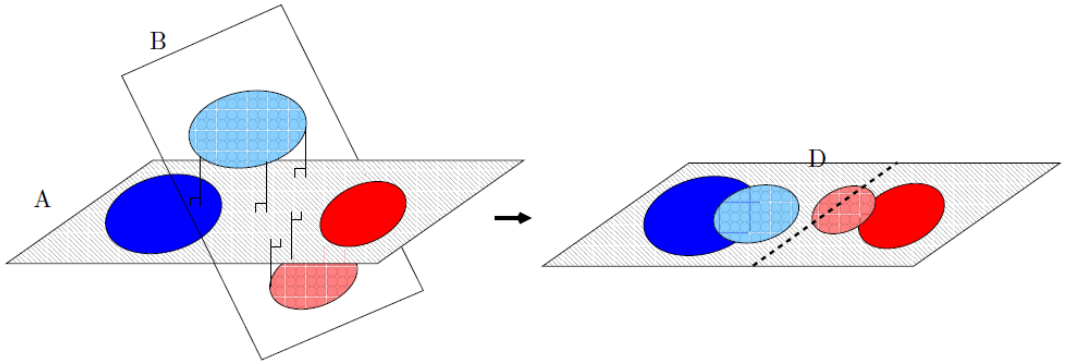
Kernels in Copenhagen

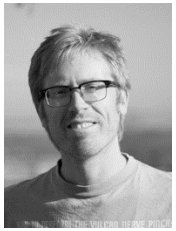
Variance inflation, explainability & spontaneous symmetry breaking

Lars Kai Hansen
DTU Compute, Technical University of Denmark



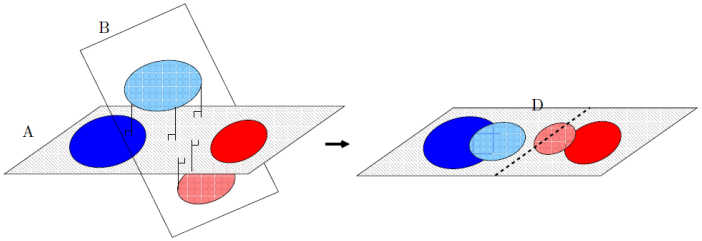


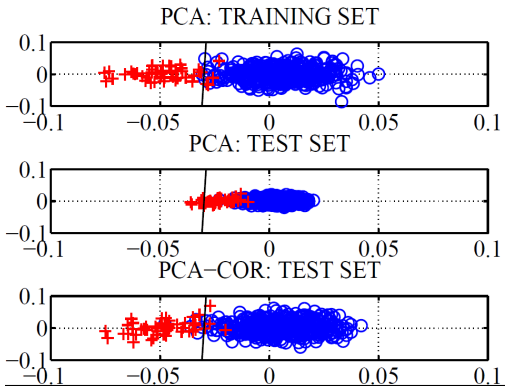




Co-workers:
Trine Abrahamsen, Ulrik Kjems, Stephen Strother, Cilie Feldager Hansen, Søren Hauberg,

OUTLINE



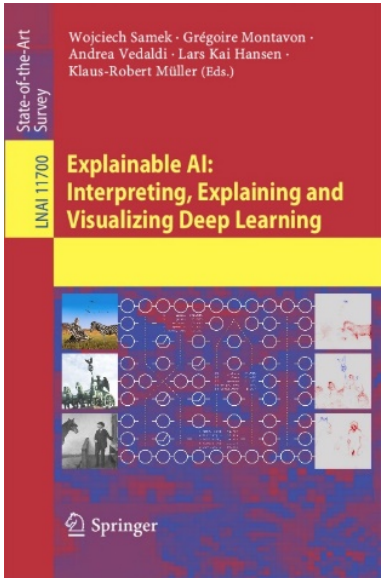


Variance inflation in

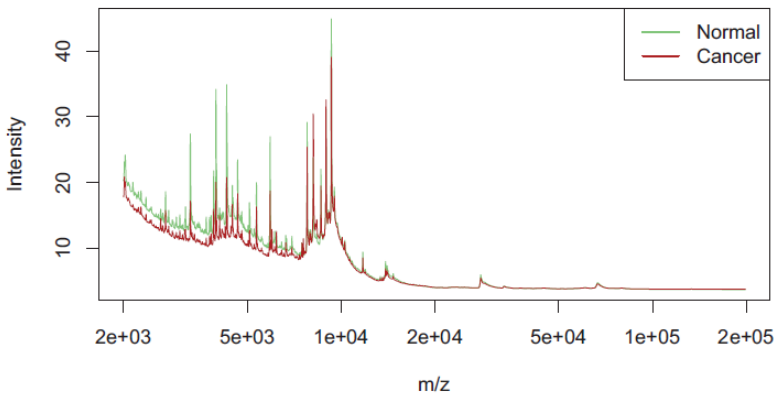
PCA, kPCA, linear regression and SVMs

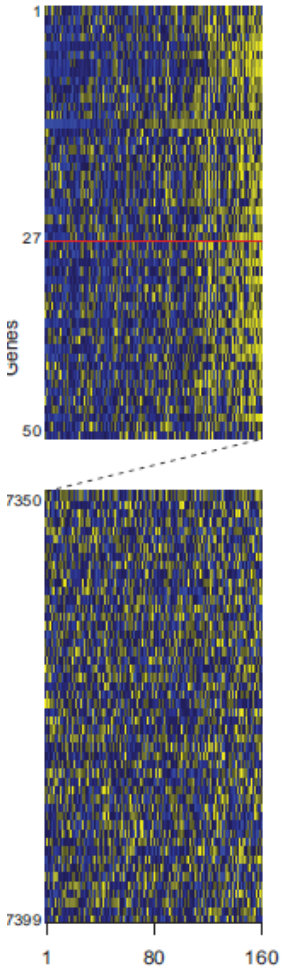
Explainability, uncertainty quantification

Spontaneous symmetry breaking in kernel reps



High dimensions – small samples (D>>N)

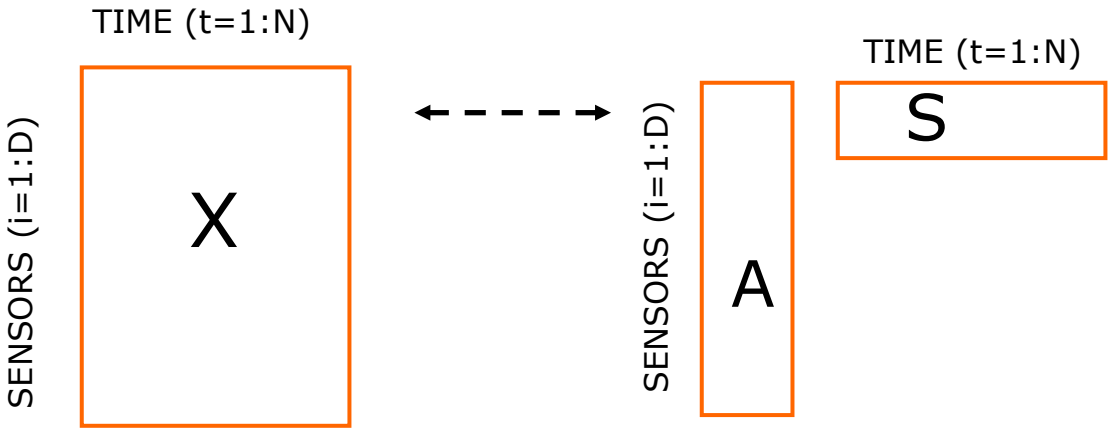




"HDLSS" high dimension, low sample size (Hall 2005, Ahn et al, 2007)
"Large p, small n" (West, 2003), "Curse of dimensionality" (Occam, 1350)
"Large underdetermined systems" (Donoho, 2001)
"Ill-posed data sets" (Kjems, Strother, LKH, 2001)

Representation learning - factor models

Represent a datamatrix by a low-dimensional approximation,
eg. linear / subspace representation



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

Unsupervised learning:
Factor analysis generative model

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

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

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

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 Computation (2002),
Neurocomputing (2002)

Matrix factorization: SVD/PCA, NMF, Clustering

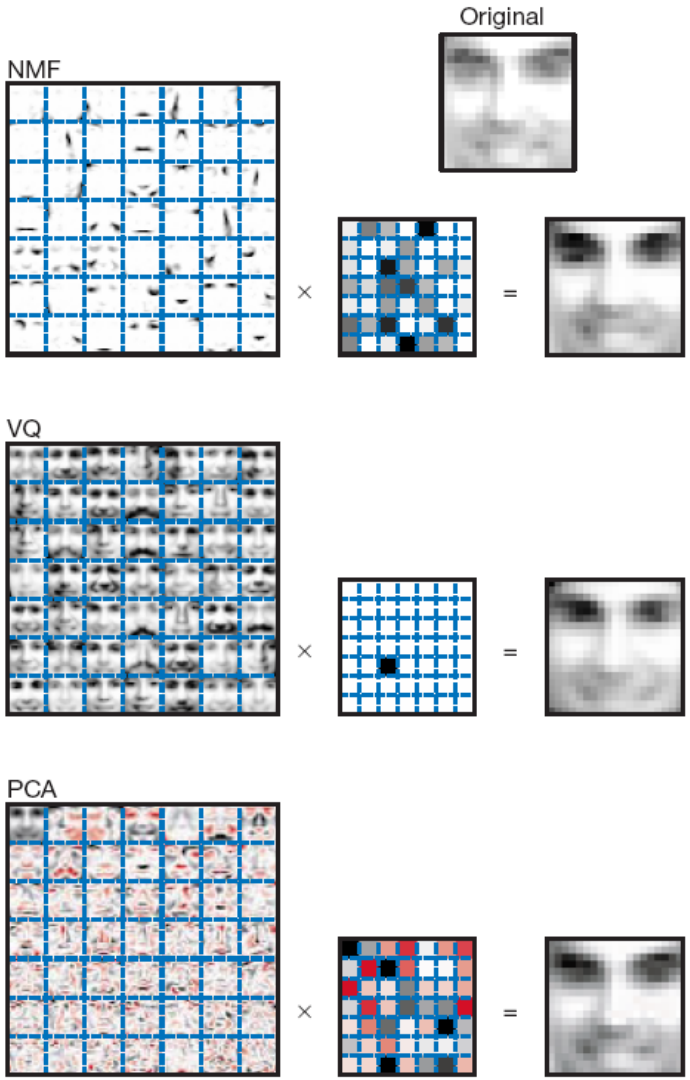


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting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

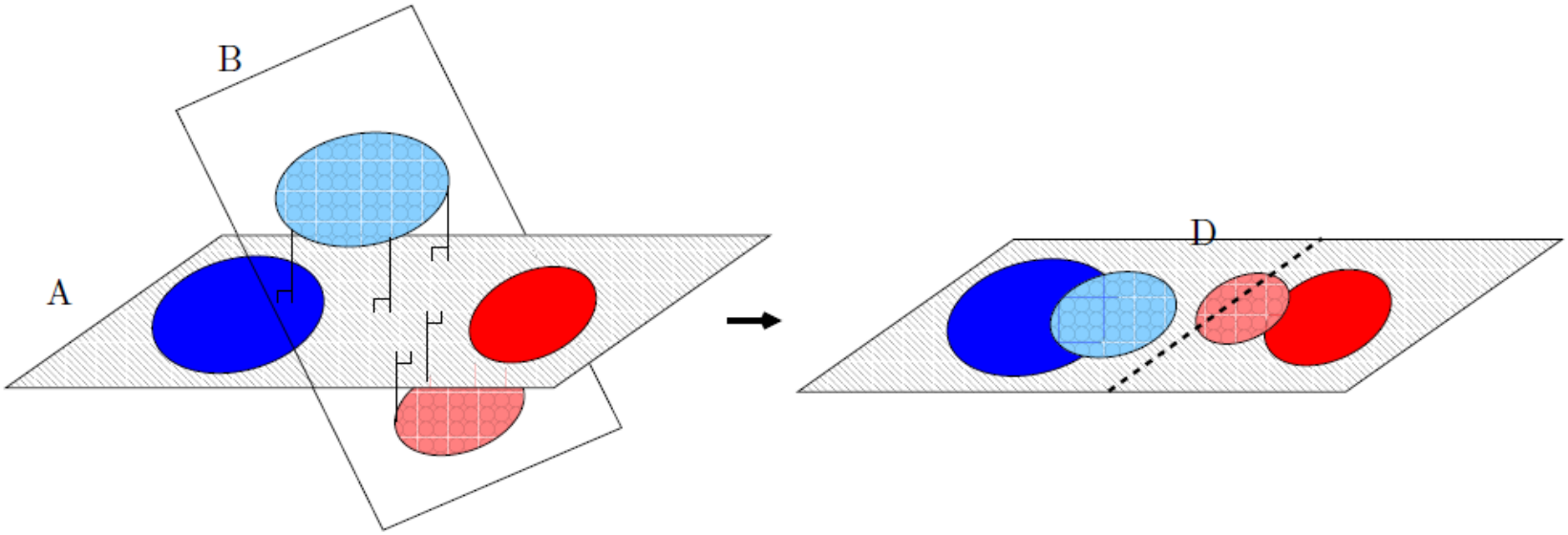
Variance inflation in PCA

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

Variance inflation in PCA



Who shrunk the test set?

