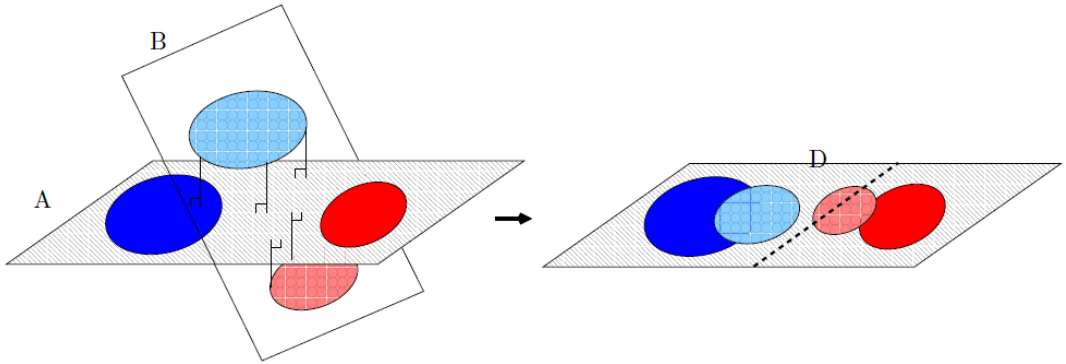
Learning from small samples in high dimensions

Lars Kai Hansen

DTU Informatics
Technical University of Denmark



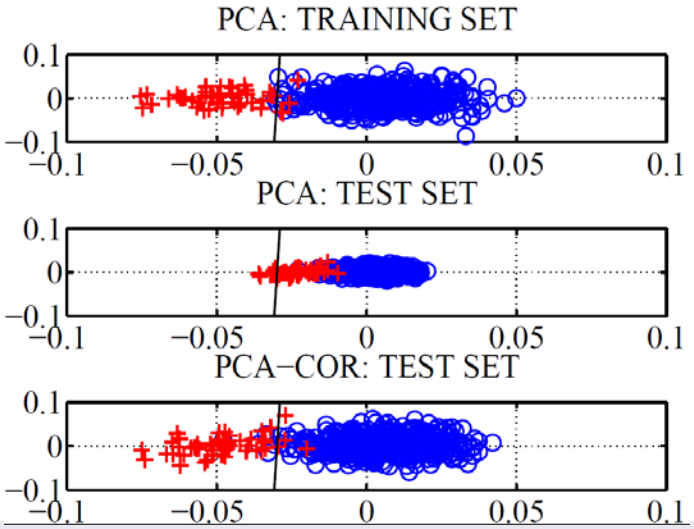
Co-workers:
Trine Abrahamsen, Stephen Strother

Cimbi
Center for integrated molecular brain imaging

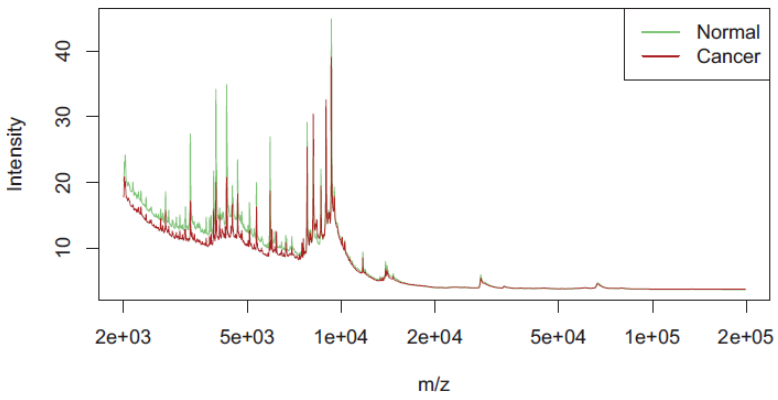
DTU

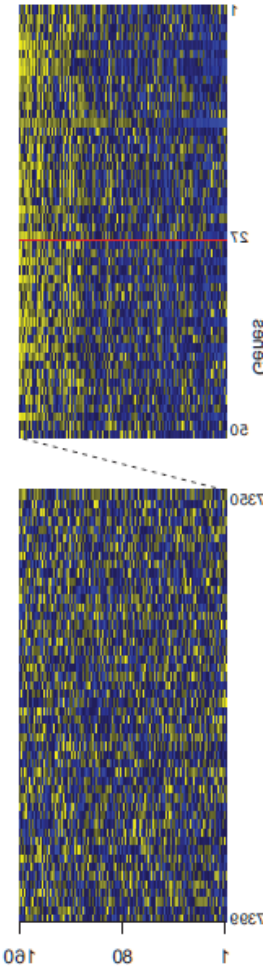
OUTLINE

- High dimensions and small samples
- Variance inflation in PCA
- Variance inflation in kPCA
- Implications for SVMs



High dimensions – small samples (D>>N)





”HDLSS” high dimension, low sample size (Hall 2005, Ahn et al, 2007)

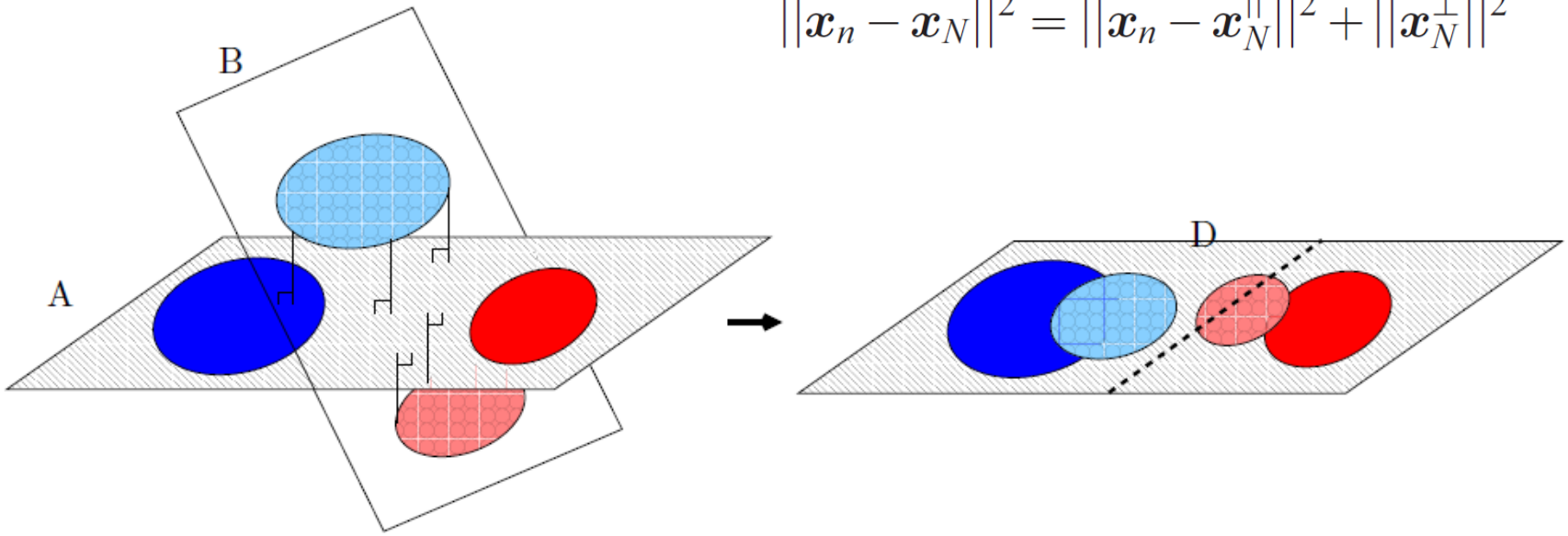
”Large p, small n” (West, 2003), ”Curse of dimensionality” (Occam, 1350)

”Large underdetermined systems” (Donoho, 2001)

”Ill-posed data sets” (Kjems, Strother, LKH, 2001)

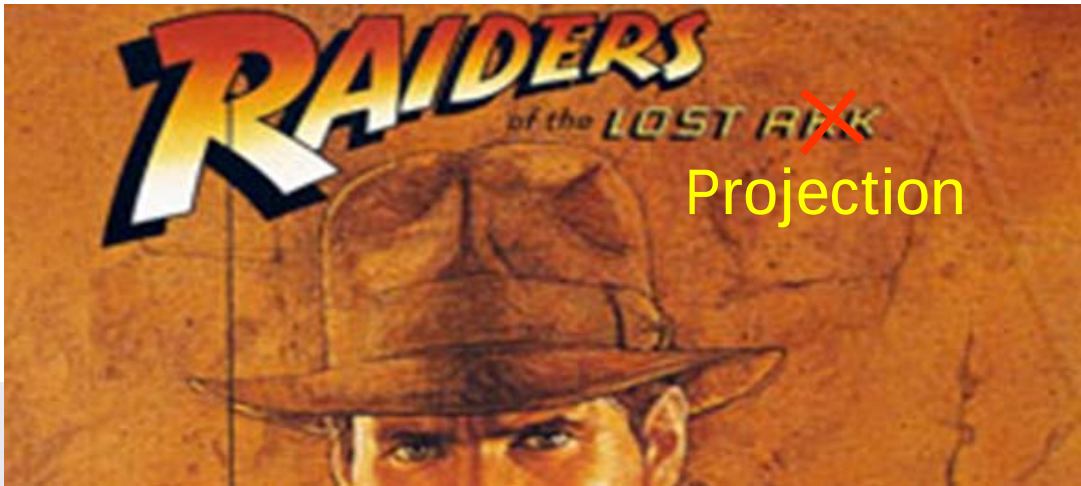
What is the challenge? The lost projection

$$||\boldsymbol{x}_n - \boldsymbol{x}_N||^2 = ||\boldsymbol{x}_n - \boldsymbol{x}_N^{||}||^2 + ||\boldsymbol{x}_N^{\perp}||^2$$



”Ill-posed data sets” (Kjems, Strother, LKH, 2001)

”Variance Inflation” (Abrahamsen, LKH, 2011)



