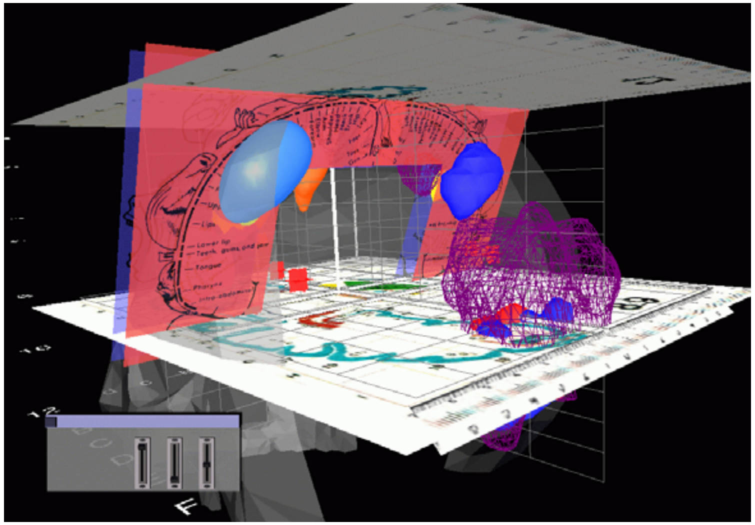
Generalization in high-dimensional factor models



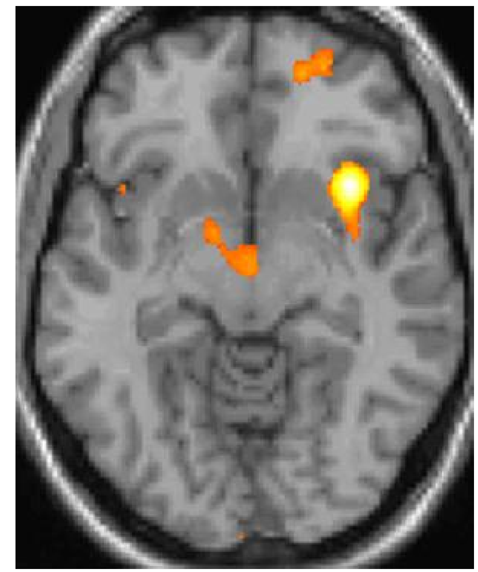
Lars Kai Hansen

DTU Informatics
Technical University of Denmark

MMDS

Modern massive data = modern ~~massive headache?~~

"Mr. Osborne, may I be excused? My brain is full."



Cluster headache: PET functional imaging shows activation of specific brain areas during pain

Do not multiply causes!

OUTLINE

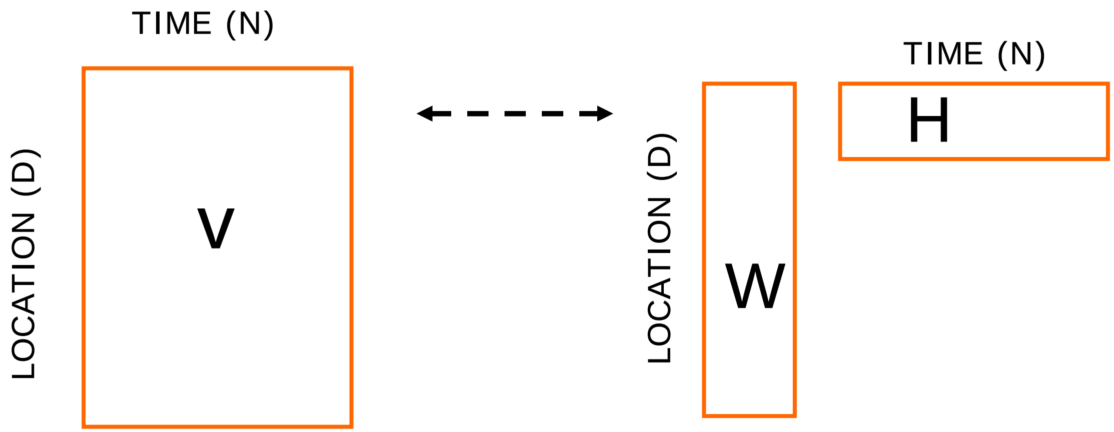
- Motivation
- Definition of generalizability
  - Operational definitions
  - Theory: Universality of learning curves

- Understanding the limits to learning in high-dimensional data
  - SVD/PCA: simple subspace models are well understood
  - ”Retarded” learning
  - What about ICA, NMF, Kmeans clustering, etc?