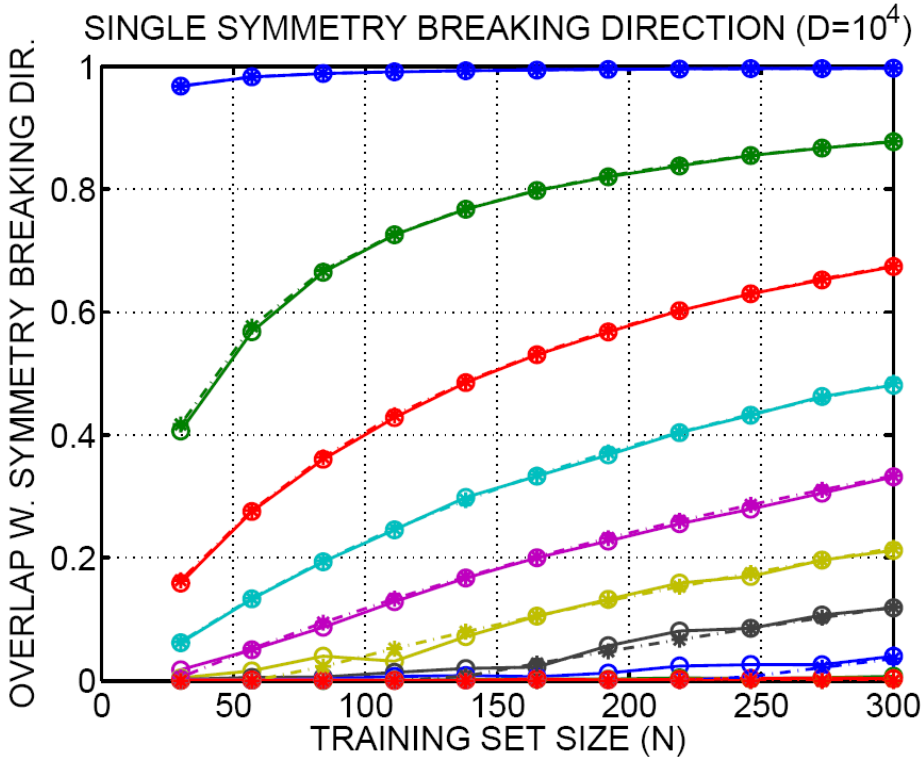
Modeling the generalizability of SVD

- 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)

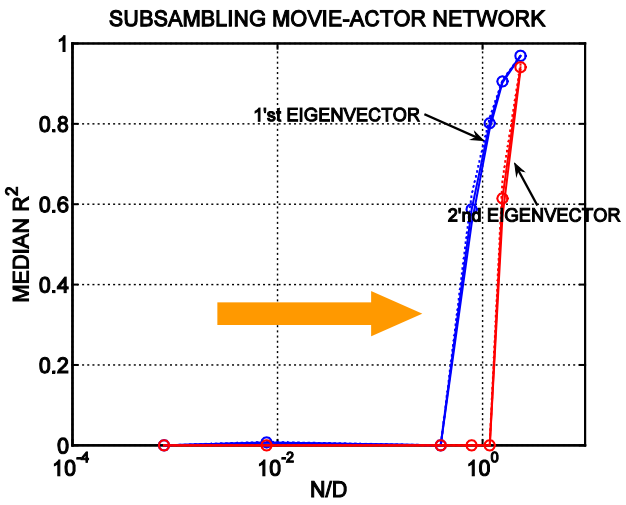


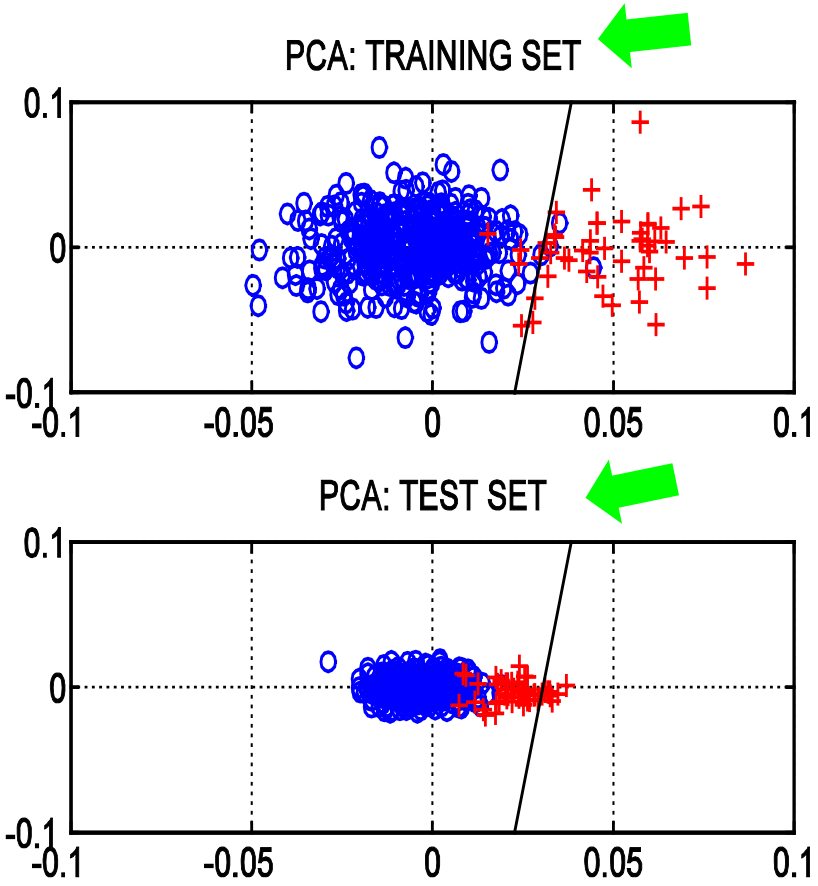
N_c = (0.0001, 0.2, 2, 9, 27, 64, 128, 234, 400, 625)

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

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 ?

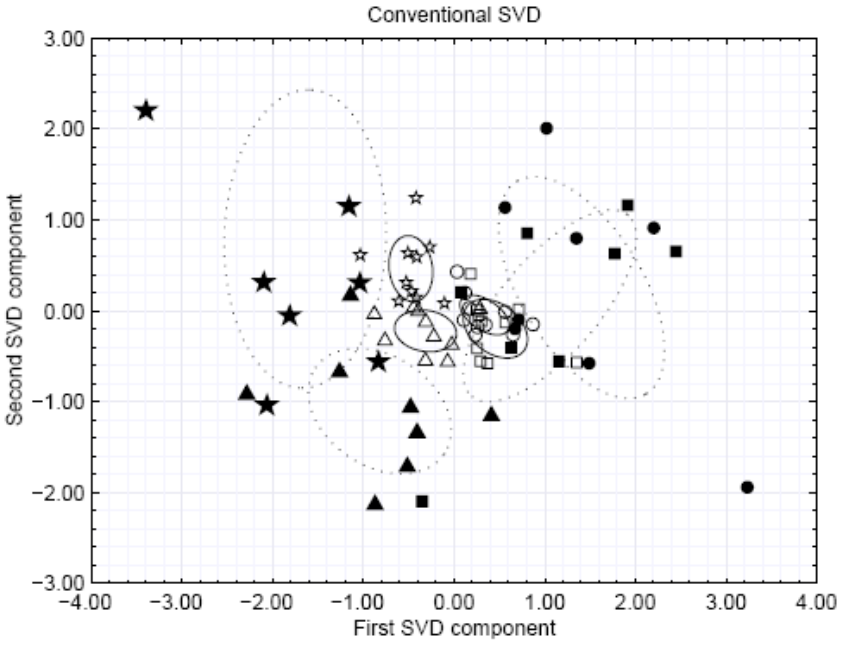


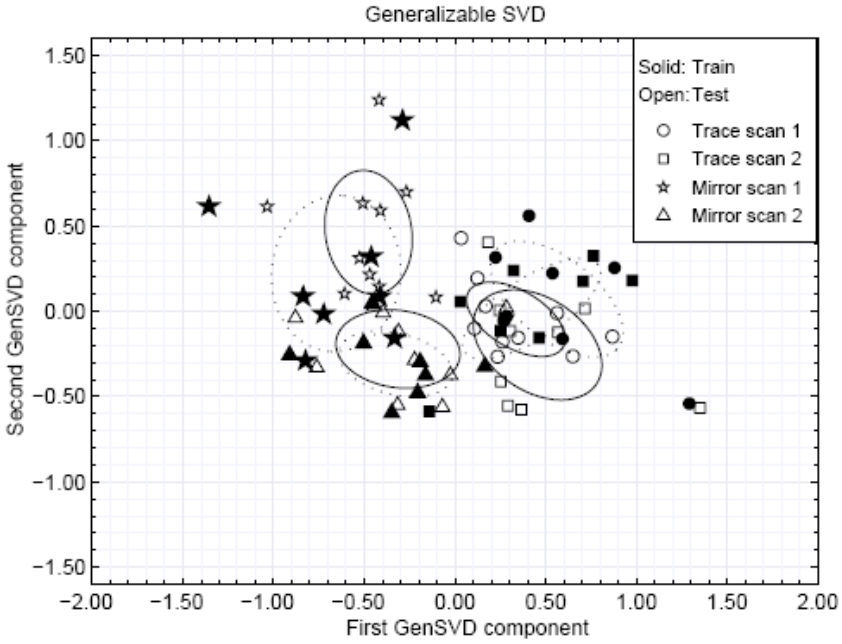


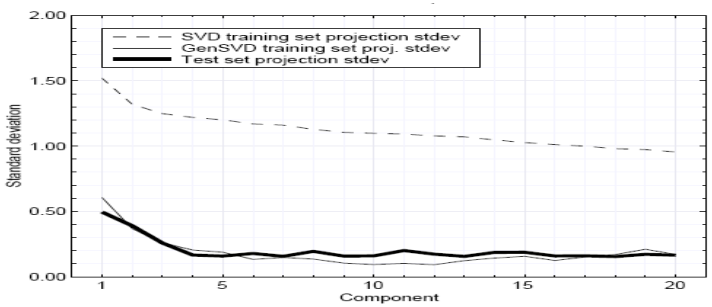
The estimated projection is offset, hence, future projections will be too small!