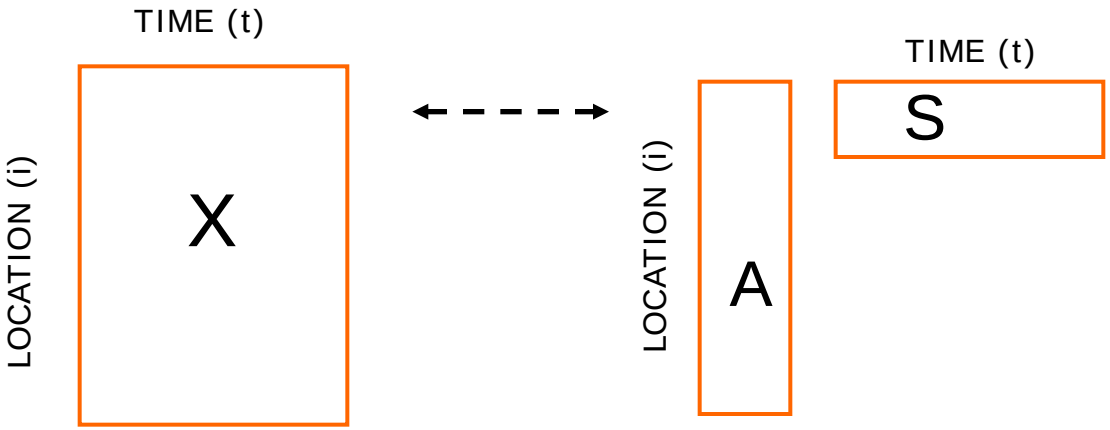
Cimbi Center for integrated molecular brain imaging

Lars Kai Hansen

IMM, Technical University of Denmark

Factor models

- Represent a datamatrix by a low-dimensional approximation



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

Unsupervised learning:
Factor analysis generative model

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

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

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

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 Computation (2002),
Neurocomputing (2002)

Matrix factorization: SVD / PCA, NMF, Clustering

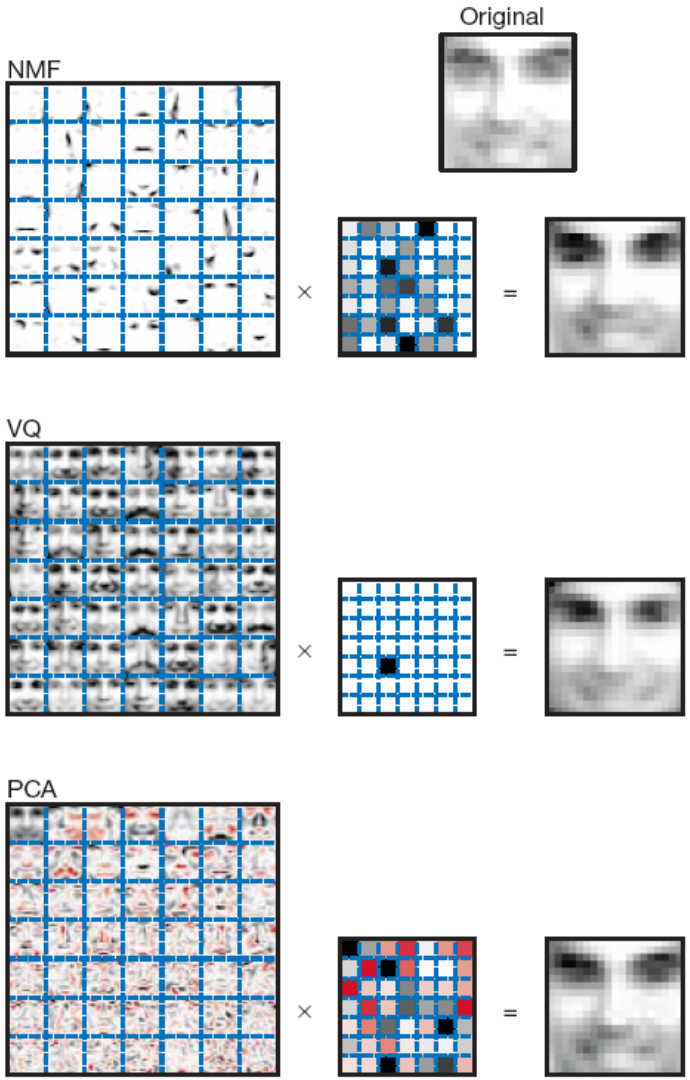


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting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

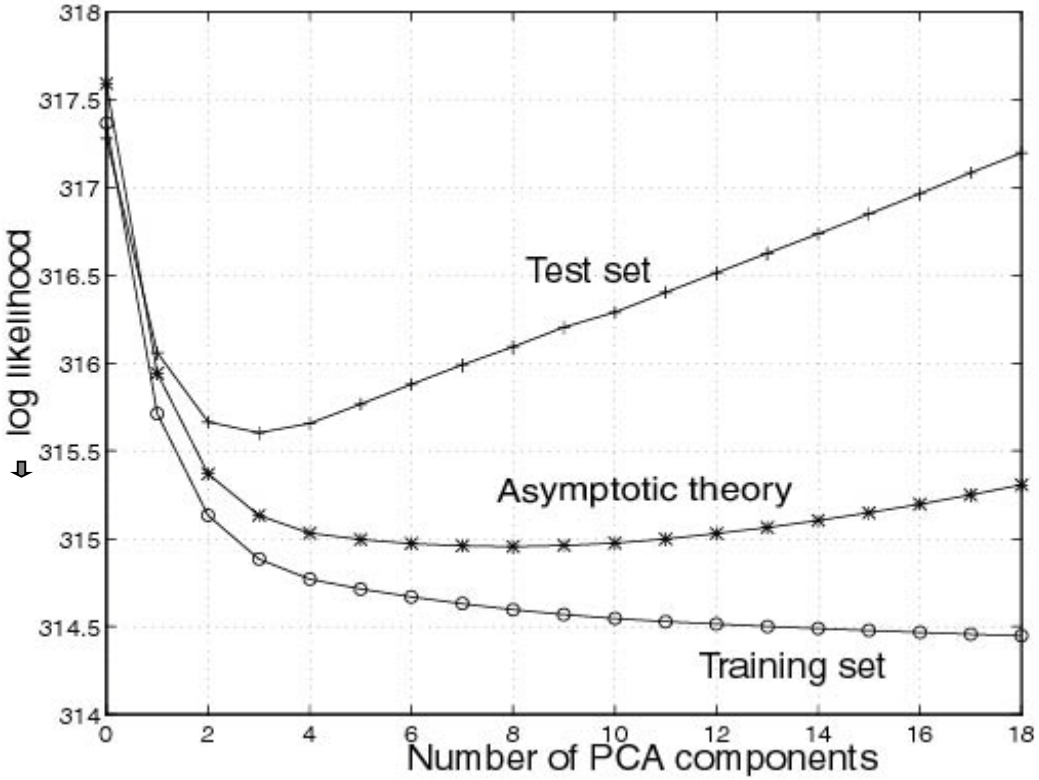
Learning the parts of objects by non-negative matrix factorization

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

Bias-variance trade-off as function of PCA dimension in fMRI data



Hansen et al. *NeuroImage* (1999)

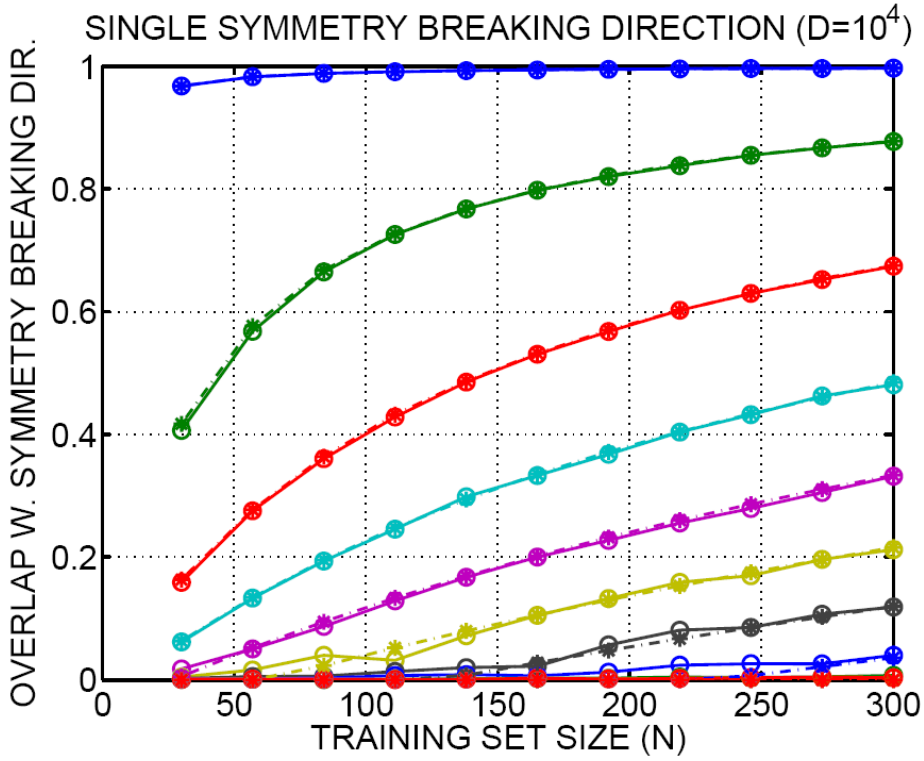
Modeling the generalizability of SVD

- 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)

Cimbi Center for integrated molecular brain imaging

Lars Kai Hansen

IMM, Technical University of Denmark

DTU

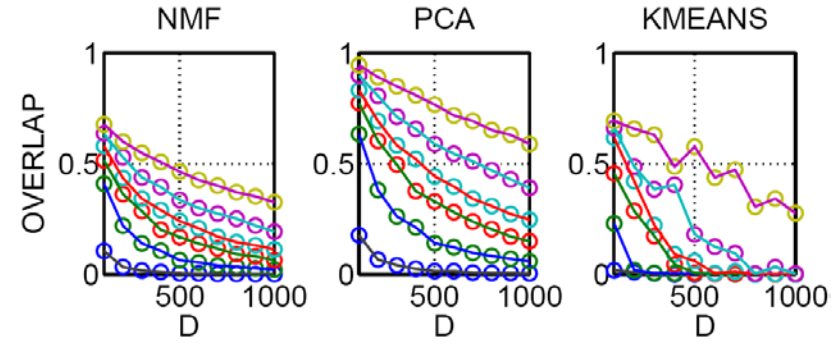
Universality in beyond PCA: NMF, Kmeans

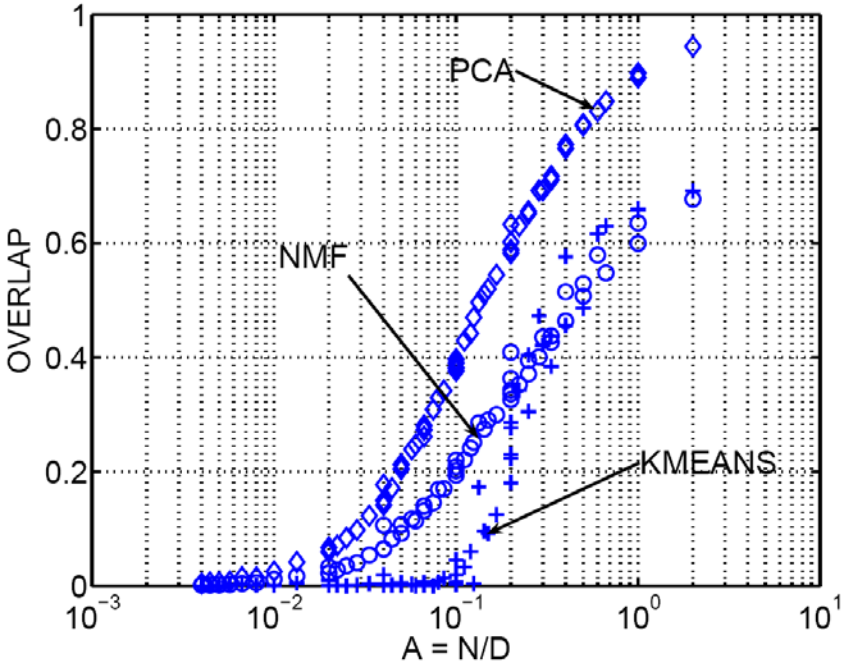
- Looking for universality by simulation
  - learning two clusters in white noise.
- Train K=2 component factor models.
- Measure overlap between line of sigth and plane spanned by the two factors.