- Heuristics to heal bad factors in poor SNR’s
  - Re-scaling projections

Factor models

- Represent a datamatrix by a low-dimensional approximation



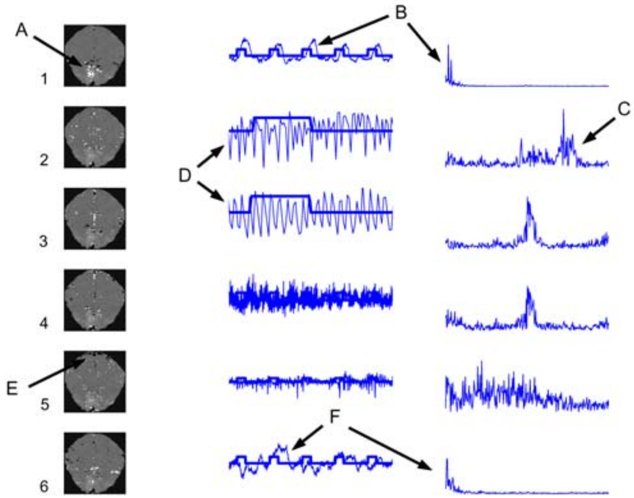
$$V(i,n) \approx \sum_{k=1}^{K} W(i,k)H(k,n)$$

....real world applications

- Many high-dimensional problems are analysed in pipelines with an initial dimension reduction step (SVD, NMF, ICA, VQ/kmeans, PLS, kOPLS etc)

- Unsupervised methods are less committed than supervised counterparts for exploratory investigations in eg.

  fMRI based ’mind reading’



(McKeown, Hansen, Sejnowski, 2003)

Matrix factorization: SVD / PCA, NMF, Clustering

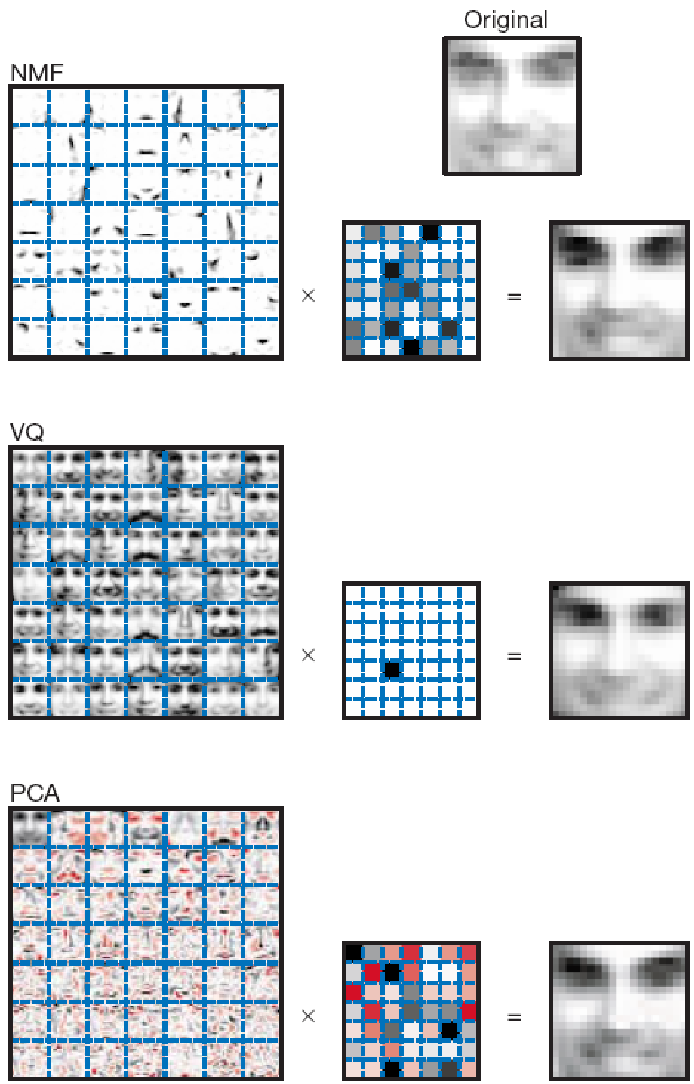


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting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

Probabilistic interpretation

(Gaussier, Goutte: Relation between PLSA and NMF.., 2005)

CASTSEARCH - CONTEXT BASED SPEECH DOCUMENT RETRIEVAL

Lasse Lohilahti Mølgaard, Kasper Winther Jørgensen, and Lars Kai Hansen

Informatics and Mathematical Modelling
Technical University of Denmark Richard Petersens Plads
Building 321, DK-2800 Kongens Lyngby, Denmark