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

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





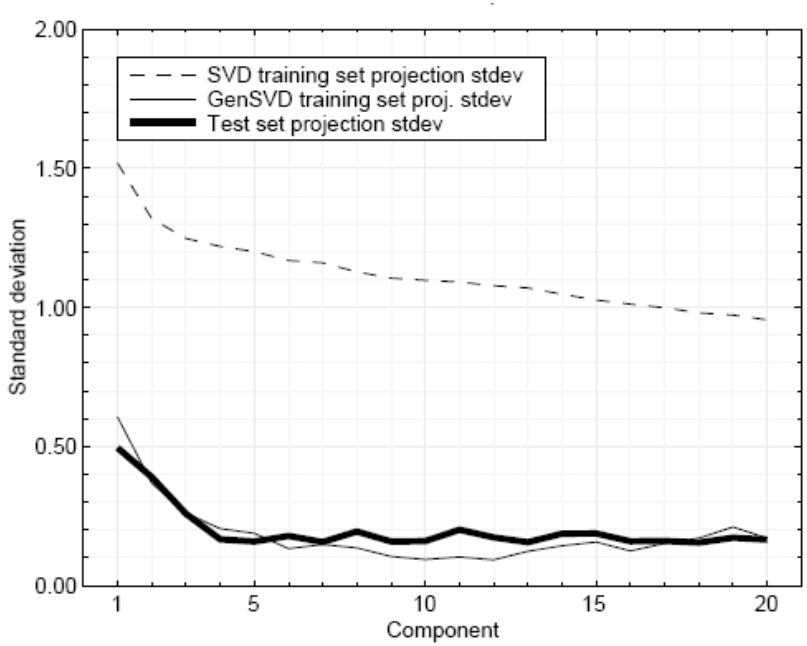


N=72, D=2.5 10^4

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

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 (…however scales like N^4)



Kjems, Hansen, Strother: NIPS (2001)

Approximating LOO (leave-one-out in N^3)

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

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

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

Projection on N-1 samples scales like N^2

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

Head-to-head comparison of two approximation scheme

Adjusting for the mean overlap using phase transition theory

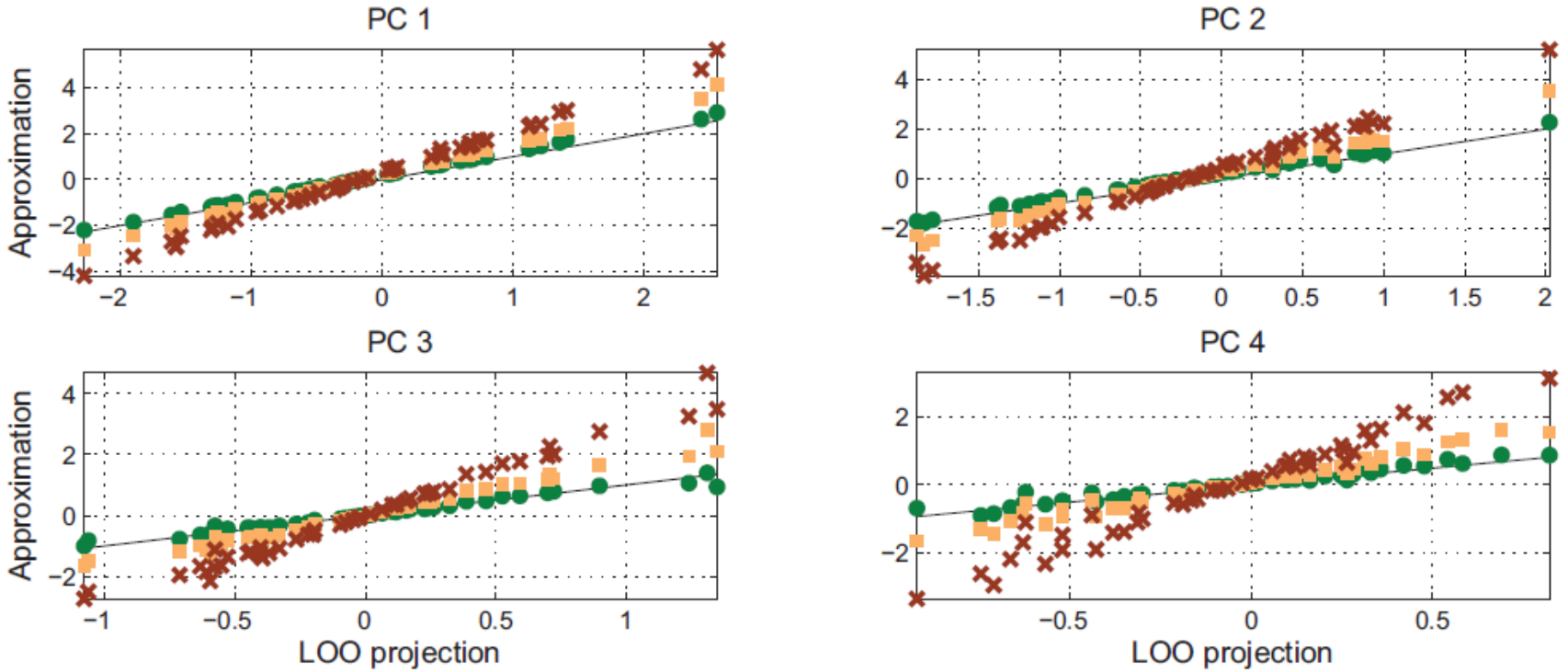
$$R^2 = \begin{cases} (\alpha S^2 - 1)/S(1+\alpha S) & \alpha > 1/S^2 \\ 0 & \alpha \le 1/S^2 \end{cases}$$

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

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

Adjusting for lost projection

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

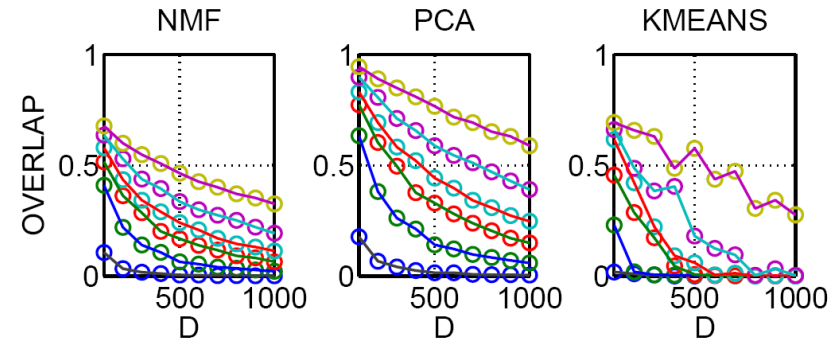


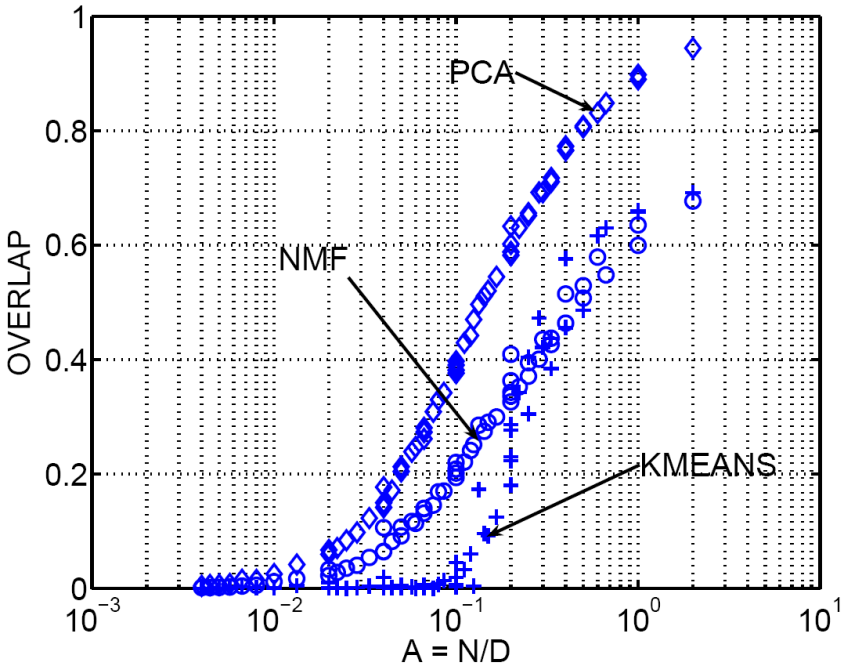
Approximating the leave-one-out (LOO) procedure. Here we simulate data with four normal independent signal components, $x = \Sigma_{k=1}^{4} \eta_k u_k + \epsilon$ of strengths $(1.4, 1.2, 1.0, 0.8,$ embedded in i.i.d. normal noise $\epsilon \sim N(0, \sigma^2 1)$, with $\sigma = 0.2$. The dimension 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).

Specific to PCA? No…universality also in NMF, Kmeans

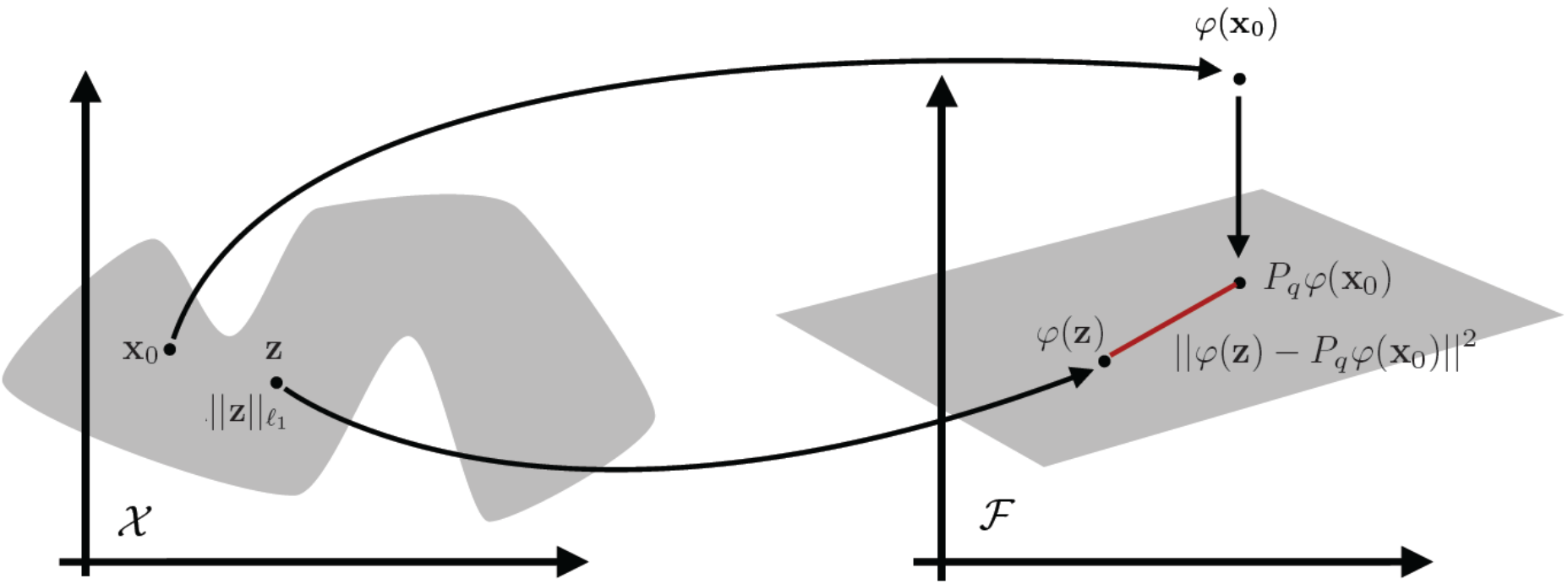
- Looking for universality by simulation
  - learning two clusters in white noise.
- Train K=2 component factor models.
- Measure overlap between line of sigth and plane spanned by the two factors.

