoising distance

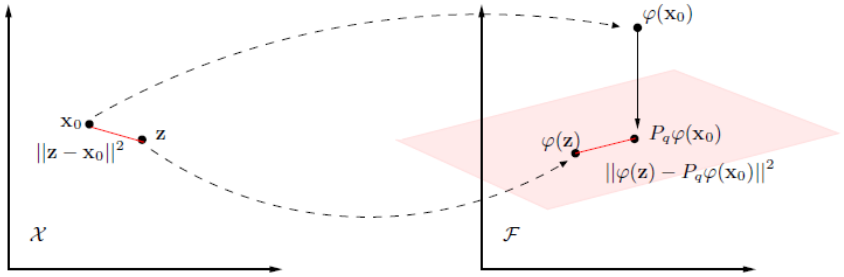


Figure 4.10: The pre-image problem in kernel PCA de-noising concerns estimating $\mathbf{z}$ from $\mathbf{x}_0$, through the projection of the image onto the principal subspace. Presently available methods for pre-image estimation lead to unstable pre-images because the inverse is ill-posed. We show that simple input space regularization, with a penalty based on the distance $||\mathbf{z}-\mathbf{x}_0||$ leads to a stable pre-image.

L1 regularization on pre-image

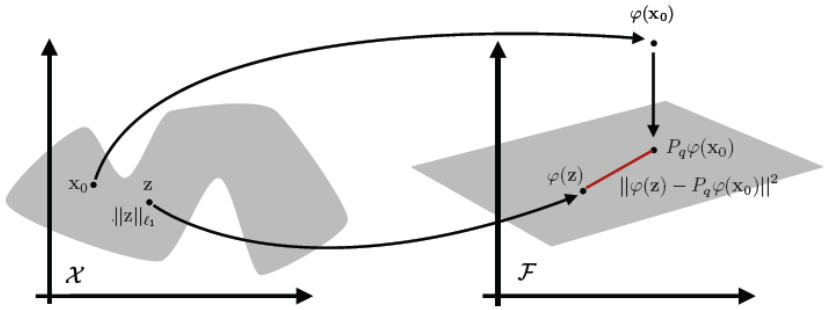


Figure 1: The pre-image problem in kernel PCA denoising concerns estimating $\mathbf{z}$ from $\mathbf{x}_0$, through the projection of the image onto the principal subspace in feature space, $\mathcal{F}$.

Beyond the linear model: Optimizing denoising using the PR-plot

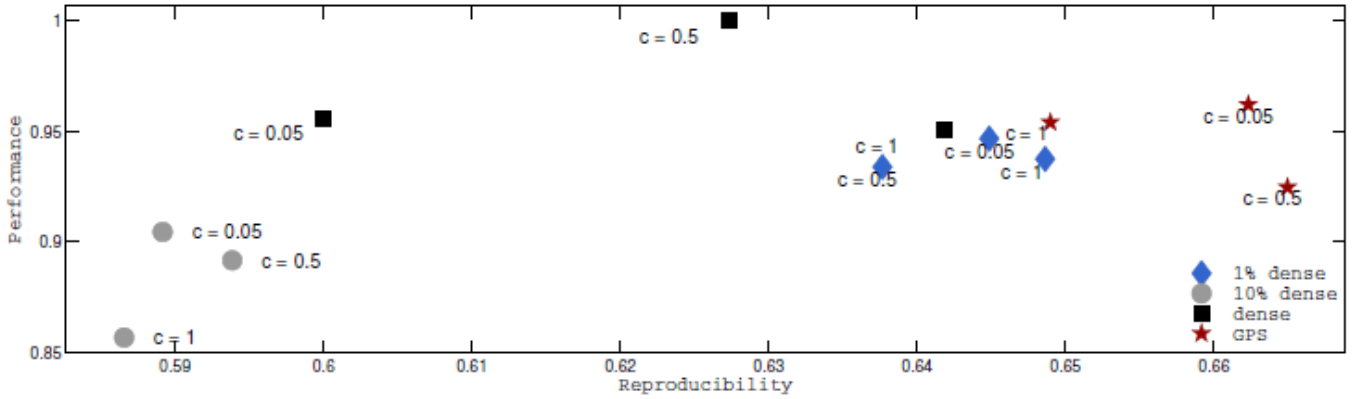


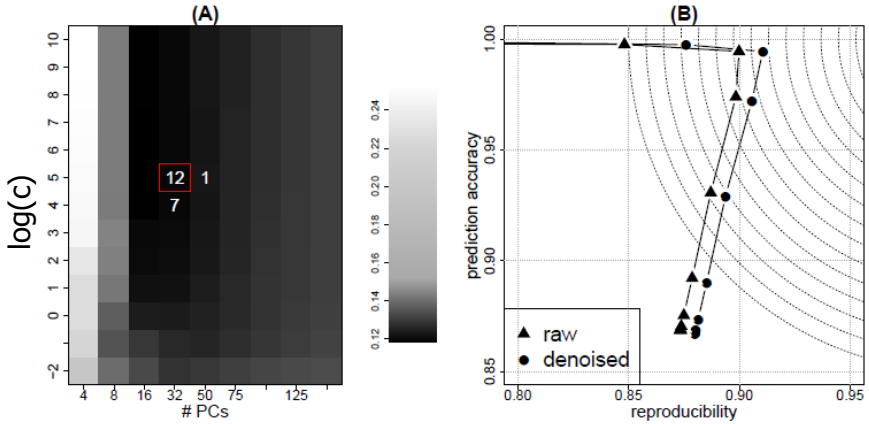
Figure 2: Prediction/reproducibility plots using all scans for the single slice fMRI visual block activation experiment. The GPS estimate when using a non-linear kernel are seen to outperform all other estimates in terms of combined prediction and reproducibility measures. Location in the upper right corner is preferred.