Experiment

Variable: N, D

Fixed: SNR

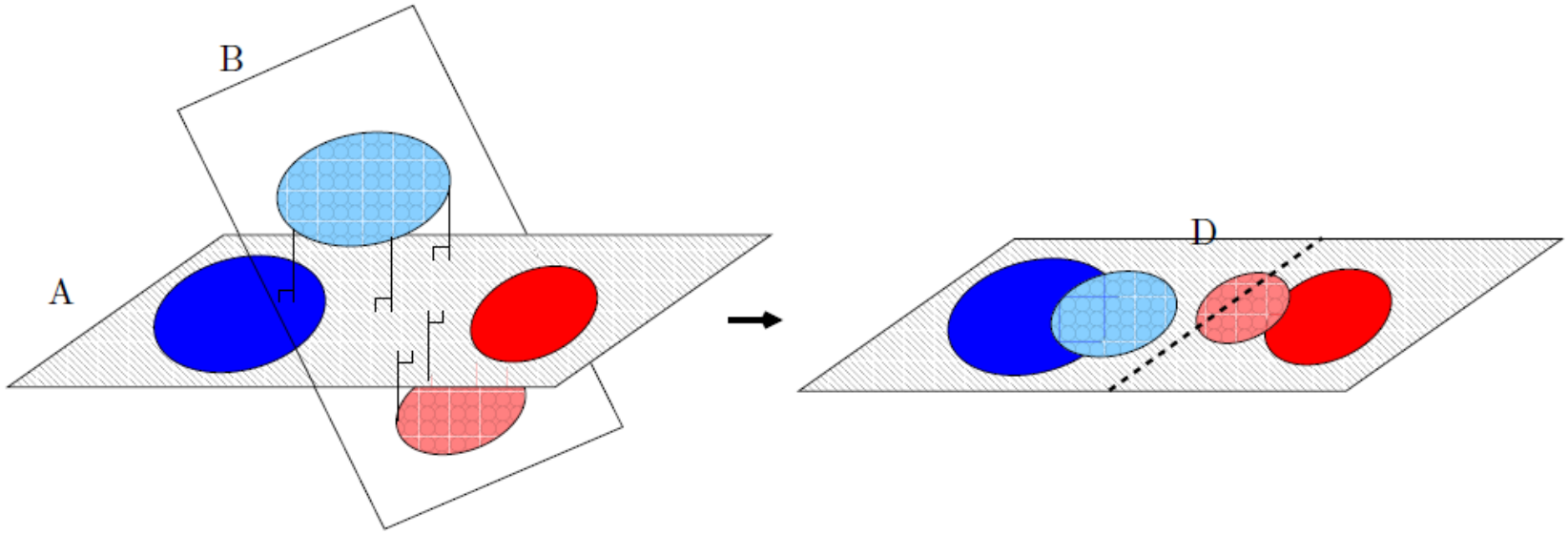




Lars Kai Hansen

IMM, Technical University of Denmark

Variance inflation in PCA



Journal of Machine Learning Research 12 (2011) 2027-2044 Submitted 1/11; Published 6/11

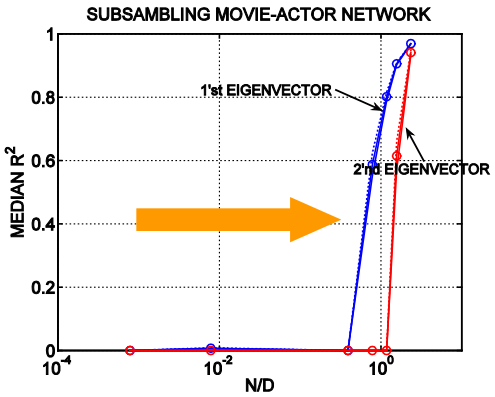
A Cure for Variance Inflation in High Dimensional Kernel Principal Component Analysis

Trine Julie Abrahamsen TJAB@IMM.DTU.DK
Lars Kai Hansen LKH@IMM.DTU.DK
DTU Informatics
Technical University of Denmark
Richard Petersens Plads, 2800 Lyngby, Denmark

Who shrunk the test set?

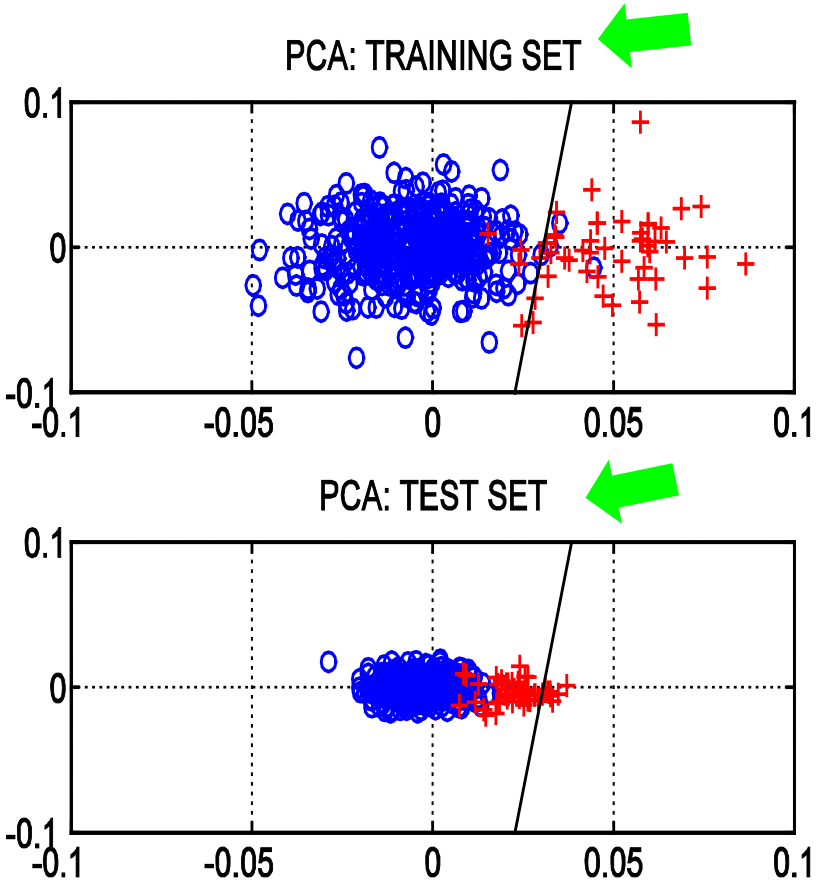
Restoring the generalizability of SVD

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

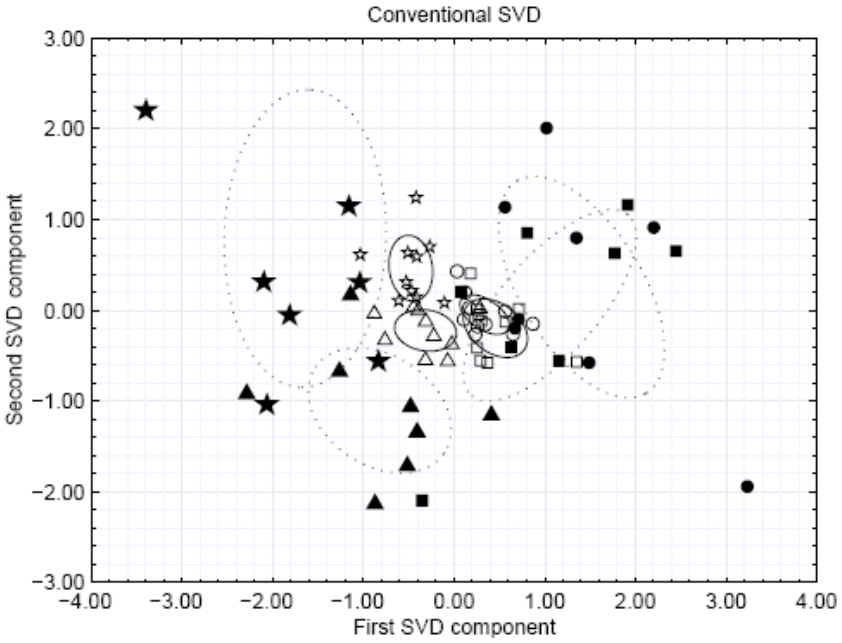


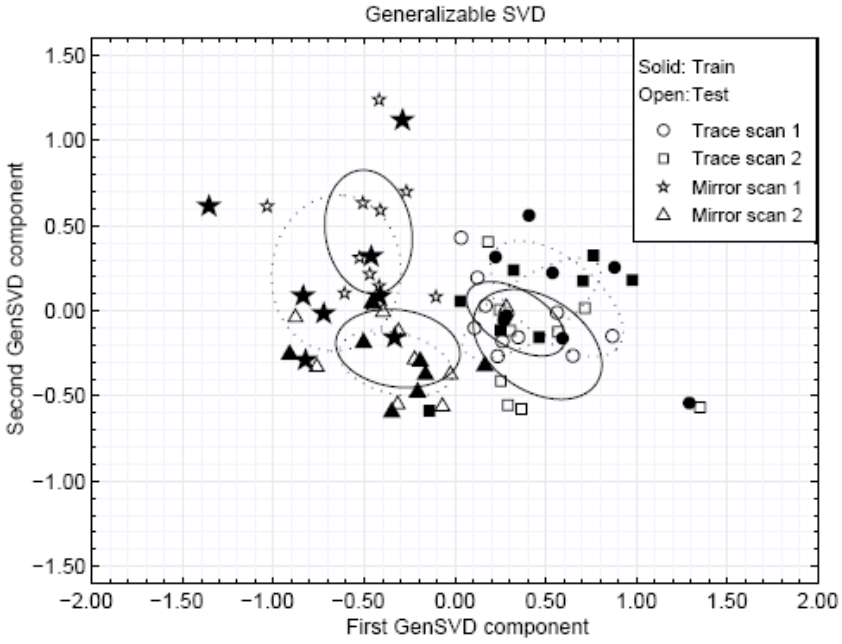
- The estimated projection is offset, hence, future projections will be too small!