Experiment
Variable: N, D
Fixed: SNR





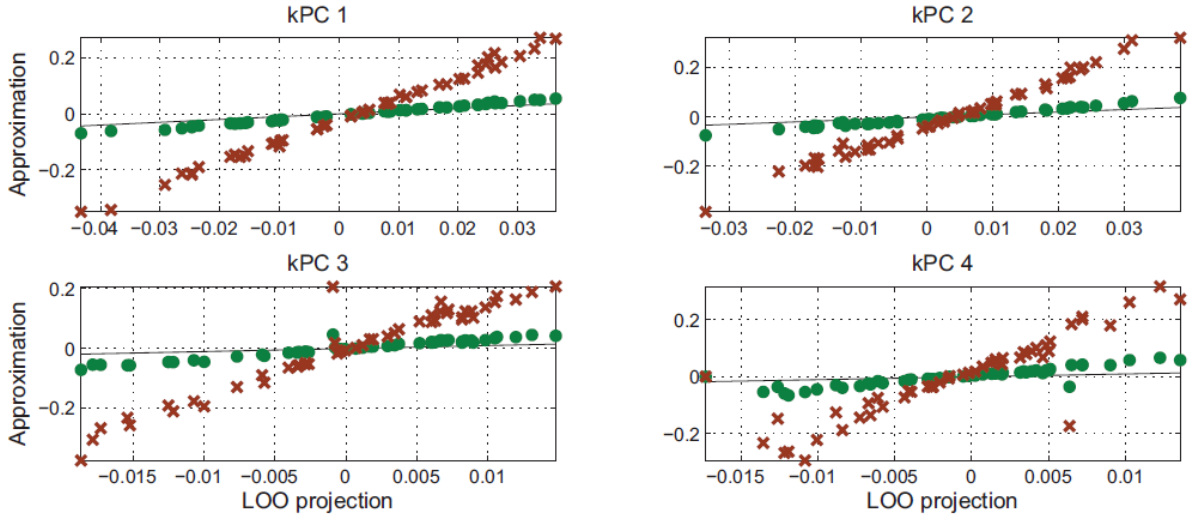
Beyond the linear model: Non-linear denoising and manifold representations



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

$$K_{n,n'} = \varphi_n^T \varphi_{n'} = \exp\left(-\frac{||\mathbf{x}_n - \mathbf{x}_{n'}||^2}{c}\right)$$

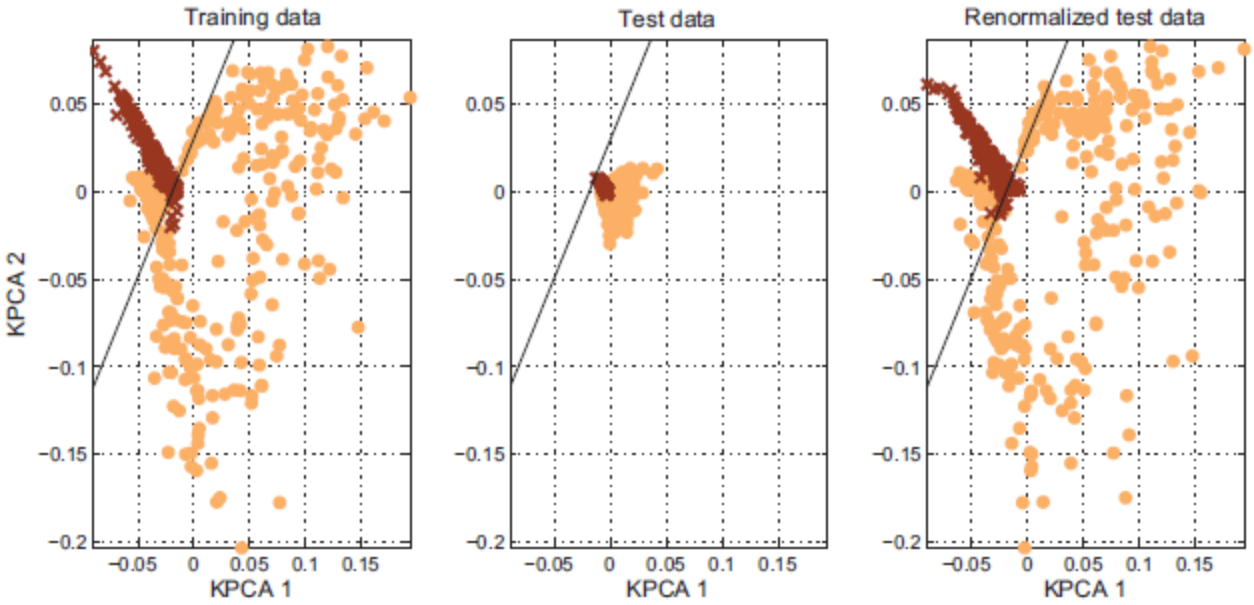
$$||\boldsymbol{x}_n - \boldsymbol{x}_N||^2 = ||\boldsymbol{x}_n - \boldsymbol{x}_N^{||}||^2 + ||\boldsymbol{x}_N^{\perp}||^2$$



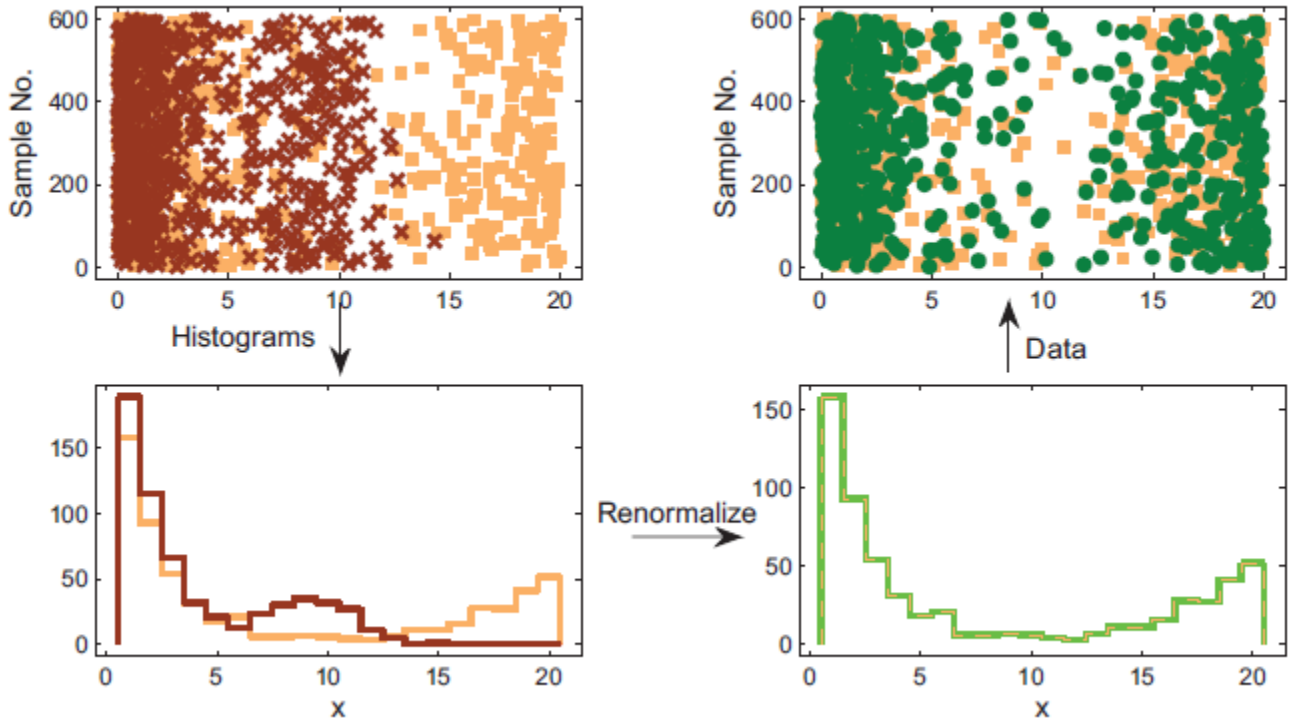
$$\beta_i(\boldsymbol{x}_N) = \sum_{n=1}^{N-1} \alpha_{in} \tilde{k}(\boldsymbol{x}_N, \boldsymbol{x}_n) = \exp\left(-\frac{1}{c}||\boldsymbol{x}_N^{\perp}||^2\right) \sum_{n=1}^{N-1} \alpha_{in} \tilde{k}(\boldsymbol{x}_N^{||}, \boldsymbol{x}_n)$$

Exponential amplification

Application to classification of high-dimensional data on manifolds

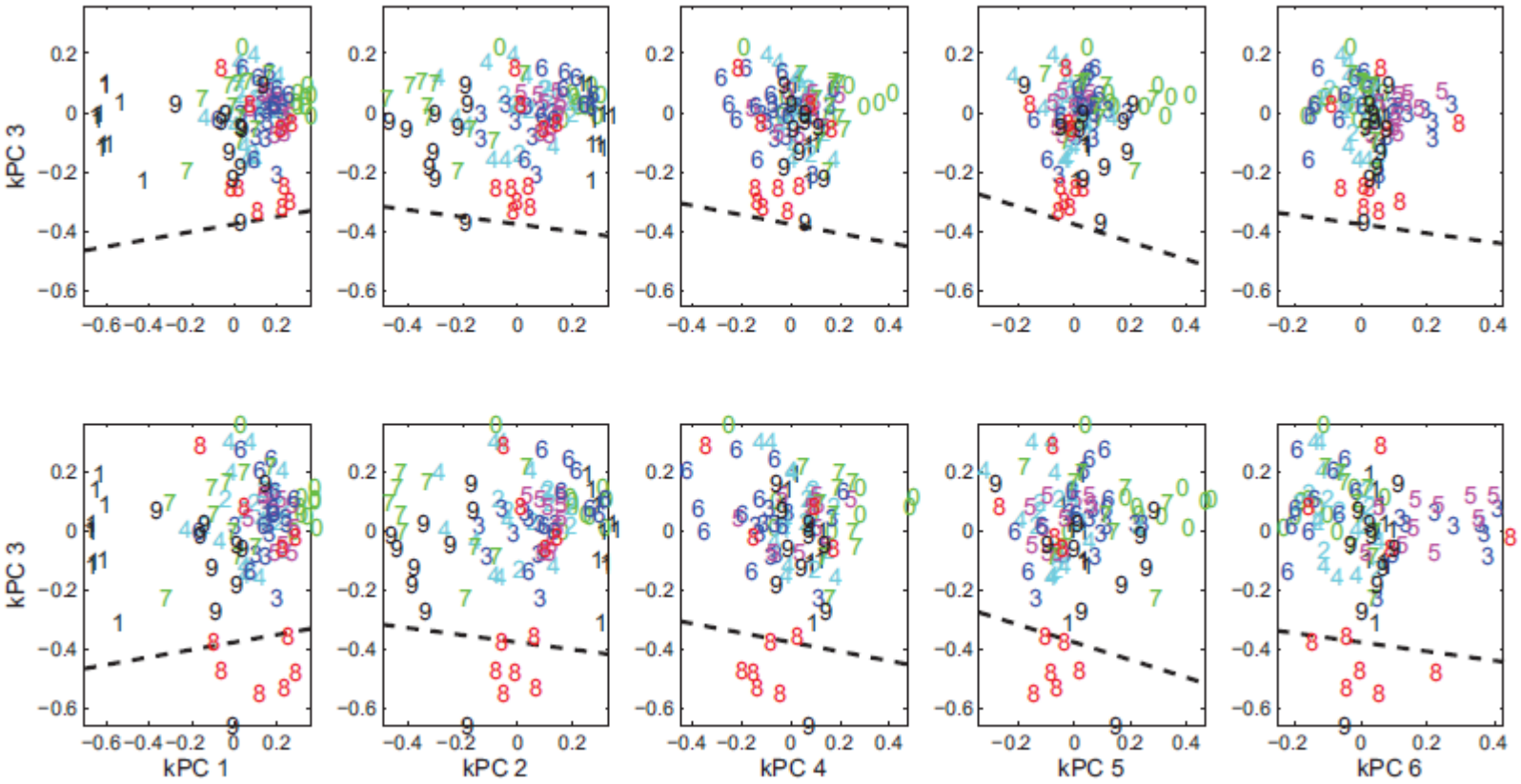


The "cure": Non-parametric histogram equalization




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

Test prior to scaling (learning ”8 vs rest”)



Test post scaling

Variance inflation in linear regression

$$y = \mathbf{w}^\top \mathbf{x} + \epsilon = \sum_{d=1}^{D} w_d x_d + \epsilon$$

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)^2$$

$$G(N) = E_{y,\mathbf{x}} \left\{ E_N \left\{ \left(y - \mathbf{w}_N^\top \mathbf{x}\right)^2 \right\} \right\}$$

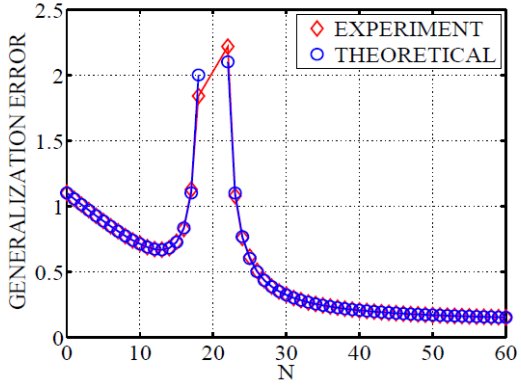


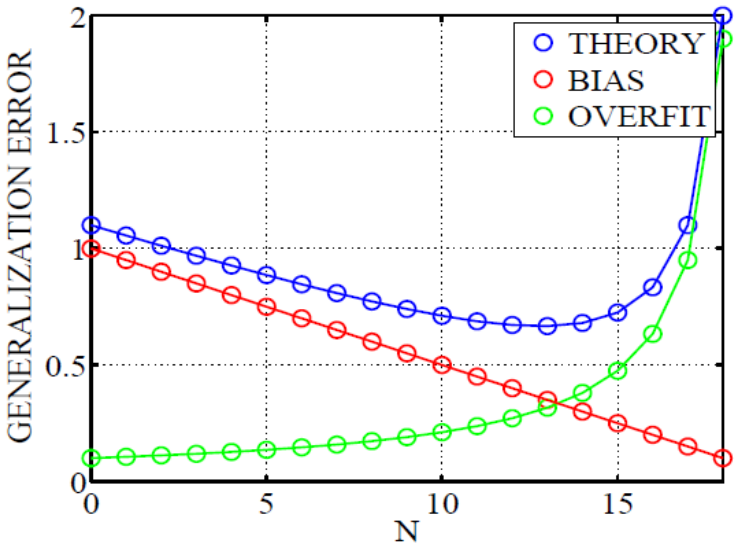
Fig. 1. Experimental and theoretical learning curves for the case $D = 20$ with $\sigma^2 = 0.1, ||\mathbf{w}_0||^2 = 1$. The theoretical result for $N > D+1$ is given in Hansen (1993). The sample size for the minimal error (for $N < D-1$) is located at $N_{\min} = [D - 1 - \sqrt{D(D-1)}\sqrt{\frac{\sigma^2}{||\mathbf{w}_0||^2}}] = 13$. The results are based on 10000 simulated data sets.