$$\mathbf{z} = \underset{\mathbf{z}\in\mathcal{X}}{\operatorname{argmin}} \, ||\varphi(\mathbf{z}) - P_q\varphi(\mathbf{x}_0)||^2 + \lambda||\mathbf{z}||_{\ell_1}.$$

GPS = General Path Seeking, generalization of the Lasso method

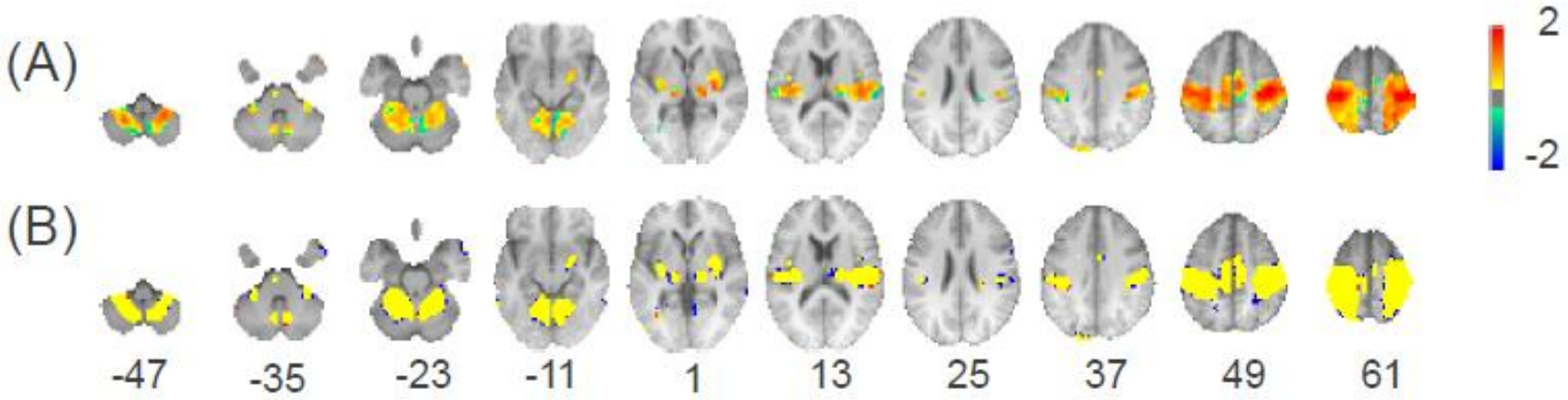
Jerome Friedman. Fast sparse regression and classification. Technical report, Department of Statistics, Stanford University, 2008.

Does denoising by kernel PCA help fMRI decoding?



(A) Comparison of resampling z-score = z-kPCA – z-Raw

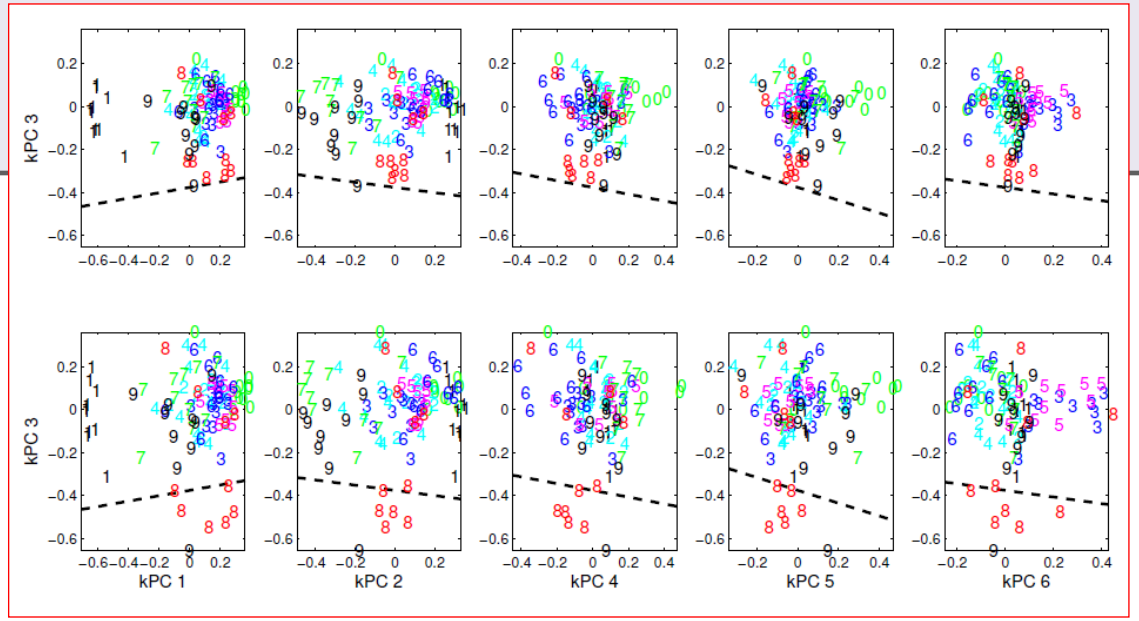
(B) FDR corrected:
yellow: consensus,
blue: only kPCA,
red: only raw