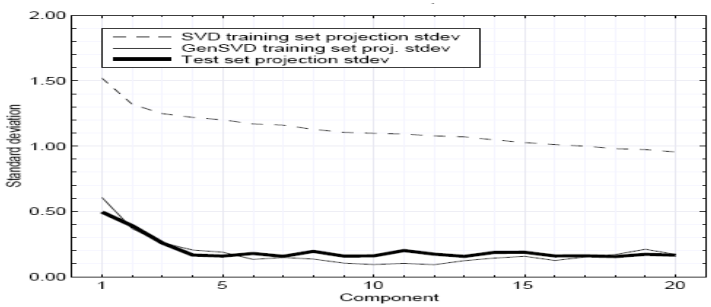
- ...problem if discriminant is optimized for unbalanced classes in the training data!



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

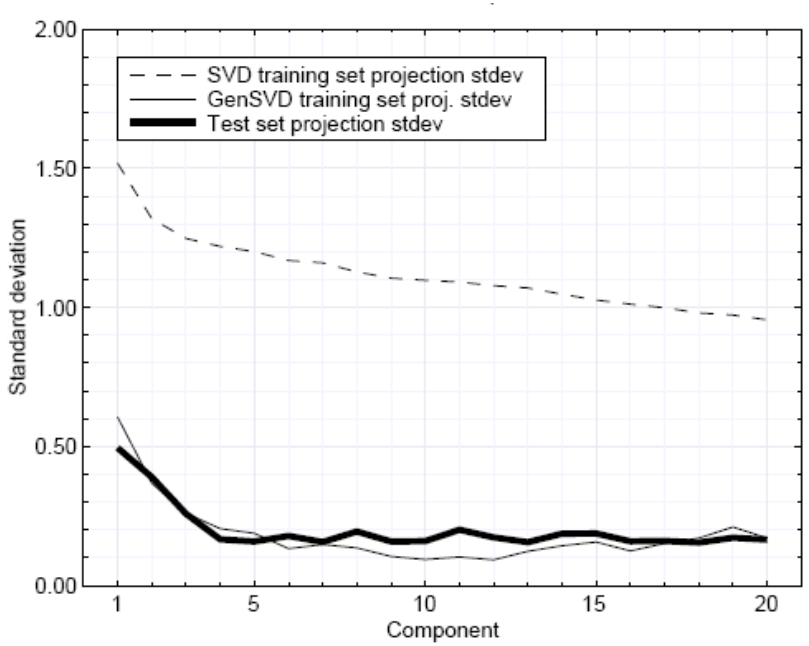
Lars Kai Hansen

IMM, Technical University of Denmark

DTU

Re-scaling the component variances by leave one out

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



Kjems, Hansen, Strother: NIPS (2001)

Approximating LOO (leave-one-out: ”N”)

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be N training data points in a D dimensional input space

$$\boldsymbol{x}_N = \boldsymbol{x}_N^{\perp} + \boldsymbol{x}_N^{\parallel}, \qquad \boldsymbol{u}_{N-1,k}^T \cdot \boldsymbol{x}_N = \boldsymbol{u}_{N-1,k}^T \cdot \boldsymbol{x}_N^{\parallel} ,$$

$$\boldsymbol{u}_{N-1,k}^T \cdot \boldsymbol{x}_N = \boldsymbol{u}_{N-1,k}^T \cdot \boldsymbol{x}_N^{\parallel} \approx \boldsymbol{u}_{N,k}^T \cdot \boldsymbol{x}_N^{\parallel}$$

Two approximations

Adjusting for the mean overlap

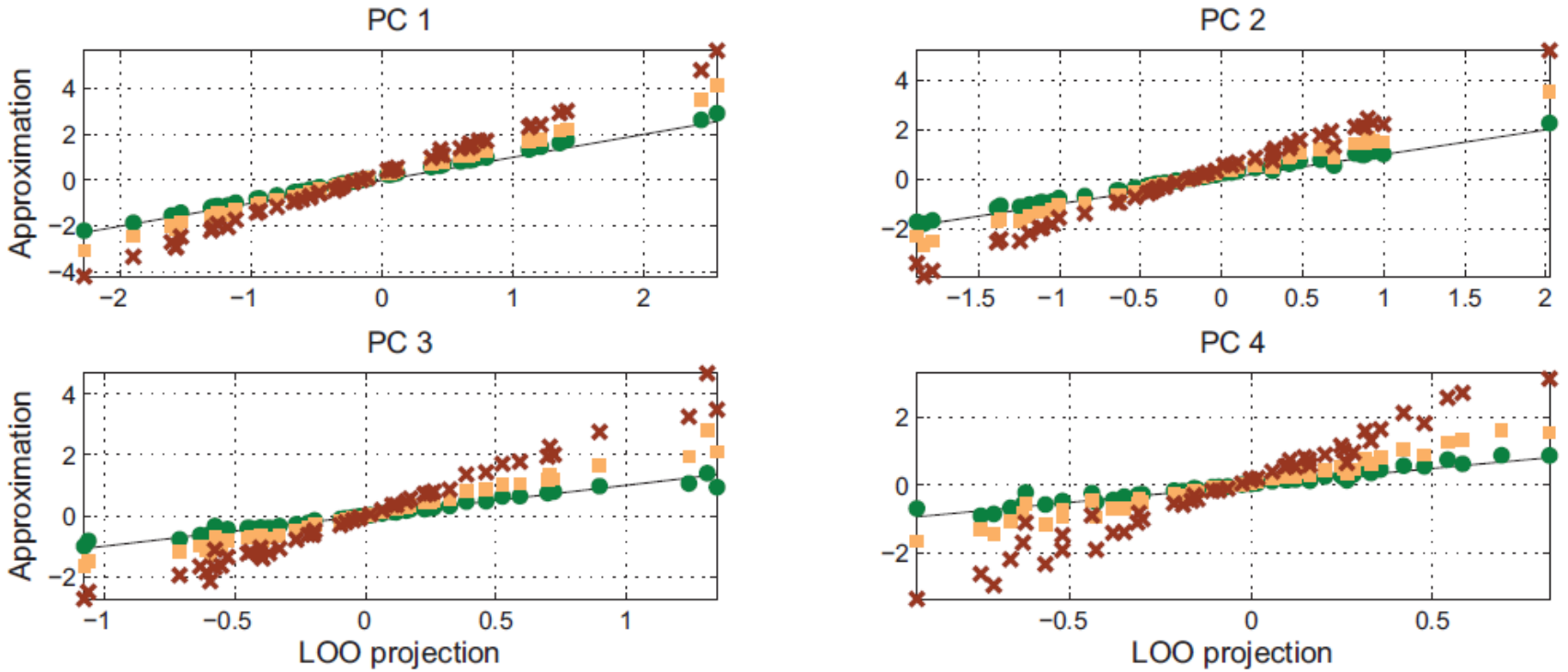
$$R^2 = \begin{cases} (\alpha S^2 - 1)/S(1+\alpha S) & \alpha > 1/S^2 \\ 0 & \alpha \le 1/S^2 \end{cases}$$

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

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

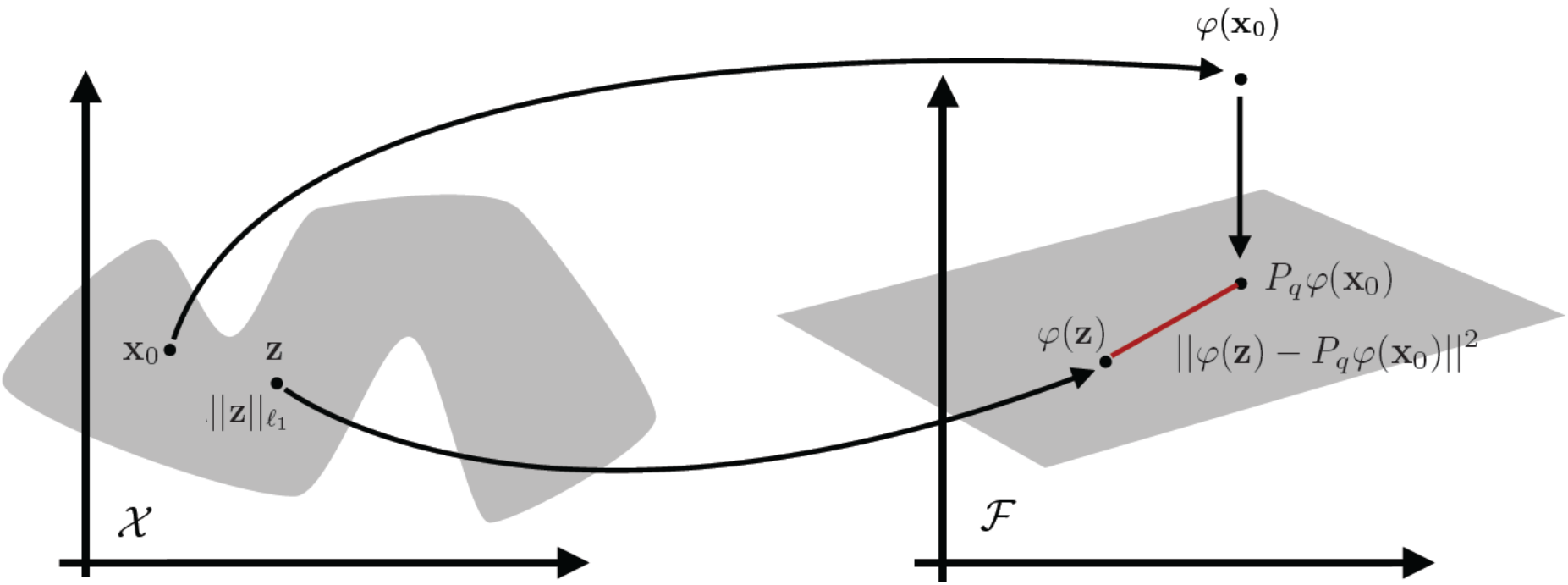
Adjusting for lost projection

$$\boldsymbol{u}_{N-1,k}^T \cdot \boldsymbol{x}_N = \boldsymbol{u}_{N-1,k}^T \cdot \boldsymbol{x}_N^{\parallel} \approx \boldsymbol{u}_{N,k}^T \cdot \boldsymbol{x}_N^{\parallel}$$



Approximating the leave-one-out (LOO) procedure. Here we simulate data with four normal independent signal components, $x = \Sigma_{k=1}^{4} \eta_k u_k + \epsilon$ of strengths $(1.4, 1.2, 1.0, 0.8,$ embedded in i.i.d. normal noise $\epsilon \sim N(0, \sigma^2 1)$, with $\sigma = 0.2$. The dimension was $D = 2000$ and the sample size was $N = 50$. In the four panels we show the training set projections (red crosses), the projections corrected for the theoretical mean overlap (Hoyle and Rattray, 2007) (yellow squares) and the geometric approximation in Equation (1) (green dots) versus the exact LOO projections (black line).