$$G(N) = \begin{cases} \left(1 - \frac{N}{D}\right) ||\mathbf{w}_0||^2 + \frac{D-1}{D-N-1}\sigma^2 & N < D-1, \\ \infty & D-1 \leq N \leq D+1 \\ \frac{N-1}{N-D-1}\sigma^2 & N > D+1. \end{cases}$$

Hansen, L. K. Stochastic linear learning: Exact test and training error averages. *Neural Networks* 6(3): 393–396 (1993)
Barber, D., D. Saad, and P. Sollich. Test error fluctuations in finite linear perceptrons. *Neural computation* 7(4): 809-821 (1995)

Variance inflation in linear regression

$$G(N) = \begin{cases} \left(1 - \frac{N}{D}\right) ||\mathbf{w}_0||^2 + \frac{D-1}{D-N-1}\sigma^2 & N < D-1, \\ \infty & D-1 \leq N \leq D+1 \\ \frac{N-1}{N-D-1}\sigma^2 & N > D+1. \end{cases}$$



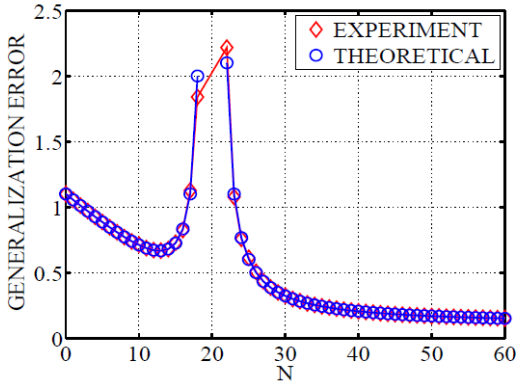


Fig. 1. Experimental and theoretical learning curves for the case $D = 20$ with $\sigma^2 = 0.1, ||\mathbf{w}_0||^2 = 1$. The theoretical result for $N > D+1$ is given in Hansen (1993). The sample size for the minimal error (for $N < D-1$) is located at $N_{\min} = [D - 1 - \sqrt{D(D-1)}\sqrt{\frac{\sigma^2}{||\mathbf{w}_0||^2}}] = 13$. The results are based on 10000 simulated data sets.

$$||\mathbf{w}_0||^2 - E_N\left\{||\widehat{\mathbf{w}}||^2\right\} = \left(1 - \frac{N}{D}\right)||\mathbf{w}_0||^2$$

Variance inflation in linear regression

$$\mathbf{w} = \sum_{n=1}^{N} \beta_n \mathbf{x}_n \qquad K_{m,n} = \mathbf{x}_m^\top \mathbf{x}_n$$

$$\widehat{\mathbf{w}} = \sum_{m,n=1}^{N} \mathbf{x}_n (K^{-1})_{n,m} y_m$$

$$\sigma^2 \left(\widehat{\mathbf{w}}^\top \mathbf{x}_n \right) = 1/N \sum_{n=1}^{N} y_n^2$$

Training set variance of predictions

$$E_N \left\{ 1/N \sum_{n=1}^{N} y_n^2 \right\} = ||\mathbf{w}_0||^2 + \sigma^2$$

Test set variance of predictions

$$E_{\mathbf{x}} \left\{ E_N \left\{ \widehat{\mathbf{w}}^\top \mathbf{x} \right\}^2 \right\} = E_N \left\{ ||\widehat{\mathbf{w}}||^2 \right\} = \frac{N}{D} ||\mathbf{w}_0||^2$$

Decision function mis-match in the SVM (MNIST)

$$\text{G-mean} = \sqrt{\text{sensitivity} \cdot \text{specificity}}$$

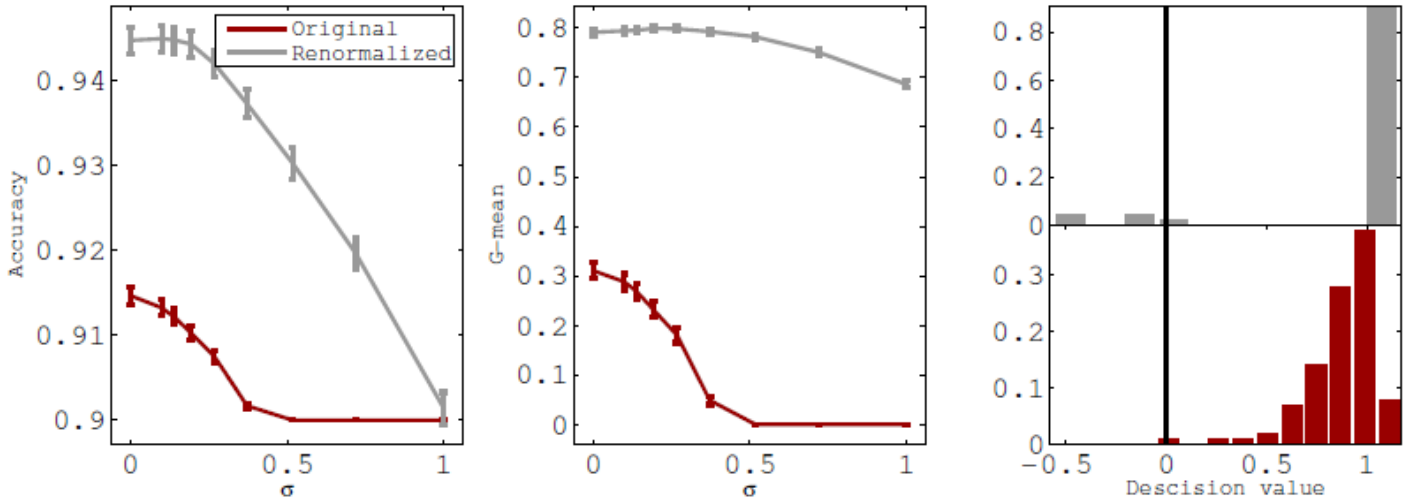
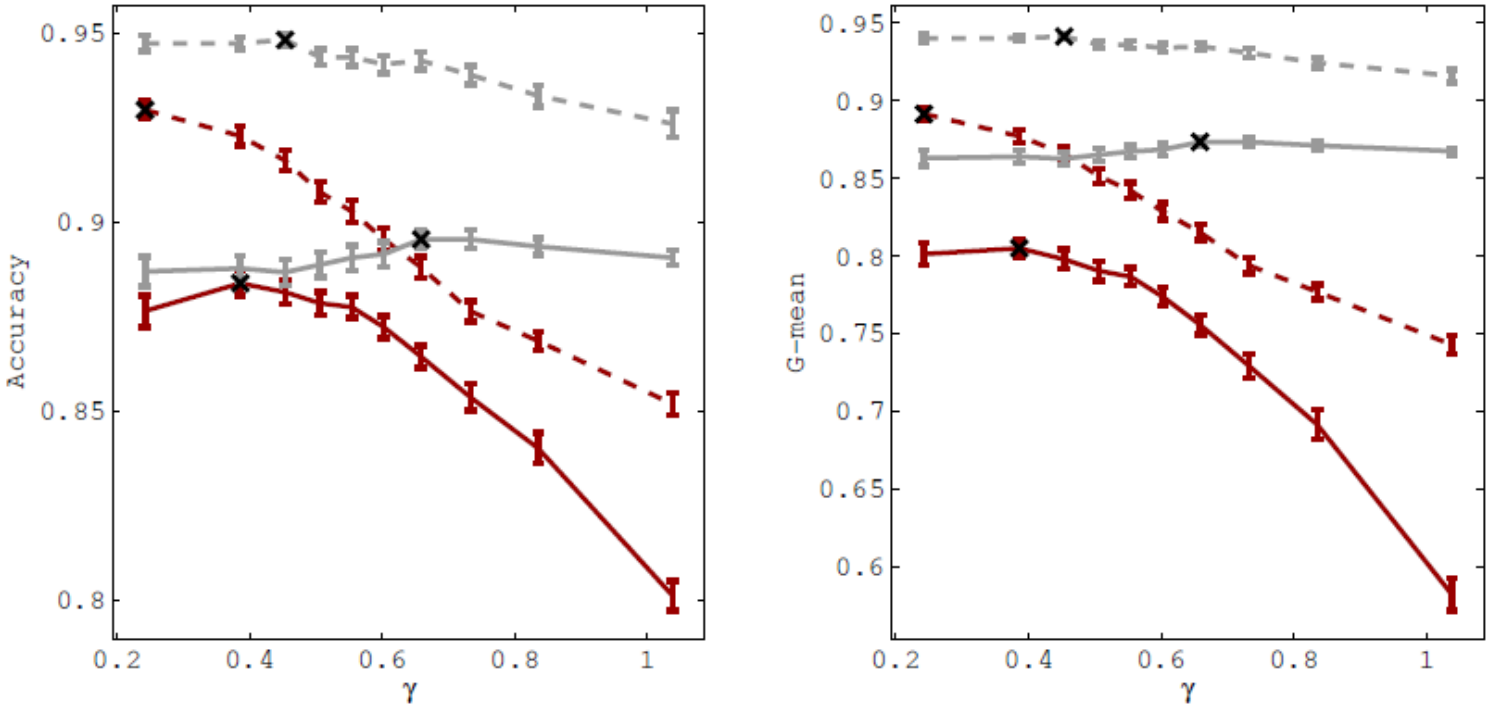


Fig. 1. Mean performance measures ± 1 std as a function of the noise level for the USPS data. The left and middle panels show the accuracy and the G-mean respectively. The test accuracy is shown in red while the renormalized test accuracy is shown in gray. The right panel shows an example of the histogram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

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

Decision function mis-match in the SVM (fMRI)