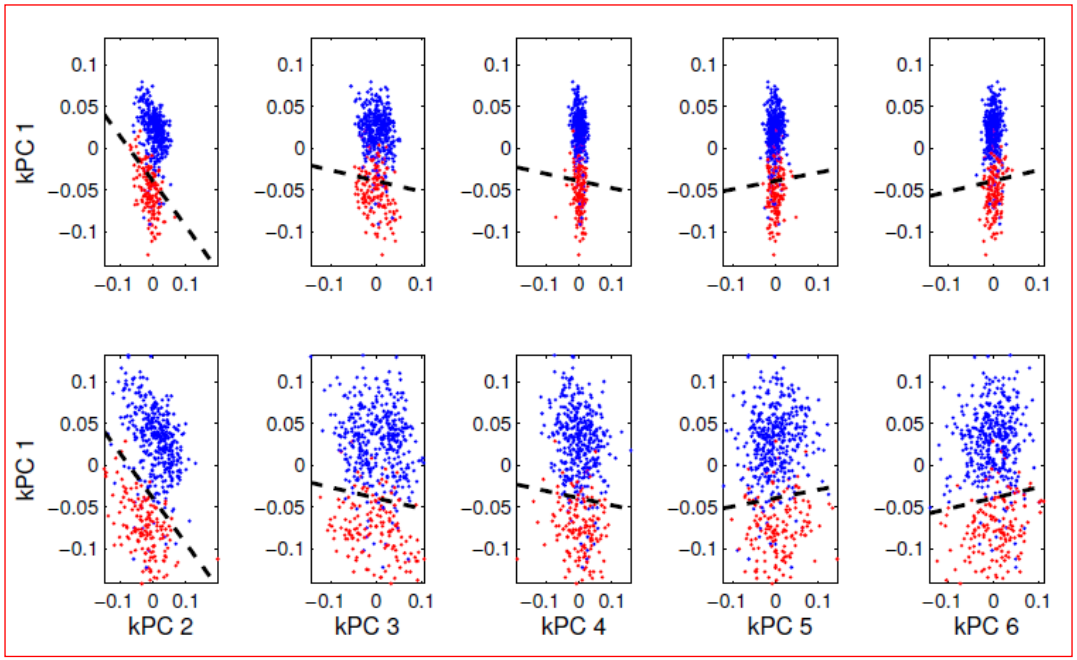


PM Rasmussen, TJ Abrahamsen, KH Madsen, LK Hansen: Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-imageestimation, NeuroImage 60(3):1807-1818 (2012).

Variance inflation in kernel PCA!

Handwritten digits:





fMRI data

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

TJ. Abrahamsen, LK. Hansen. A Cure for Variance Inflation in High Dimensional Kernel Principal Component Analysis. Journal of Machine Learning Research 12:2027-2044 (2011).

Cimbi Center for integrated molecular brain imaging

Lars Kai Hansen

IMM, Technical University of Denmark

Challenges from more complex data structures …Data represented as multiway arrays