Beyond the linear model: Non-linear denoising



TJ Abrahamsen and LK Hansen. Sparse non-linear denoising: Generalization performance and pattern reproducibility in functional MRI Pattern Recognition Letters 32(15) 2080-2085 2011

Beyond the linear model:

- 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:

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

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

Schölkopf et al. : Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Comp (1998)

Approximating the LOO cure for kPCA

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be N training data points

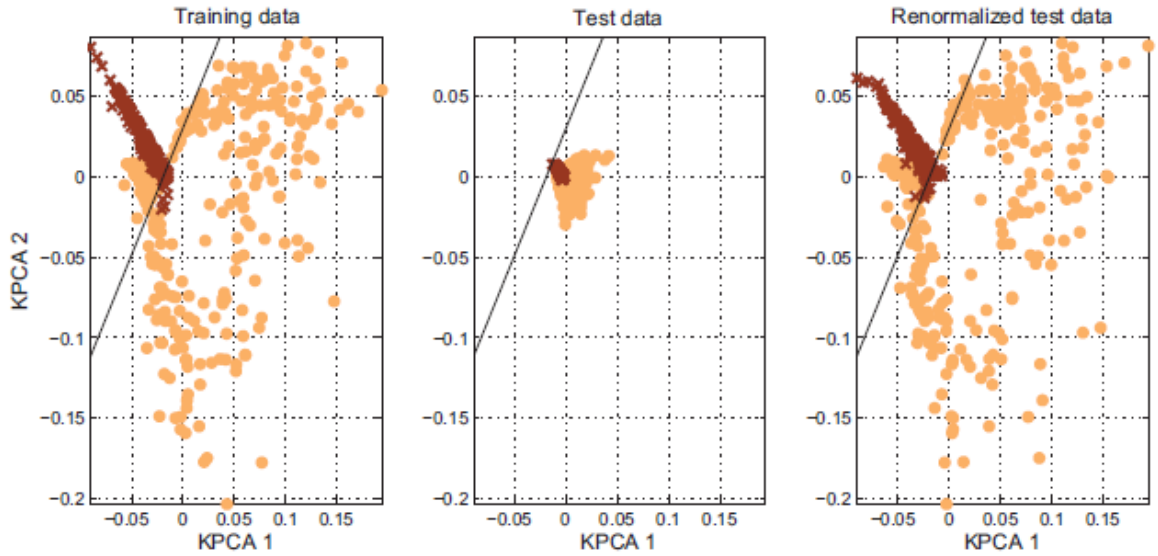
$$\tilde{\varphi}(\boldsymbol{x}) = \varphi(\boldsymbol{x}) - \bar{\varphi}.$$

$$\widetilde{\boldsymbol{K}} = \boldsymbol{K} - \frac{1}{N}\mathbf{1}_{NN}\boldsymbol{K} - \frac{1}{N}\boldsymbol{K}\mathbf{1}_{NN} + \frac{1}{N^2}\mathbf{1}_{NN}\boldsymbol{K}\mathbf{1}_{NN}$$

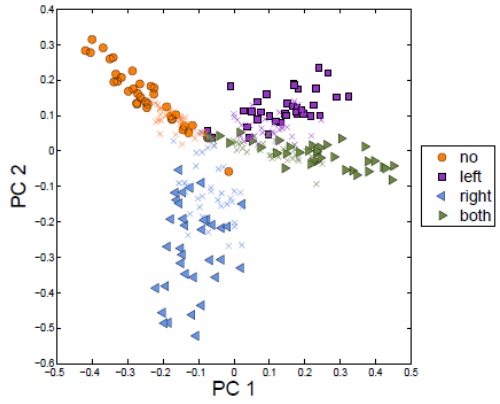
$$\widetilde{\boldsymbol{K}}\boldsymbol{\alpha}_i = \lambda_i \boldsymbol{\alpha}_i$$

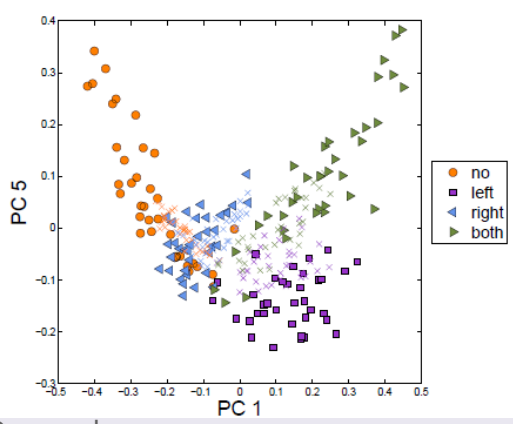
$$\beta_i = \tilde{\varphi}(\boldsymbol{x})^T \boldsymbol{v}_i = \sum_{n=1}^{N} \alpha_{in} \tilde{\varphi}(\boldsymbol{x})^T \tilde{\varphi}(\boldsymbol{x}_n) = \sum_{n=1}^{N} \alpha_{in} \tilde{k}(\boldsymbol{x}, \boldsymbol{x}_n)$$

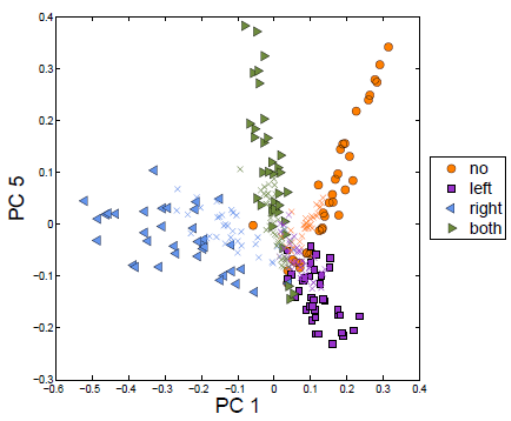
Application to classification of high-dimensional data on manifolds



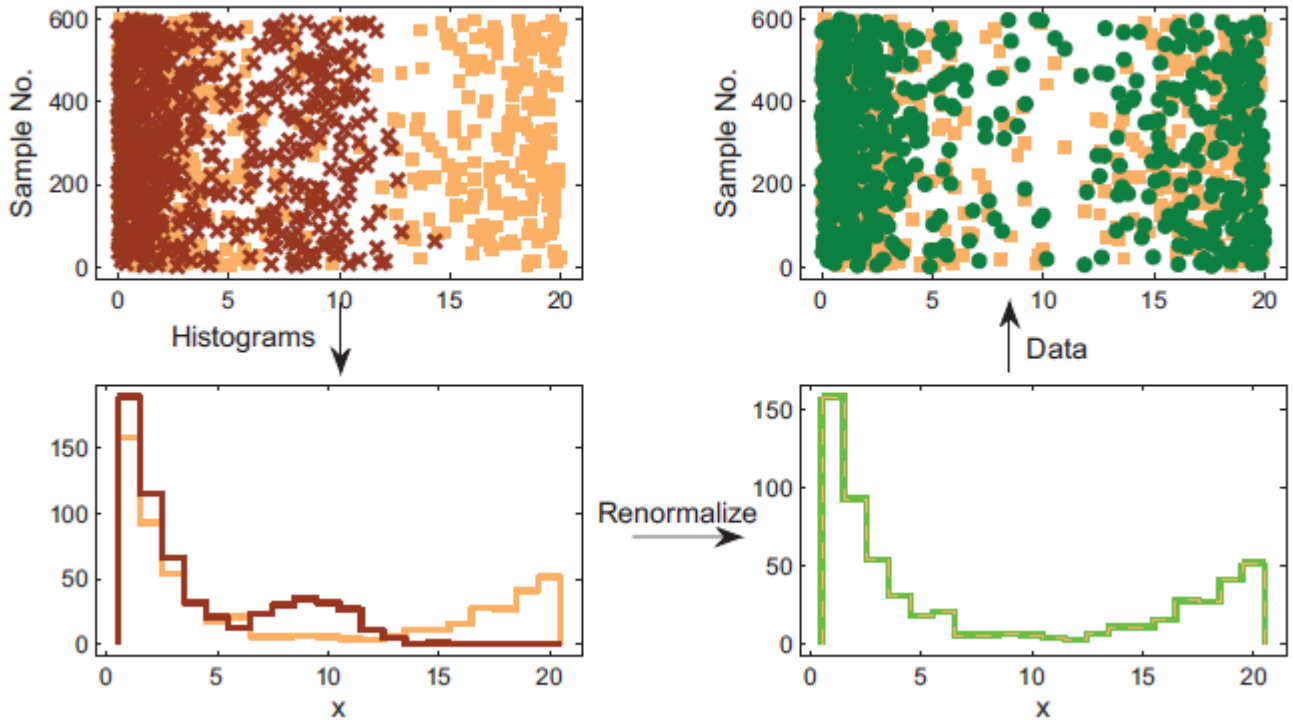
XOR fMRI: D= 75,257 N=576





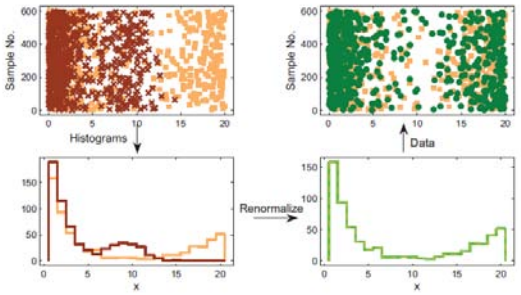


Non-parametric histogram equalization




```
>> [as,ia]=sort(a);
>> [bs,ib]=sort(b);
>> b(ix)=as;
```

Non-parametric histogram equalization



Algorithm 1 Approximate renormalization in kernel PCA

Require: $\boldsymbol{X}_{tr}$ and $\boldsymbol{X}_{te}$ to be $N_{tr} \times D$ and $N_{te} \times D$ respectively