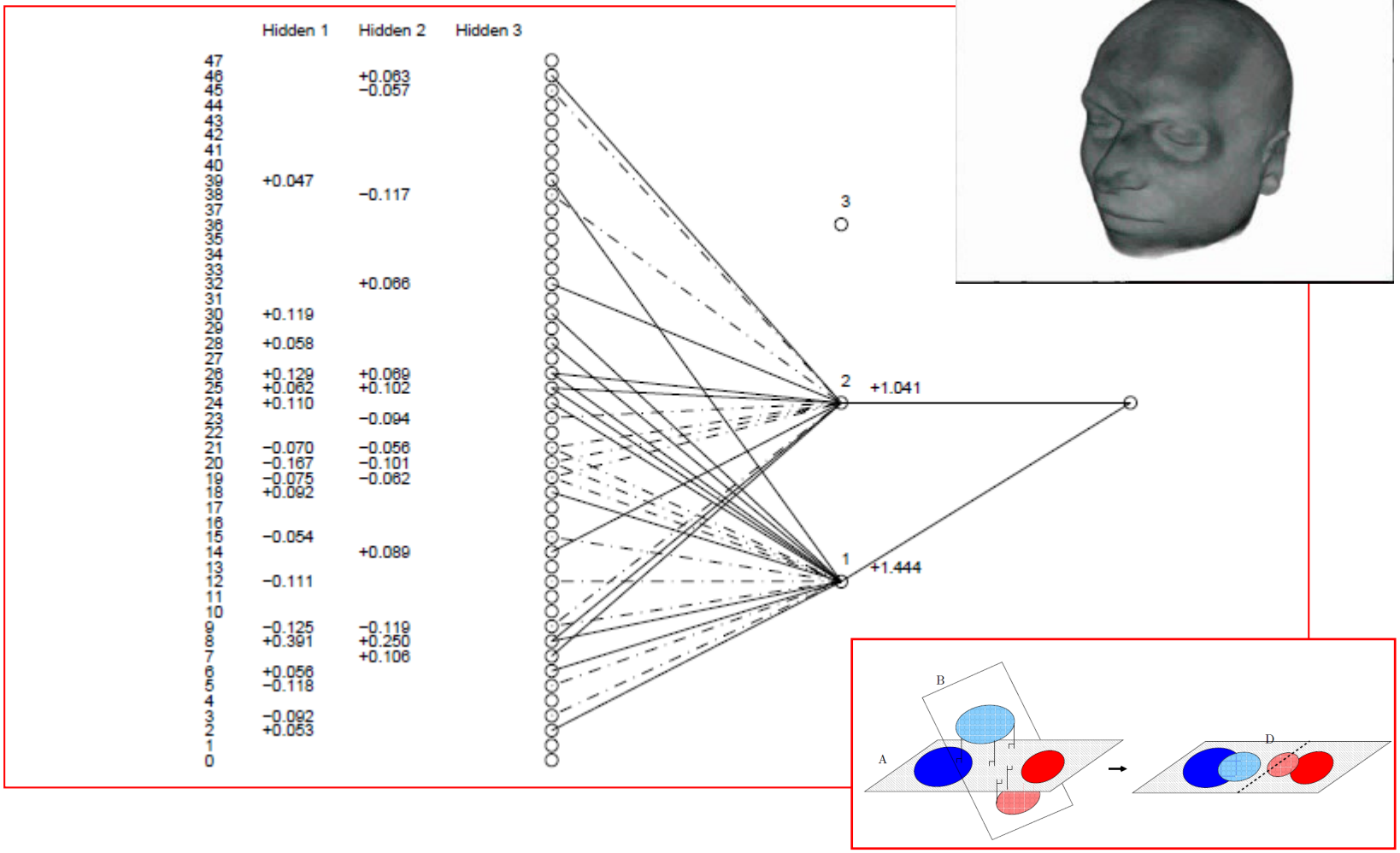


γ=1/c

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.

Explaining machine learning is possible (and has been for some time…)

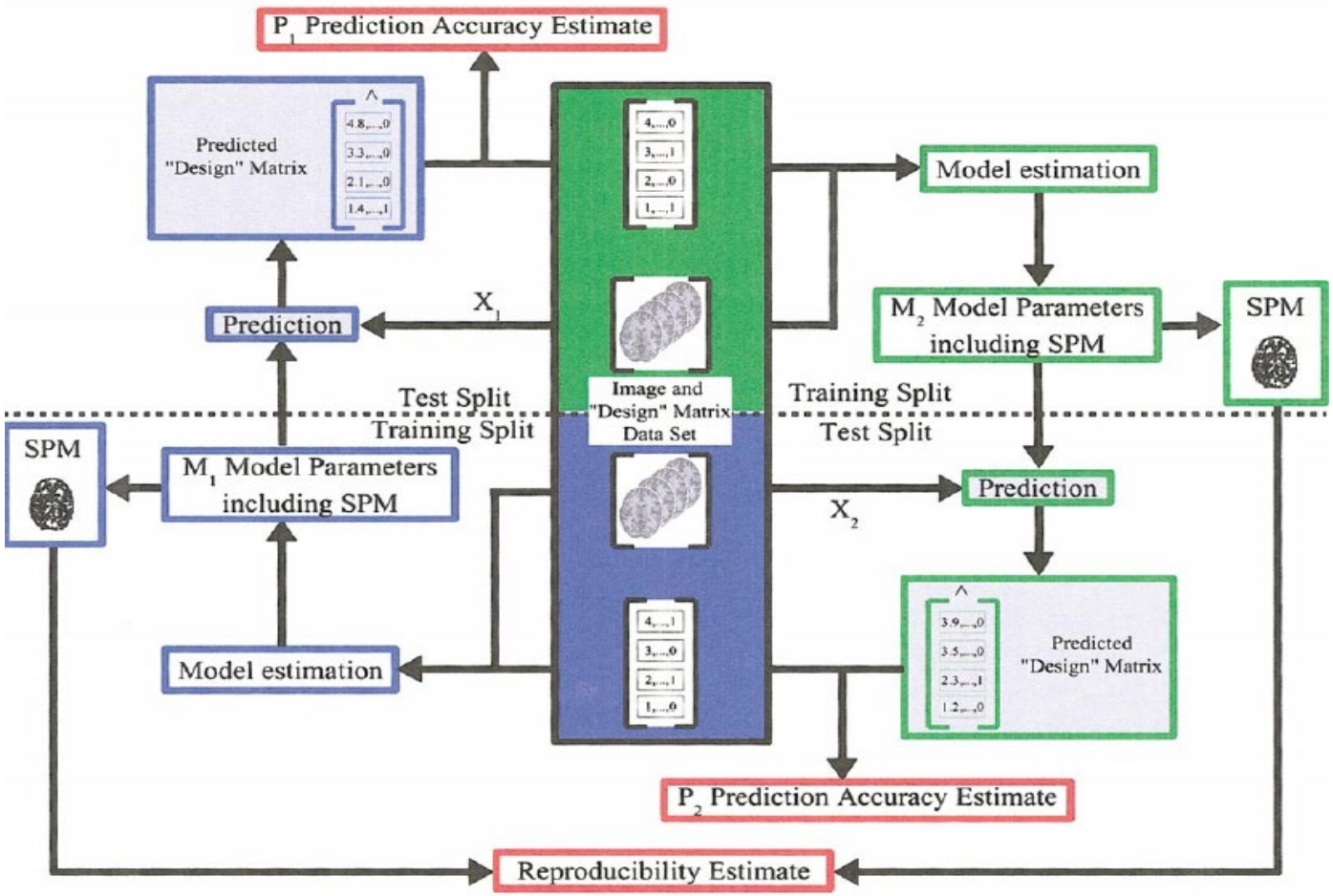
(probably) the first example… decoding PET brain scans (1994)



Lautrup, B., Hansen, L. K., Law, I., Mørch, N., Svarer, C., & Strother, S. C. (1994). Massive weight sharing: a cure for extremely ill-posed problems. In *Workshop on supercomputing in brain research: From tomography to neural networks* (pp. 137-144). "EARLY (but nor first) USE KERNEL TRICK"

Assume we have tuned ML performance – what does it do?

NPAIRS: Understanding ML performance & latent v'ble uncertainty



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), Andersen et al (2014)
Brain and Language: Hansen (2007)

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

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

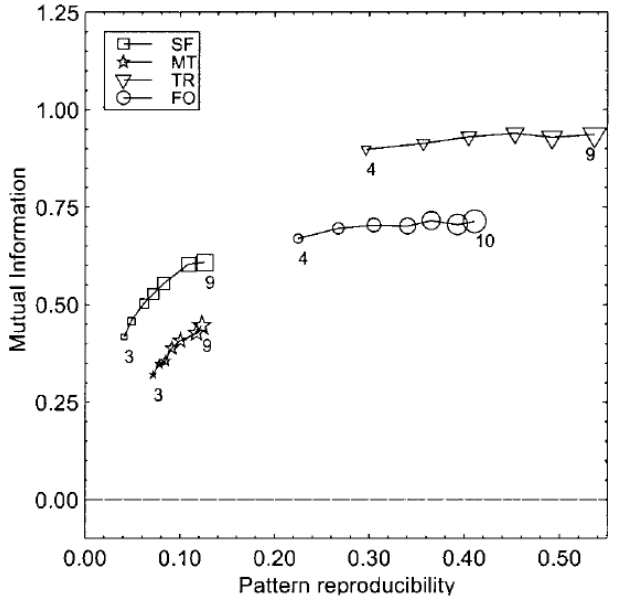
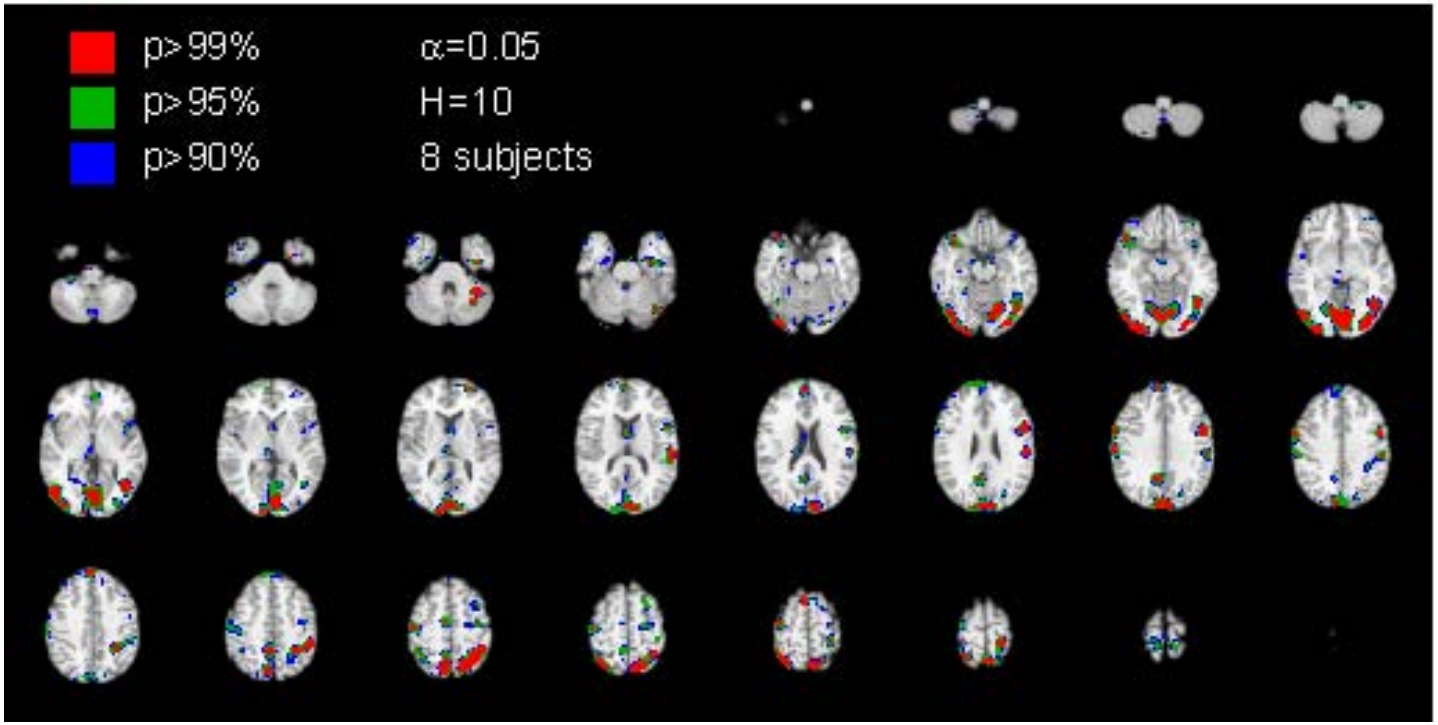


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both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

Reproducibility of internal representations

Predicting applied static force
with visual feed-back



Split-half resampling provides unbiased estimate of reproducibility of SPMs

NeuroImage: Strother et al (2002), Kjems et al. (2002), LaConte et al (2003), Strother et al (2004), …

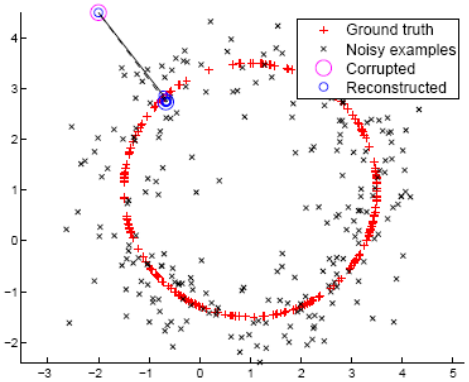
Visualization of latent manifold de-noising: The pre-image problem

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, (ii) if it does - there is no reason that it should be unique!