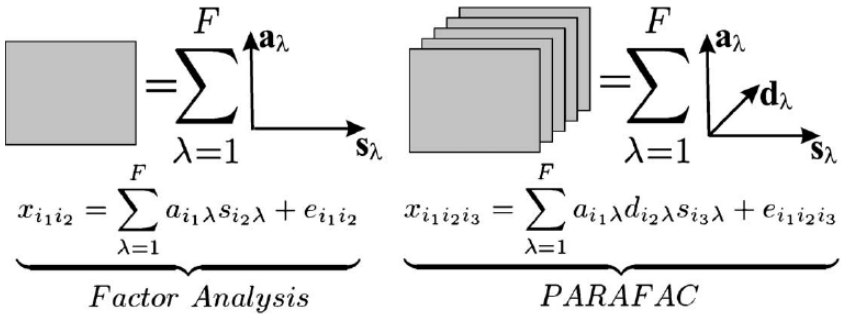


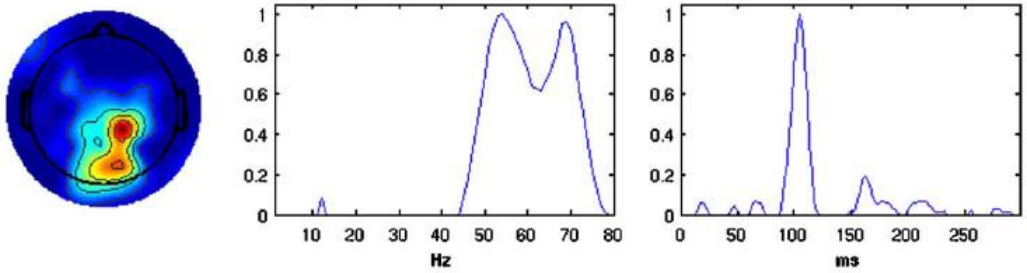
Fig. 1. Graphical representation of the factor analysis to the left and the PARAFAC decomposition of a 3-way array to the right. Like the factor analysis, PARAFAC decomposes the data into factor effects pertaining to each modality. F denotes the number of factors.

A full linear factor model may be overparametrized

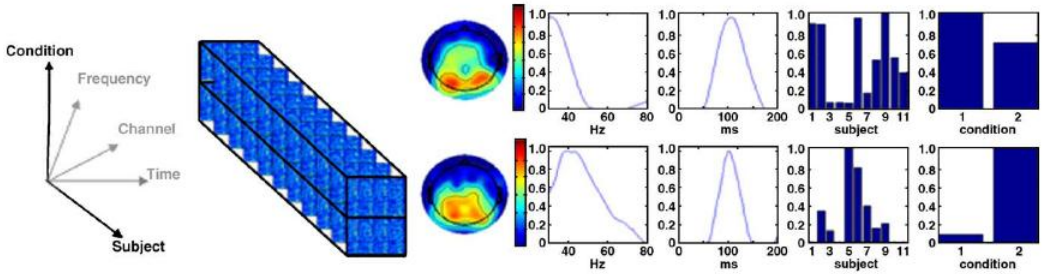
Example

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



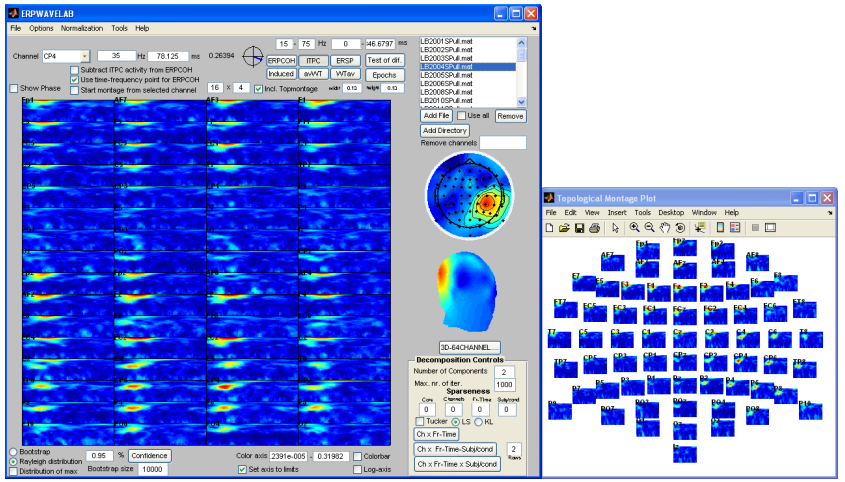
Mørup et al. NeuroImage (2005), NeuroImage (2008)

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),

Supervised learning

Retrieval of relevant patterns p(s|x)

Generalizable supervised models - 'mind reading'

- Non-linear kernel machines, SVM

Local voting +/-

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

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

S. Mika, C. Schäfer, P. Laskov, D. Tax, K. R. Müller, Support Vector Machines Handbook of Computational Statistics, Springer, Berlin, 2004

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 (may be fixed though!).