Compute $\widetilde{\boldsymbol{K}}_{tr}$ using Equation (2) and find the eigenvectors, $\alpha_1, \ldots, \alpha_q$

for $i = 1$ to N_{tr} **do**

$f_{tr}^{i,:} \leftarrow P_q(x_{tr}^{i,:}) = \tilde{k}_{x_i}^T \alpha^{i:q}$ {see Equation (3)}

end for

for $j = 1$ to N_{te} **do**

$f_{te}^{j,:} \leftarrow P_q(x_{te}^{j,:}) = \tilde{k}_{x_j}^T \alpha^{i:q}$ {see Equation (3)}

end for

for $d = 1$ to q **do**

$[f_{sort}, \ ] \leftarrow \text{sort}(f_{tr}^{:,d})$ {ascending order}

$[\ , I] \leftarrow \text{sort}(f_{te}^{:,d})$ {ascending order}

if $N_{tr} = N_{te}$ **then**

$h \leftarrow f_{sort}$

else $\{N_{tr} \neq N_{te}\}$

$h \leftarrow \text{spline}\big([1 : N_{tr}], f_{sort}, \text{linspace}(1, N_{tr}, N_{te})\big)$ {interpolate to create N_{te} values of f_{sort} in the interval $[1 : N_{tr}]$}

end if

for $n = 1$ to N_{te} **do**

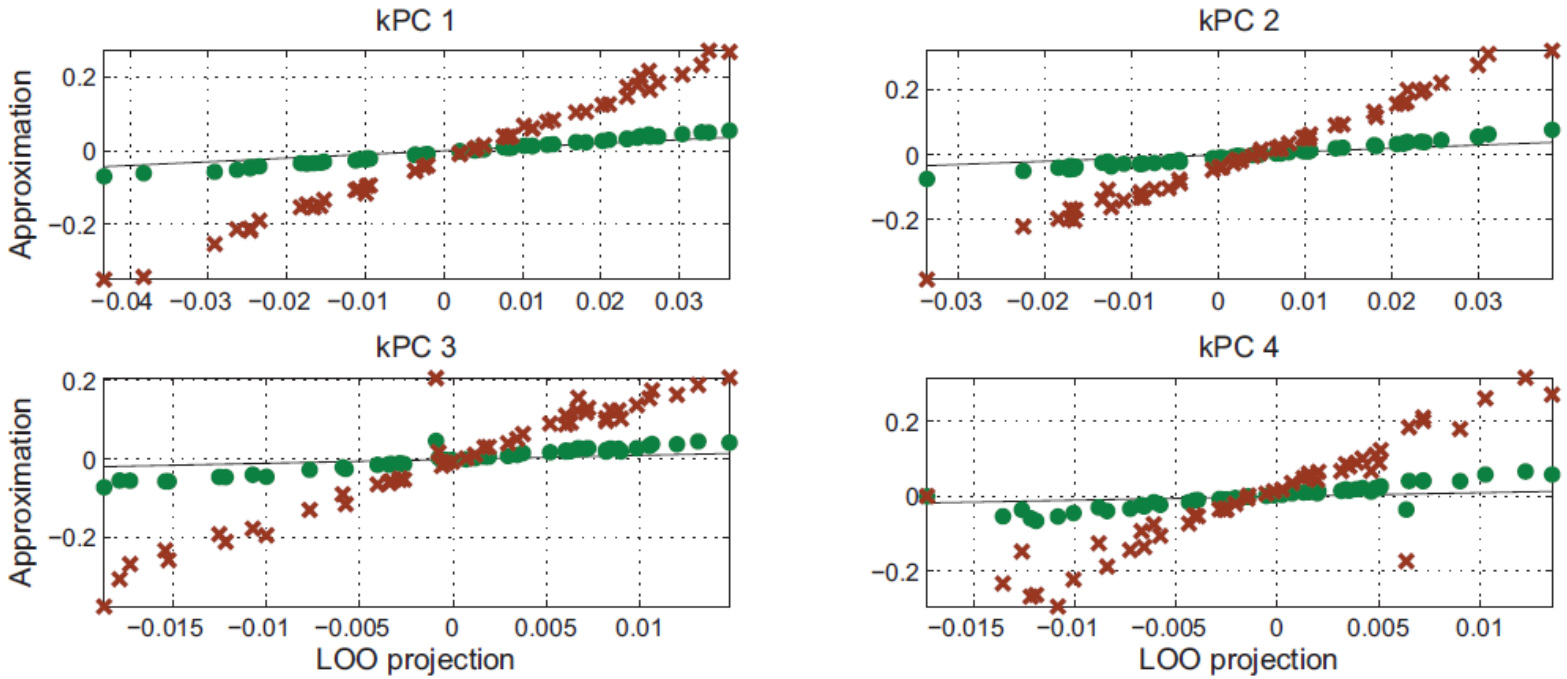
$\tilde{g}_{te}^{I(n),d} \leftarrow h^{n,d}$ {renormalized test data in the principal subspace, see Equation (4)}

end for

end for

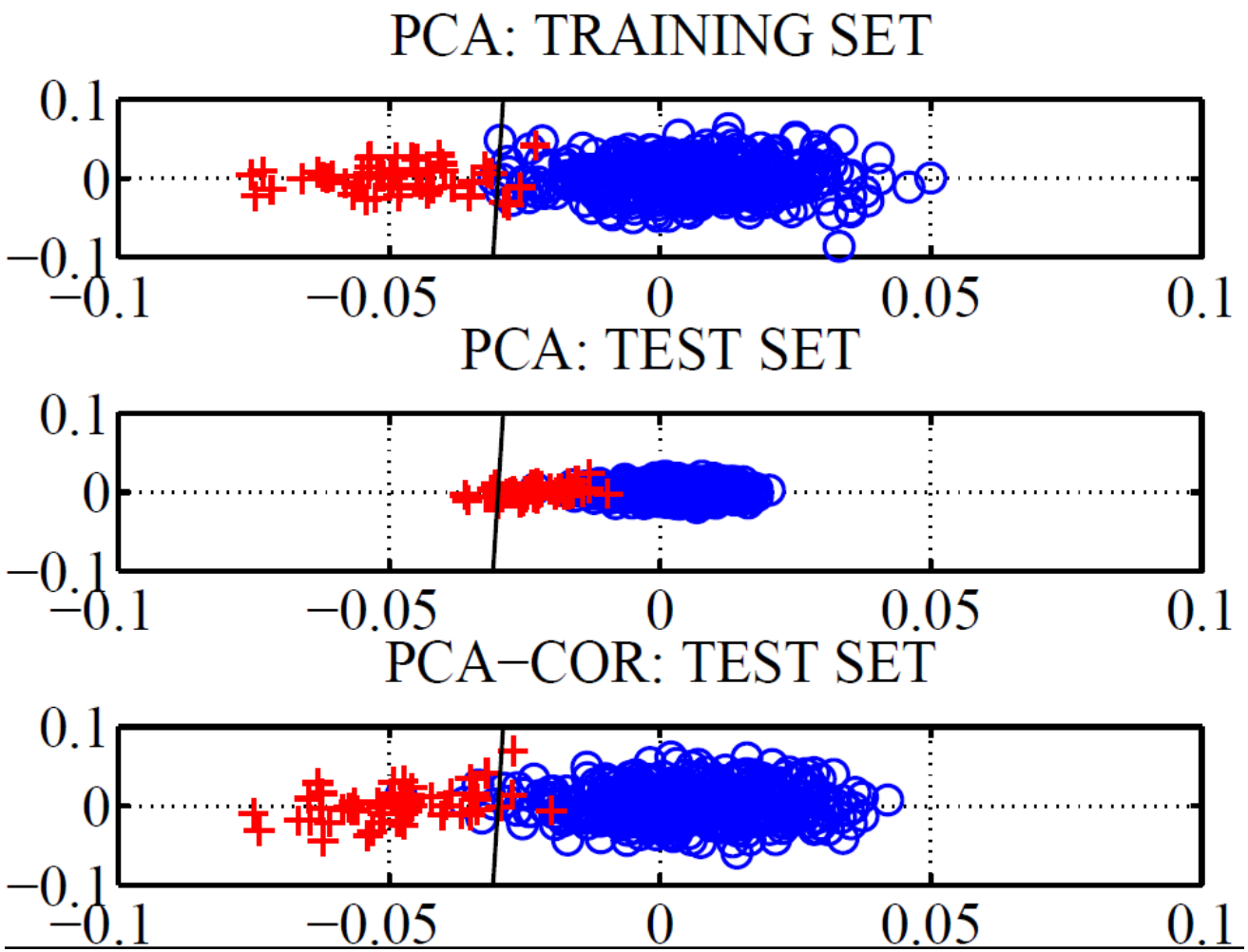
The lost projection

$$||\boldsymbol{x}_n - \boldsymbol{x}_N||^2 = ||\boldsymbol{x}_n - \boldsymbol{x}_N^{||}||^2 + ||\boldsymbol{x}_N^{\perp}||^2$$