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



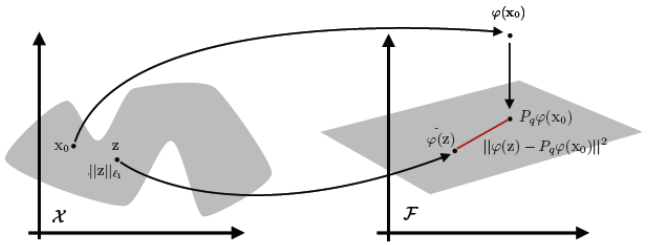


Figure 1: The pre-image problem in kernel PCA denoising concerns estimating z from $\mathbf{x}_0$, through the projection of the image onto the principal subspace in feature space, $\mathcal{F}$.

Mika, S., Schölkopf, B., Smola, A., Müller, K. R., Scholz, M., Rätsch, G. Kernel PCA and de-noising in feature spaces. In NIPS 11:536–542 (1999).

Regularization mechanisms for pre-image estimation in fMRI denoising

L2 regularization on denoising distance

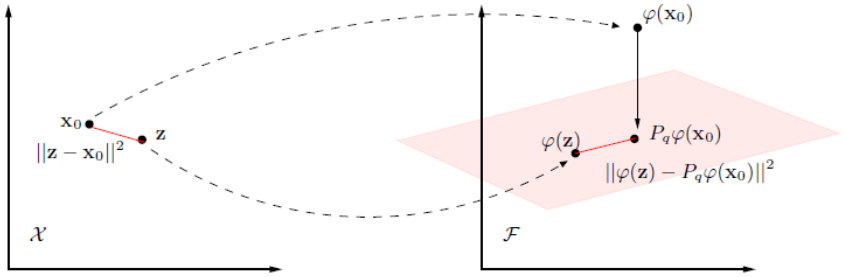


Figure 4.10: The pre-image problem in kernel PCA de-noising concerns estimating $\mathbf{z}$ from $\mathbf{x}_0$, through the projection of the image onto the principal subspace. Presently available methods for pre-image estimation lead to unstable pre-images because the inverse is ill-posed. We show 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

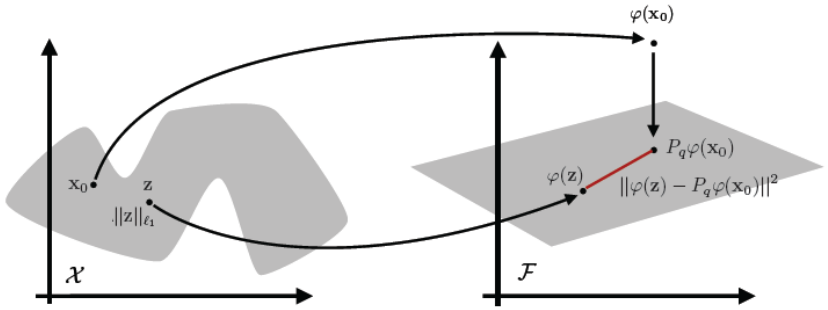


Figure 1: The pre-image problem in kernel PCA denoising concerns estimating $\mathbf{z}$ from $\mathbf{x}_0$, through the projection of the image onto the principal subspace in feature space, $\mathcal{F}$.

Optimizing denoising using the PR-plot: Sparsity, non-linearity

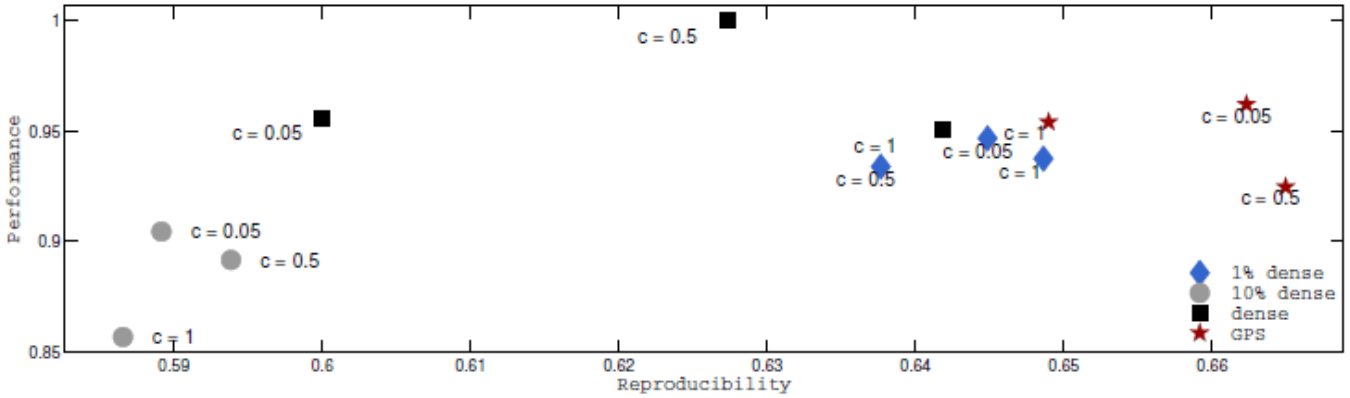


Figure 2: Prediction/reproducibility plots using all scans for the single slice fMRI visual block activation experiment. The GPS estimate when using a non-linear kernel are seen to outperform all other estimates in terms of combined prediction and reproducibility measures. Location in the upper right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

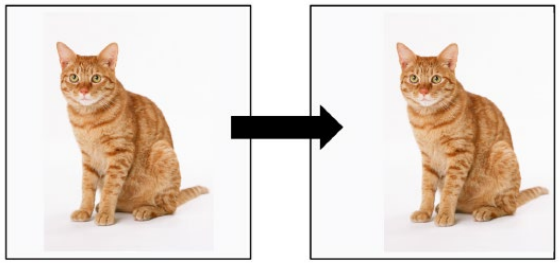
T.J. Abrahamsen and L.K. Hansen. Sparse non-linear denoising: Generalization performance and pattern reproducibility in functional MRI. Pattern Recognition Letters 32(15):2080-2085 (2011).

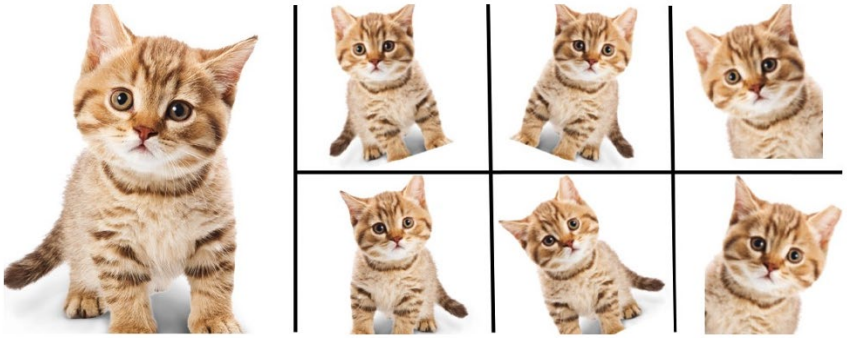
Spontaneous symmetry breaking

Understanding symmetry is of theoretical and practical interest:

Alex Krizhevsky, Ilya Sutskever, and Geofrey E Hinton. Imagenet classication with deep convolutional neural networks. In Advances in Neural Information Processing Systems 2012 - Cited by 56120

"*Without* data augmentation*, our network suffers from substantial overfitting, which would have forced us to use much smaller networks*."



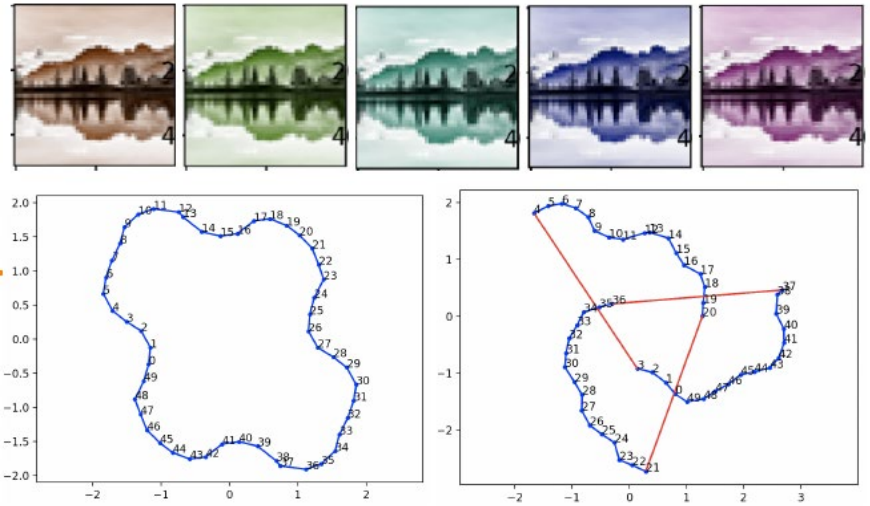


Latent variables –

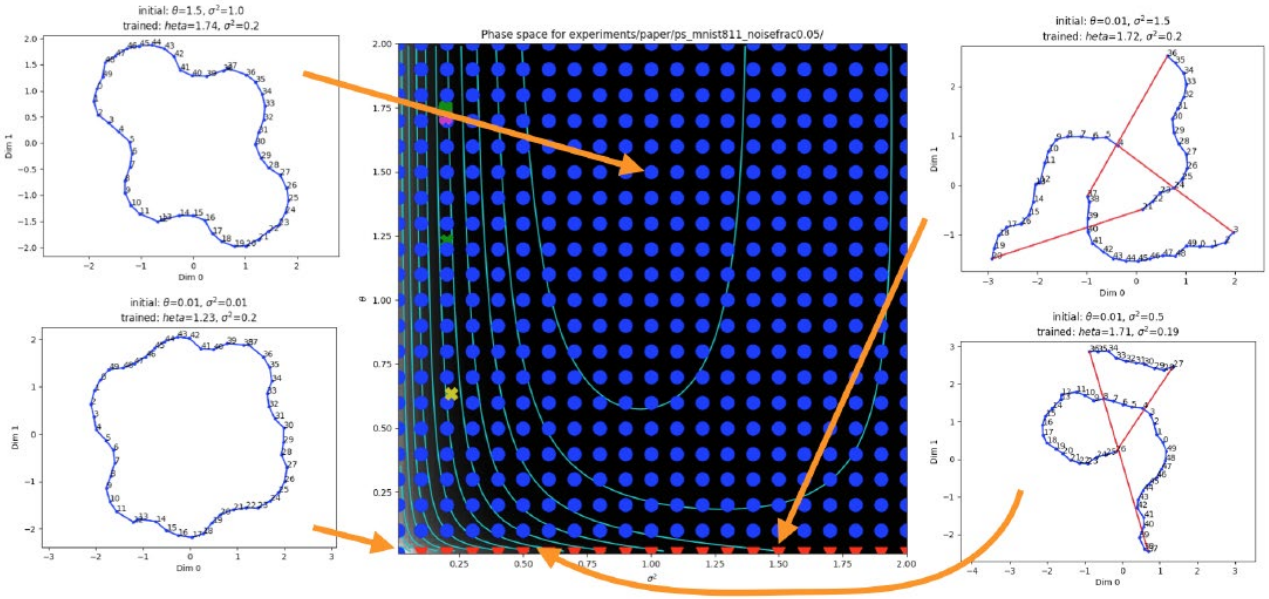
invariant – supervised learning

equivariant – representation learning

Symmetry breaking in kernel reps (GPLVM)