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

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

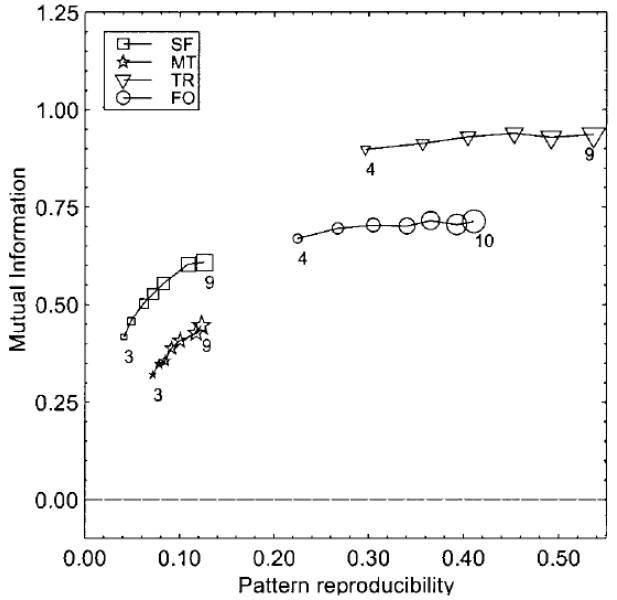
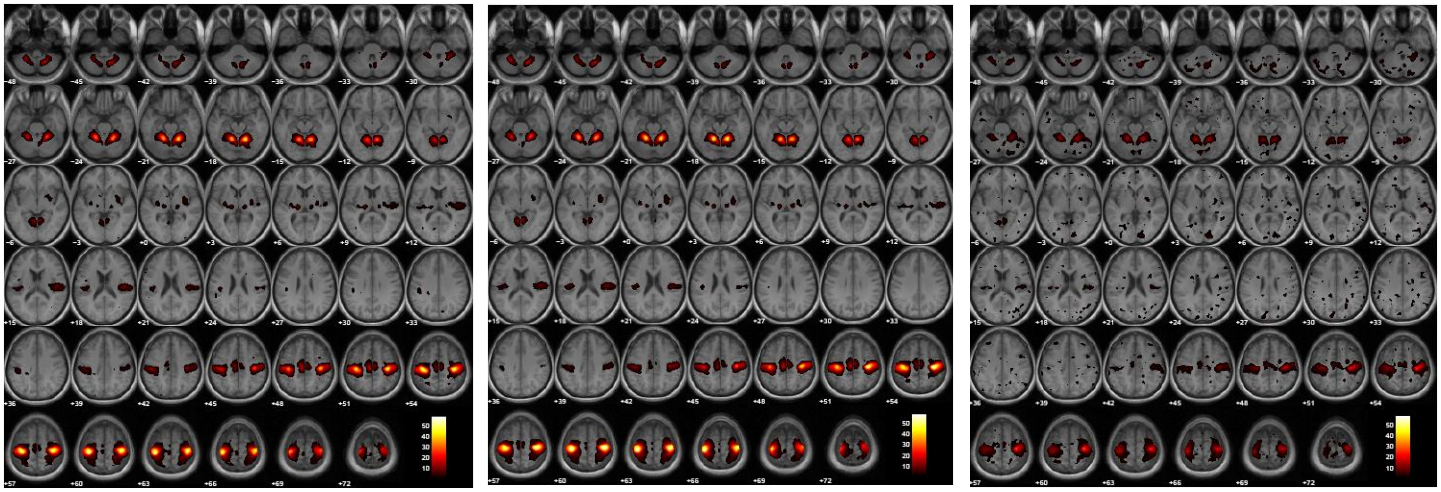


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both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

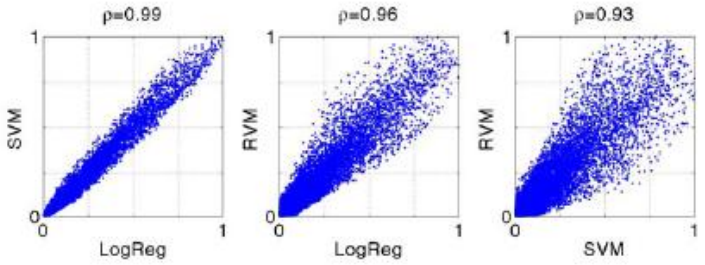


Figure 7: fMRI fingertapping experiment - consensus analysis. The plots show the extend of consensus in the average rSPI among the three models. The rSPI for LogReg was scaled by its maximum value. Hereafter the rSPIs from the SVM and RVM were transformed to match the histogram of that of LogReg. Correlation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

Sensitivity maps for non-linear kernel regression

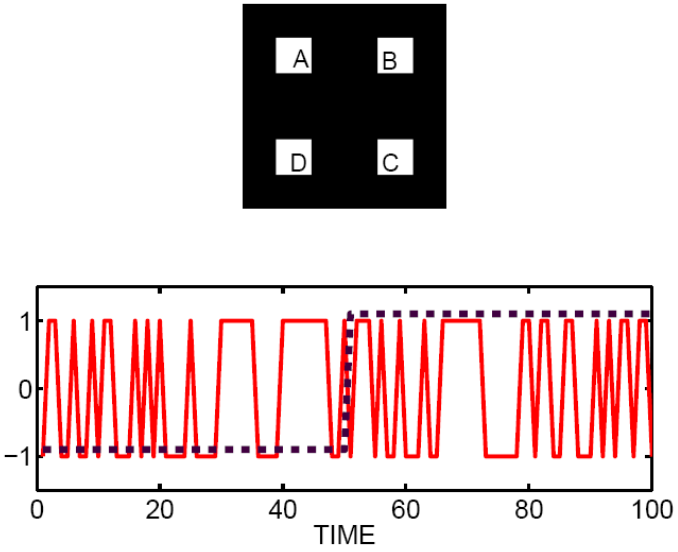


Fig. 1. XOR-image set define by four activated regions (A,B,C,D). Initially we let regions (A,B,D) be activated by random sequence taking values ± 1, as shown in example in the bottom panel (full curve). The target signal, also taking values $t_n = pm1$, and is also indicated in the bottom panel (dashed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