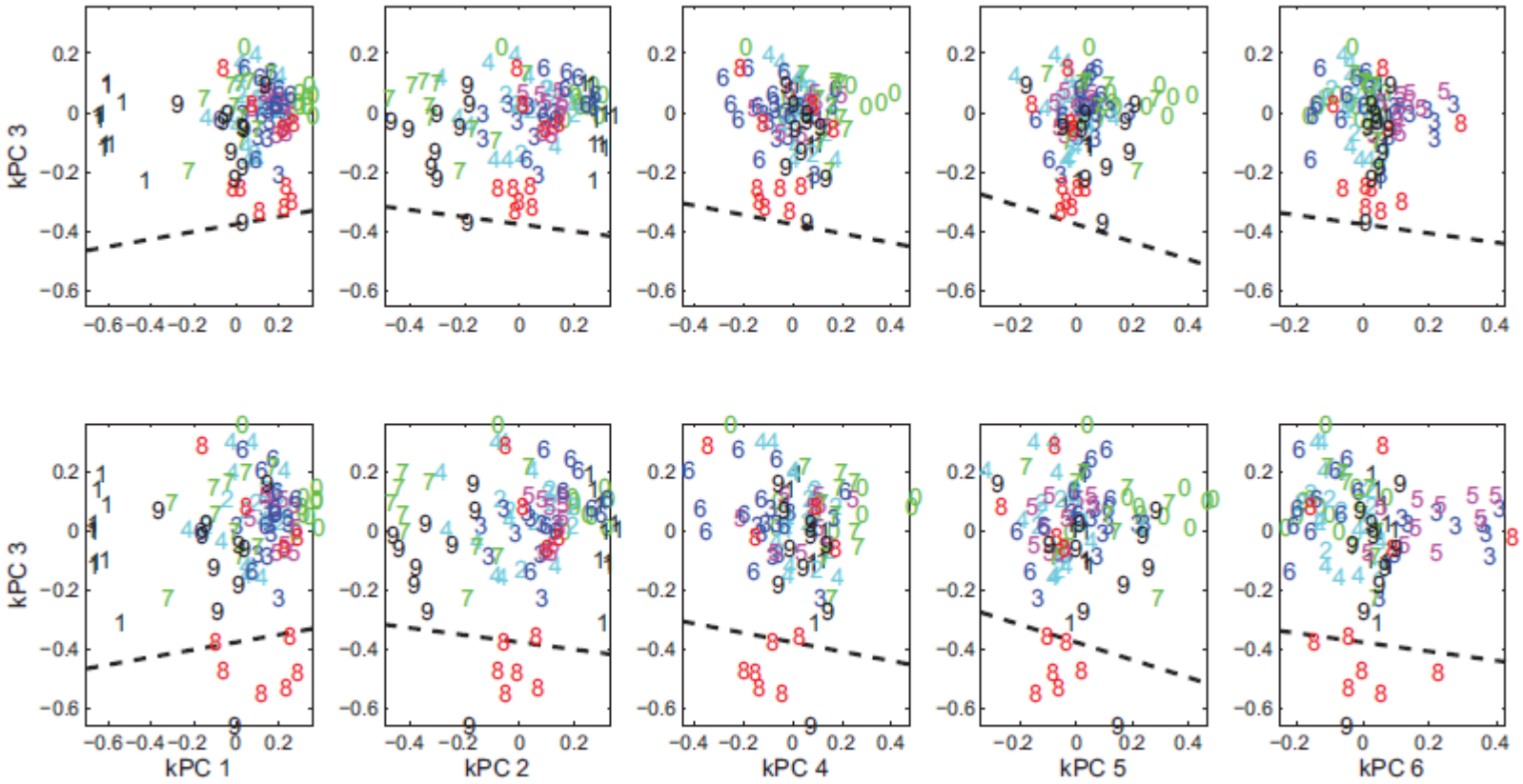


$$\beta_i(\boldsymbol{x}_N) = \sum_{n=1}^{N-1} \alpha_{in} \tilde{k}(\boldsymbol{x}_N, \boldsymbol{x}_n) = \exp\left(-\frac{1}{c}||\boldsymbol{x}_N^{\perp}||^2\right) \sum_{n=1}^{N-1} \alpha_{in} \tilde{k}(\boldsymbol{x}_N^{||}, \boldsymbol{x}_n)$$

Implication for a simple classifier

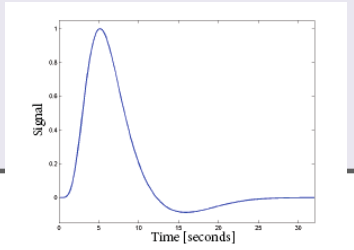


Test prior to scaling

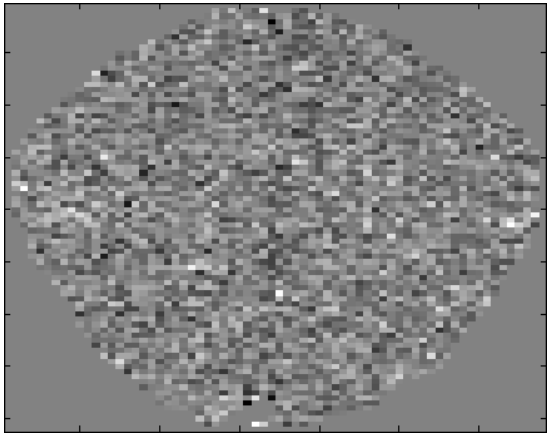


Test post scaling

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”?



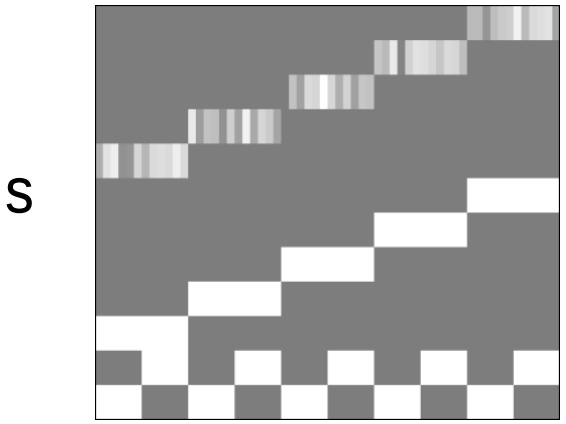
TR = 333 ms

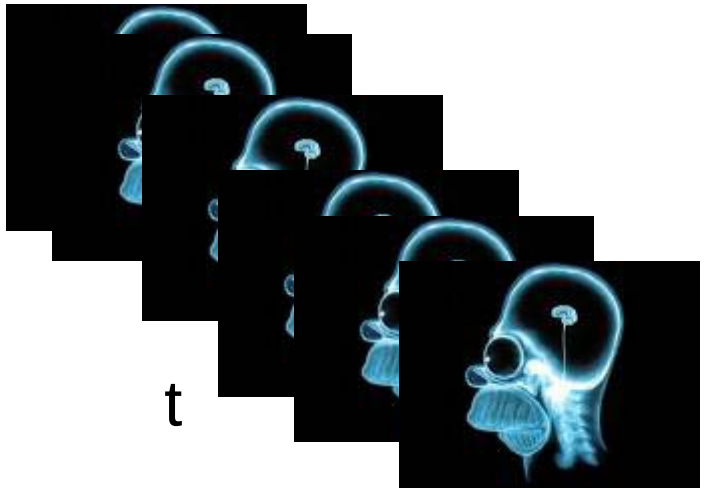
Multivariate neuroimaging models

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”…..





Application to classification of high-dimensional data on manifolds (fMRI, exceptional good SNR in raw data)

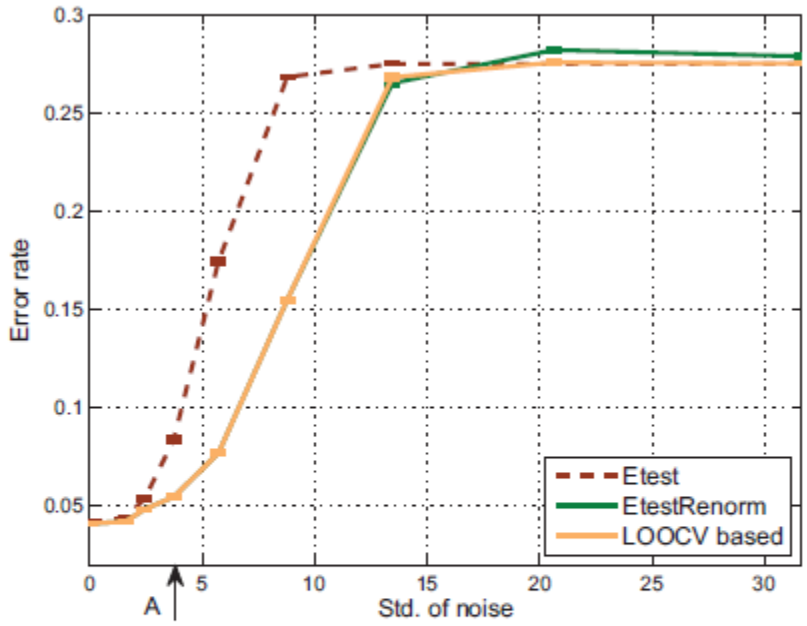


Figure 11: Mean error rates $\pm$ 1 standard deviation as a function of the noise level for fMRI data ($D = 16{,}384, N = 605$) . The test error based on conventional kernel PCA projections, renormalized projections, and a LOOCV scheme is shown. Renormalization is seen to clearly improve the performance. Arrow 'A' indicates the noise level used in Figure 12

Application to classification of high-dimensional data on manifolds (fMRI, exceptional good SNR in raw data)

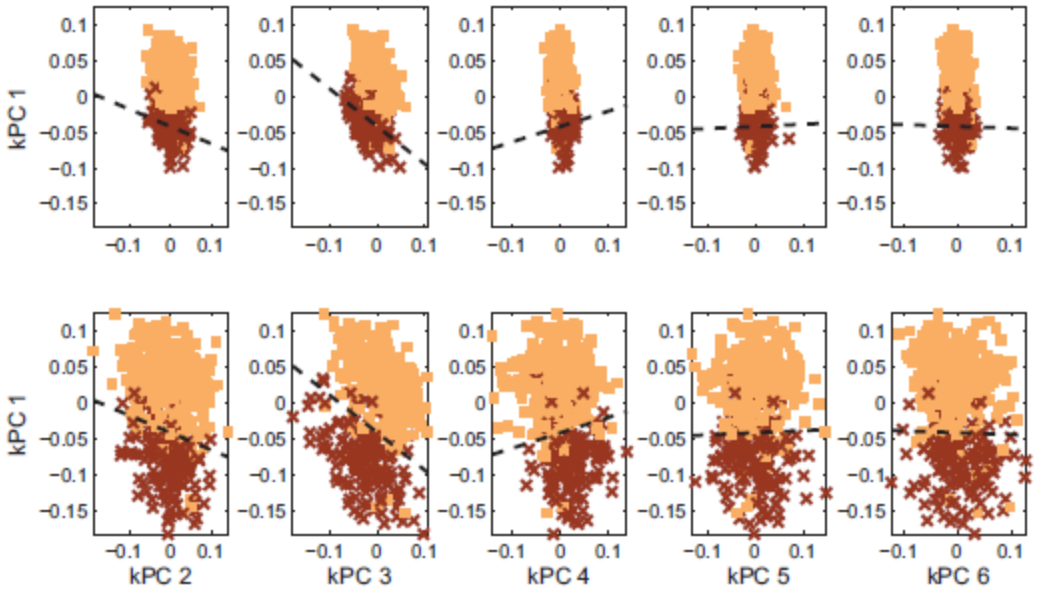


Figure 12: Test set projections of the fMRI data with Gaussian noise added as marked on Figure 11 ($\varepsilon_i = \mathcal{N}(0, 3.8^2)$). The top row shows the conventional projections, while the bottom row shows the projections after renormalization. The 'red class' indicates activation, while the blue observations are acquired during rest. The dashed line marks the linear discriminant. The scale is chosen as the 5th percentile of the mutual distances.

Implications for the SVM?

Distribution of the decision function

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0.$$

`....unlike other machine learning methods, SVMs generalization error is related not to the input dimensionality of the problem, but to the margin with which it separates the data...'