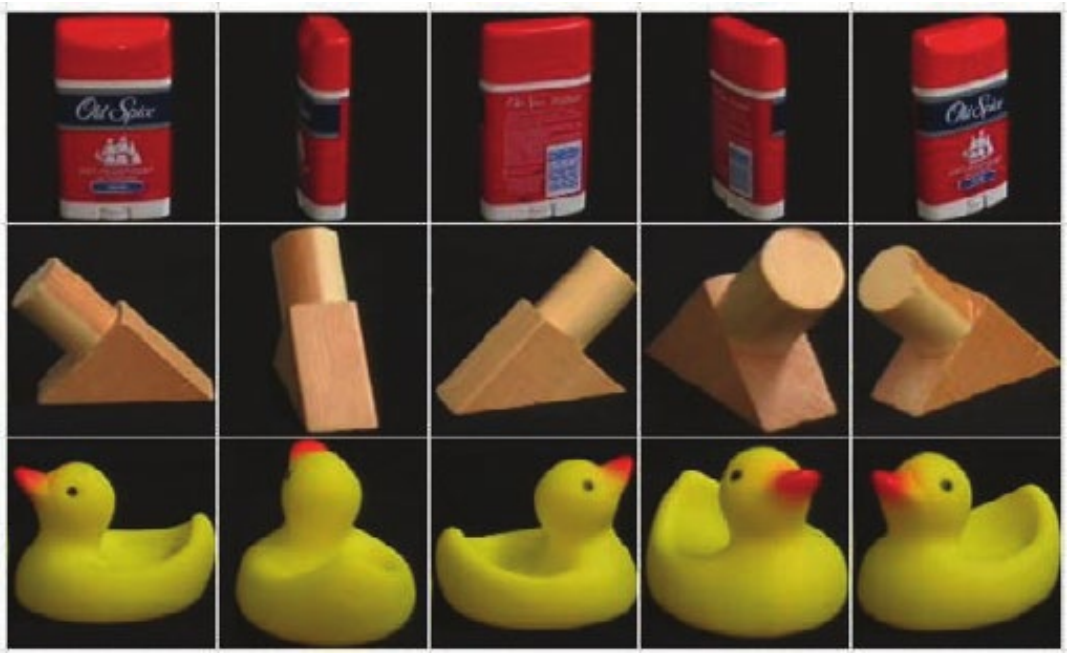


Phase space: Initialization

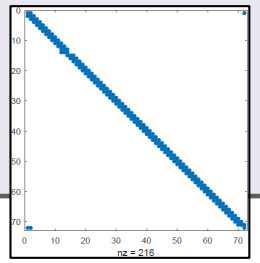


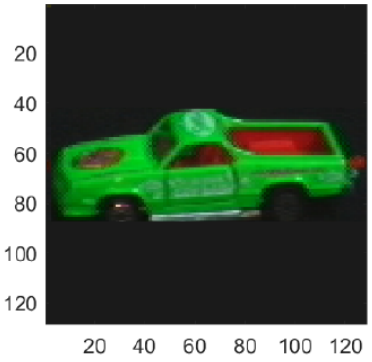
Lars Kai Hansen
Technical University of Denmark

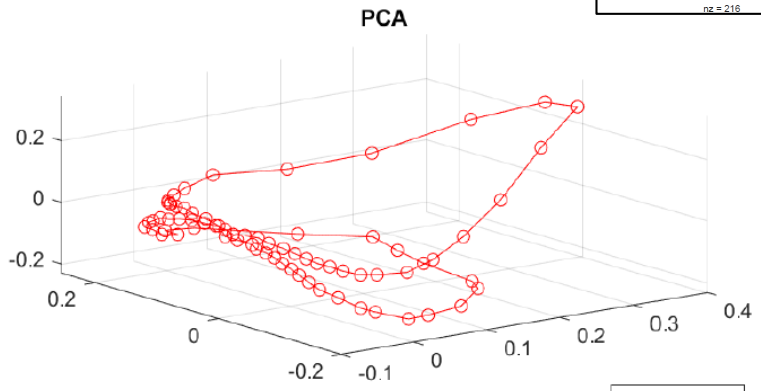
COIL 100 rotated objects

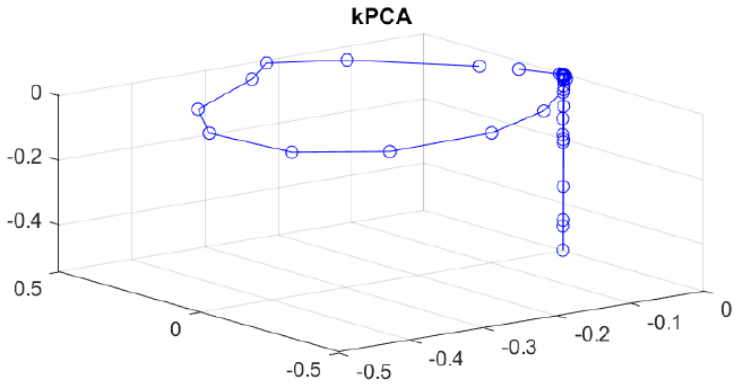


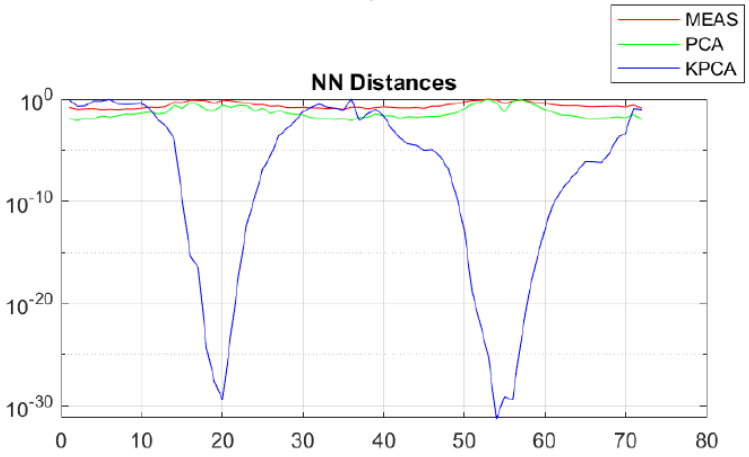
kPCA on COIL rotated objects



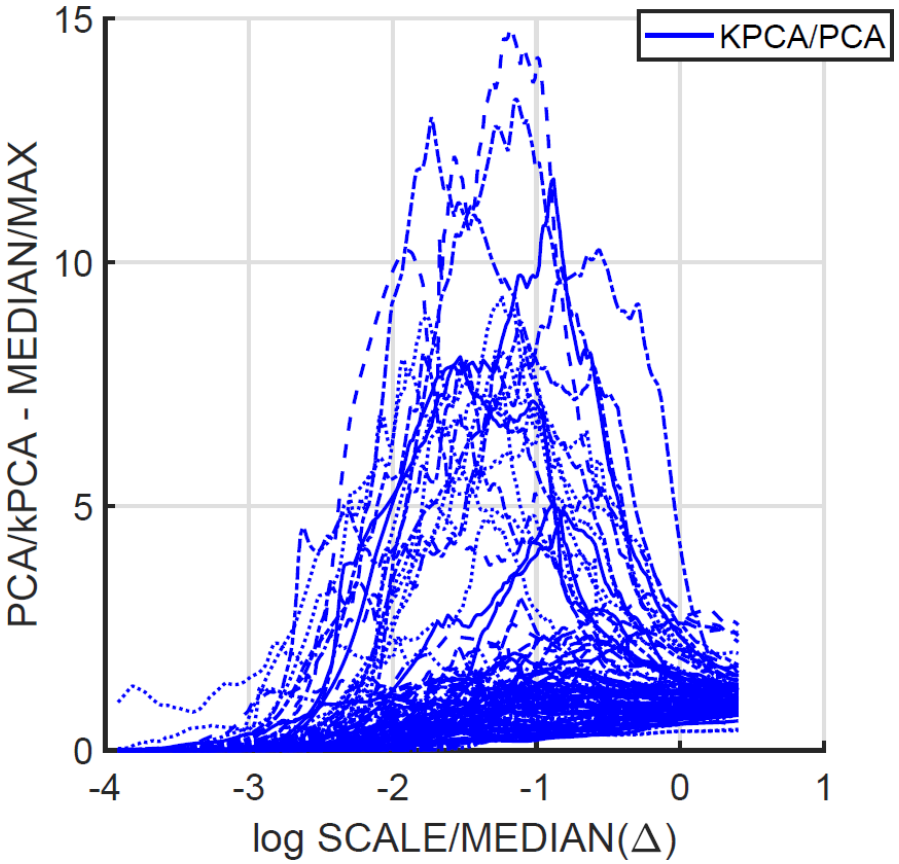


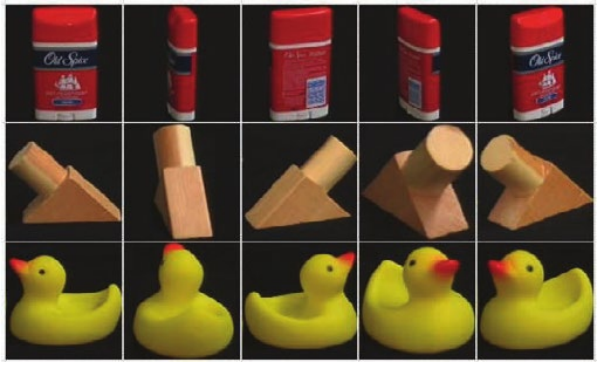


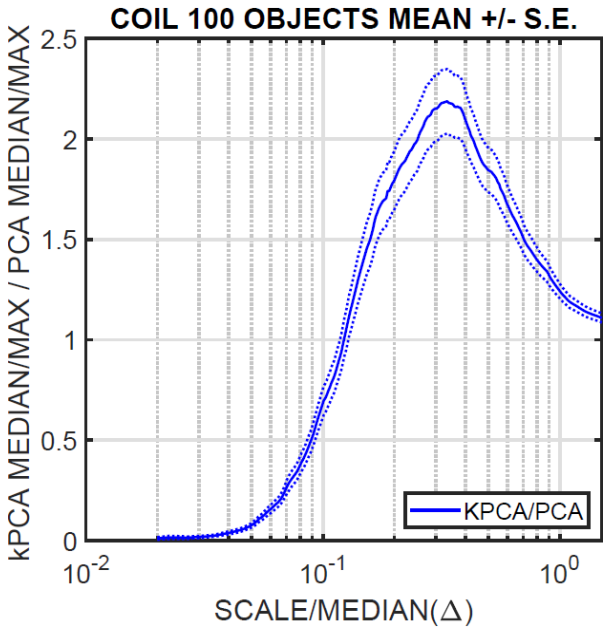




Universal pattern of symmetry breaking in KPCA on COIL

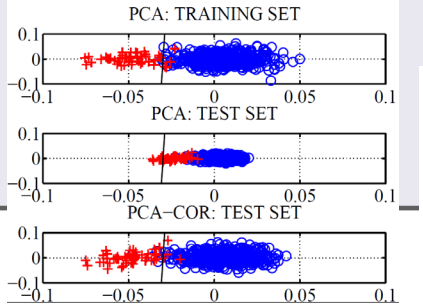


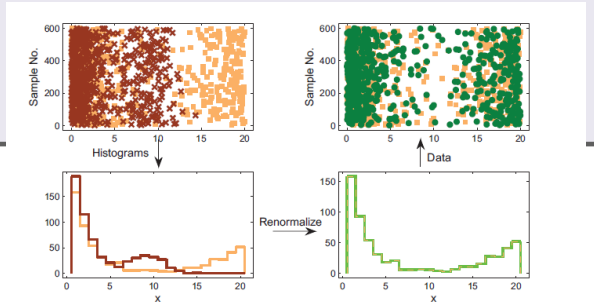




Lars Kai Hansen
Technical University of Denmark

Conclusion





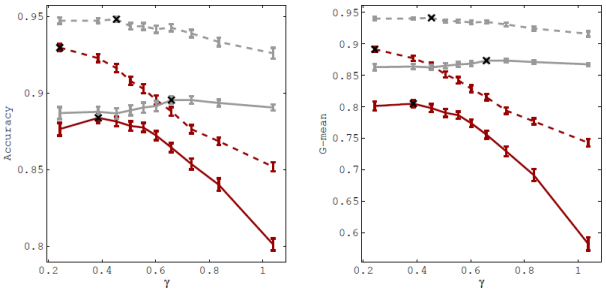
- Variance inflation in PCA

  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



- Kernel representations visualization is possible – uncertainty!

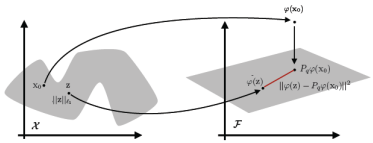
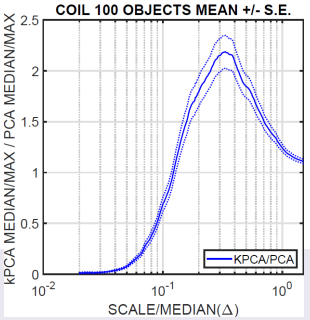


Figure 1: The pre-image problem in kernel PCA denoising concerns estimating a z f
through the projection of the image onto the principal subspace in feature space,

- Need to understand the (lack of) symmetry of latent variable models

  Is spontaneous symmetry breaking a "side effect"?



Acknowledgments

Lundbeck Foundation, Novo Nordisk Foundation
Danish Research Councils, Innovation Foundation Denmark