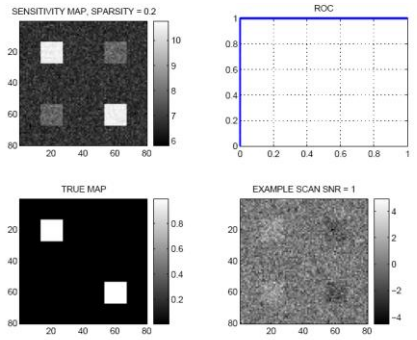


Fig. 2. XOR-image set define by four activated regions. The results of analyzing a image set with $N = 400$ examples. The image signal-to-noise ratio is $SNR = 1$, i.e., the additive noise is unit variance. The target function has in addition been contaminated by 10% random label noise. The four subplots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

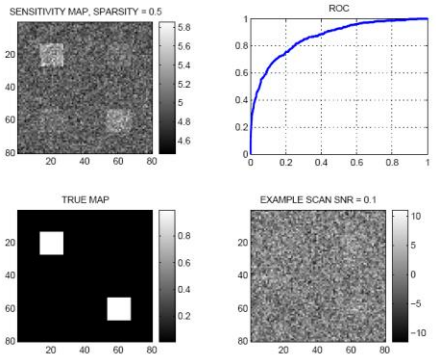
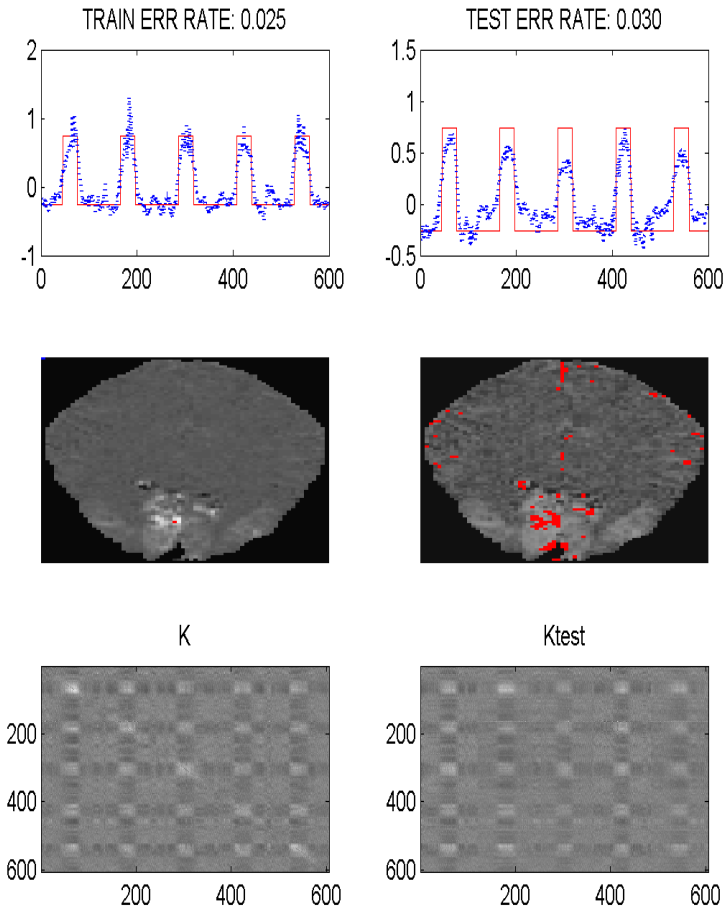


Fig. 4. XOR-image set define by four activated regions. Similar to figure 2, however the image signal-to-noise ratio is $SNR = 0.1$.

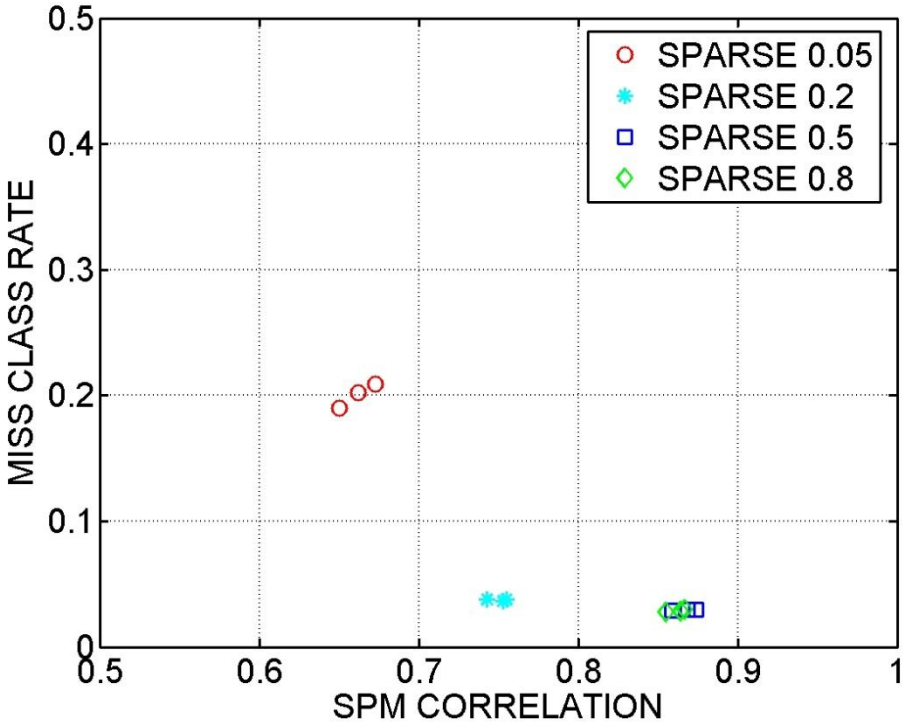
Initial dip data: Visual stimulus (TR 0.33s)

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



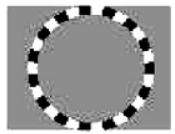
Initial dip data: Visual stimulus (TR 0.33s)

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



Non-linearity in fMRI?