J. Kwok IEEE TNN (1999)

Decision function mis-match in the SVM

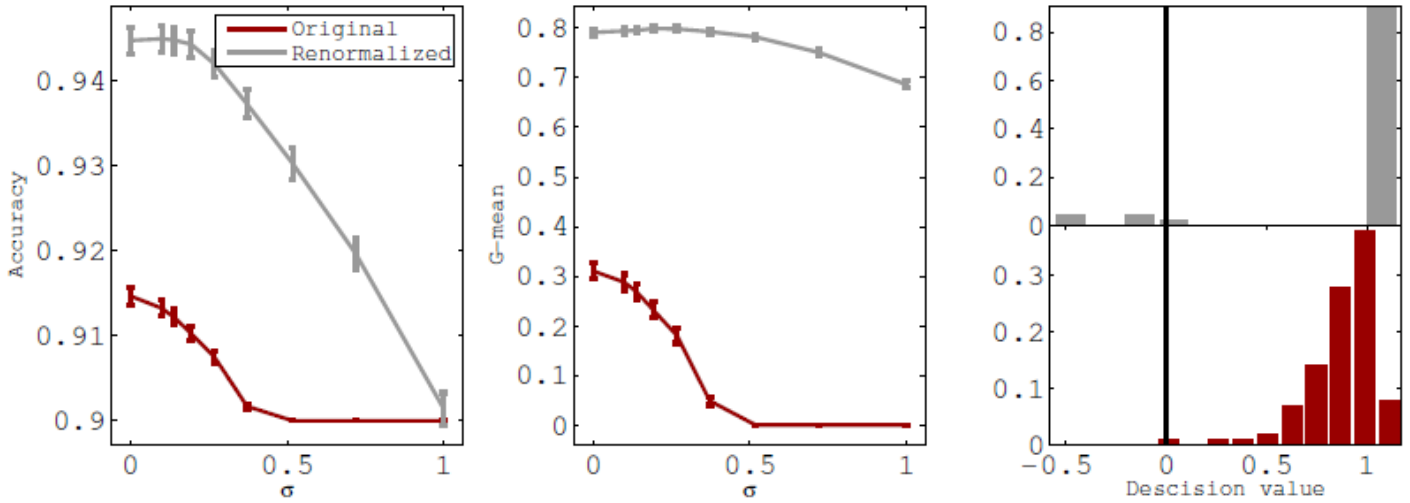


Fig. 1. Mean performance measures ±1 std as a function of the noise level for the USPS data. The left and middle panels show the accuracy and the G-mean respectively. The test accuracy is shown in red while the renormalized test accuracy is shown in gray. The right panel shows an example of the histogram before and after renormalization (for a noise level of $\sigma = 0.27$).

T.J. Abrahamsen, LK Hansen: Restoring the Generalizability of SVM based Decoding in High Dimensional Neuroimage Data NIPS Workshop: Machine Learning and Interpretation in Neuroimaging (MLINI-2011)

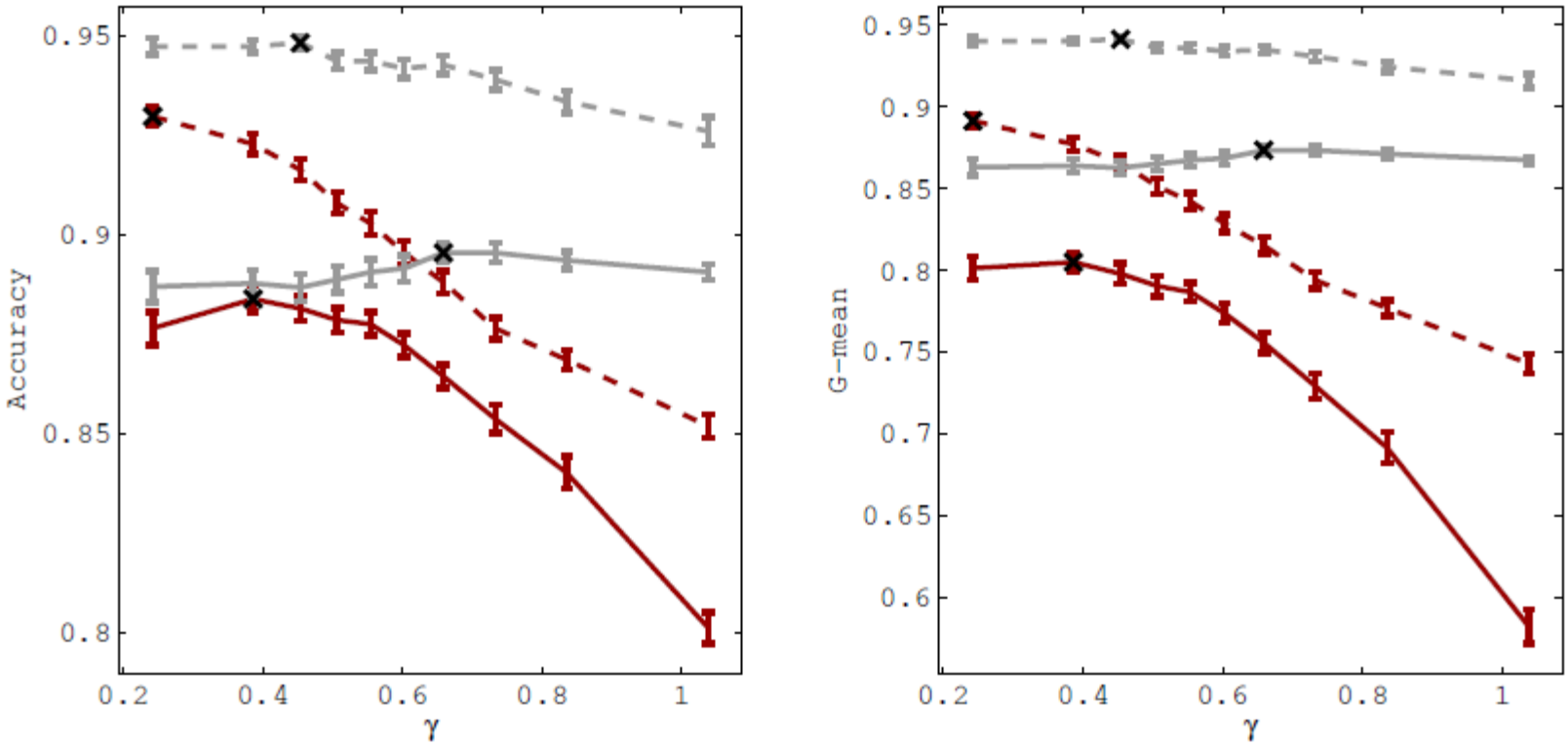
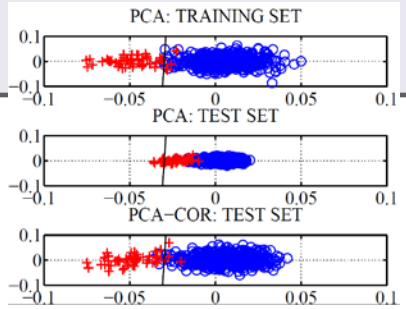


Fig. 2. Mean performance measures ± 1 std as a function of kernel hyperparameter for the fMRI data. Higher values of γ lead to more non-linear kernel embeddings. The left and right panel shows the accuracy and the G-mean respectively. The dashed lines correspond to the scheme where data with no stimuli are omitted, while the full lines show the performance on the subsampled data. The test accuracy is shown in red while the renormalized test accuracy is shown in gray. The black crosses indicate the optimal kernel hyperparameter. Renormalization is seen to improve performance and notably it leads to more non-linear optimal kernels as the optimal scale parameters chosen by cross-validation are increased.

Conclusion



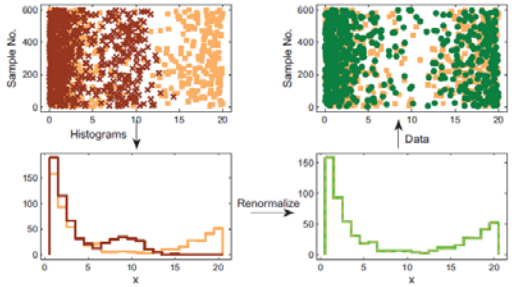
- Variance inflation in PCA

  Cure: Rescale std's

- Variance inflation in kPCA

  Cure: Non-parametric

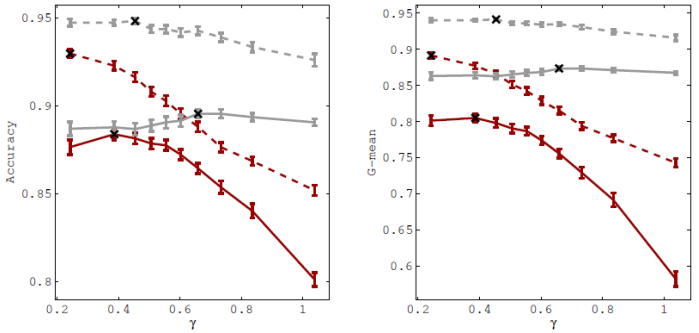
  renormalization of components



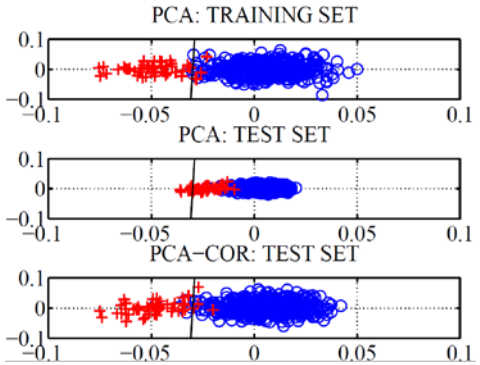
- Support Vector Machines:

  In–line renormalization seems to enable

  more non-linear classifiers in D>>N



Conclusion: Small samples in high-dimensions
Variance inflation & hubs



- Variance inflation in PCA

  Can be cured by rescaling variances

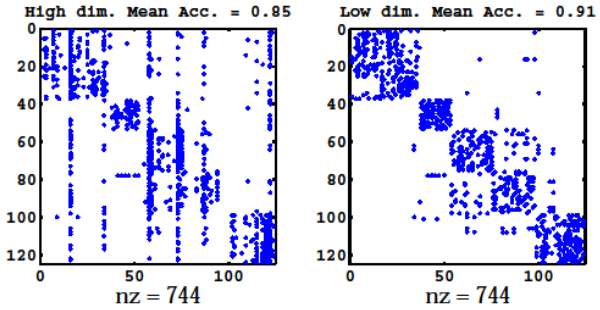
- Variance inflation in kPCA

  Can be cured by renormalization of components

- Hubs in the neighbor graph

  Hubs is a challenge for classifiers in HD,

  Research issue: development of diagnostics?



Acknowledgments

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