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



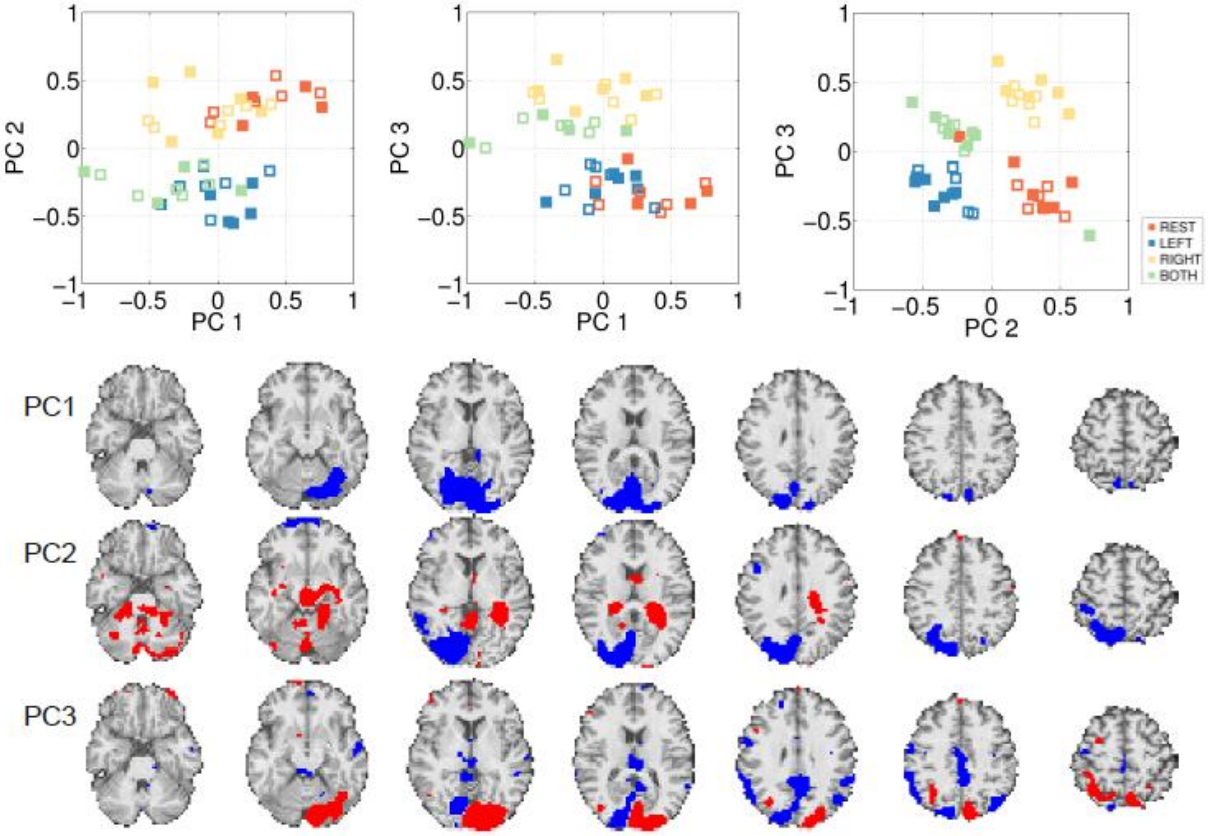
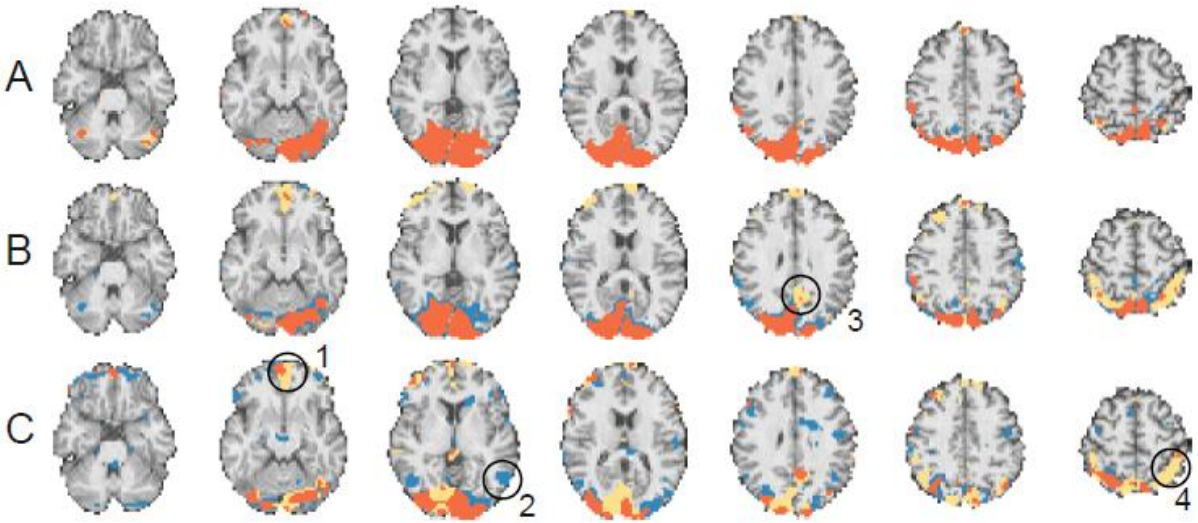
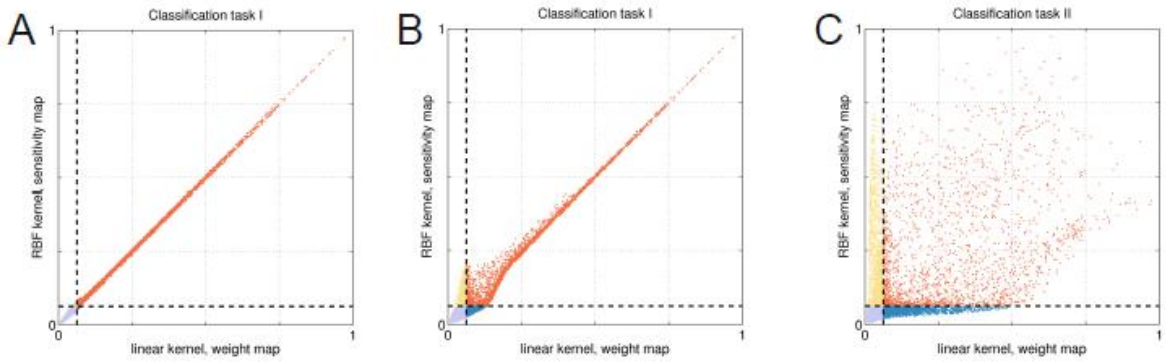


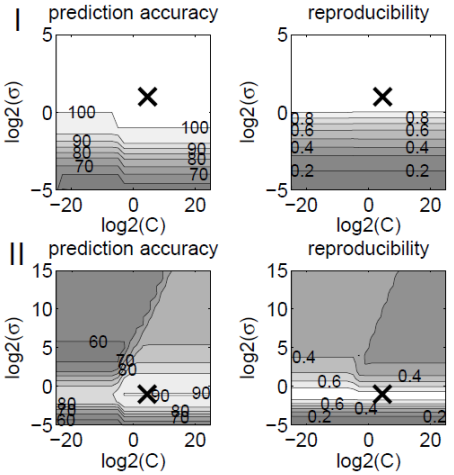
Figure 1: PCA analysis of the fMRI data set. An example of the three first PCs estimated from the training set in a NPAIRS split. The scatter plots show both training (filled markers) and test examples projected onto the PCs. The blue and red voxels on the brain slices corresponds to negative and positive 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. NeuroImage 55 (2011) 1120–1131







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

Conclusions

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

Numerous challenges remaining in neuroimage modelling

=> outlook

Outlook – future of mind reading?

- More ecological valid imaging conditions
- Long time observations in the ”wild”
- EEG real time 3D imaging for bio-feedback
- 24/7 monitoring

Fig. 1. Handheld brain scanner components. Emotiv EPOC wireless EEG headset (1), Emotiv Receiver module with USB connector (2), USB connector and adapter (3+4), and Nokia N900 mobile phone. The total cost of the system is less than USD1000.

Illustration of
HypoSafe implantable device

hypo safe®

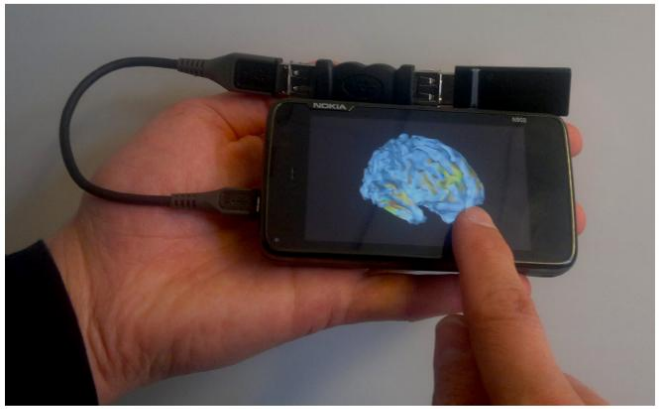


Fig. 3. A user interacting with a 3D model of the brain using the handheld brain scanner device with touch-based interaction.

Acknowledgments

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

www.imm.dtu.dk/~lkh
hendrix.imm.dtu.dk