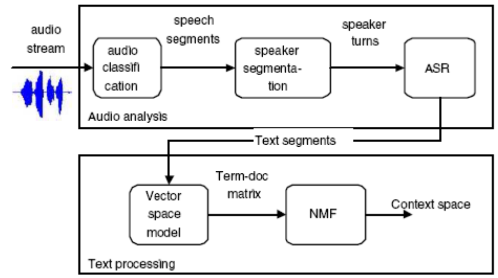
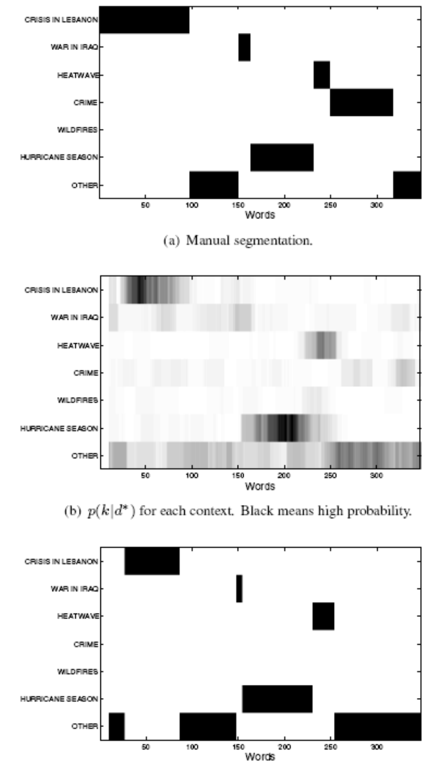


Fig. 1. The system setup. The audio stream is first processed using audio segmentation. Segments are then using an automatic speech recognition (ASR) system to produce text segments. The text is then processed using a vector representation of text and apply non-negative matrix factorization (NMF) to find a topic space.



(a) Manual segmentation.

(b) $p(k|d^*)$ for each context. Black means high probability.

(c) The segmentation based on $p(k|d^*)$.

Fig. 3. Figure 3(a) shows the manual segmentation of the news show into 7 classes. Figure 3(b) shows the distribution $p(k|d^*)$ used to do the actual segmentation shown in figure 3(c). The NMF-segmentation is in general consistent with the manual segmentation. Though, the segment that is manually segmented as 'crime' is labeled 'other' by the NMF-segmentation

Multinomial mixture model, V is a matrix of 'counts'

$$\frac{V(i,j)}{\sum_{i',j'} V(i',j')} \approx \sum_{k=1}^{K} W(i,k)H(k,j)$$

$$\frac{V(i,j)}{\sum_{i',j'} V(i',j')} \approx P(i,j) \approx \sum_{k=1}^{K} P(i,k)P(j,k)P(k)$$

Terms

Documents

(Mølgaard, Jørgensen, Hansen, ICASSP, 2007)

Generalizability

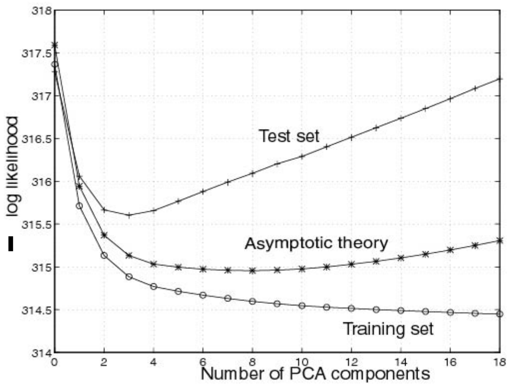
- Generalizability is defined as *the expected performance on a random new sample*
  - the av. performance of a model on a ”fresh” test data set is an unbiased estimate of generalization
  - in simulations how similar are estimated parameters to the ”true” values

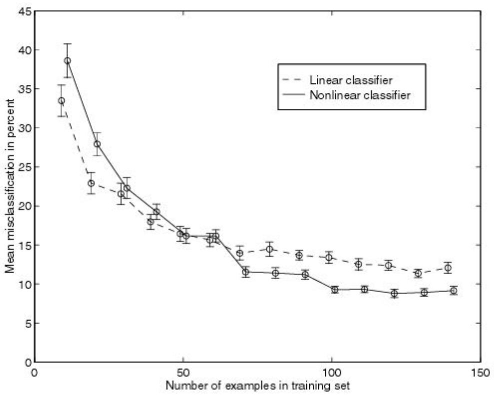
- Typical loss functions (supervised/unsupervised):

$$\langle -\log p(c \mid v, \theta) \rangle, \quad \langle -\log p(v \mid \theta) \rangle,$$

$$\langle (c - \hat{c})^2 \rangle, \quad \left\langle \log \frac{p(c, v \mid \theta)}{p(c \mid \theta)\, p(v \mid \theta)} \right\rangle$$

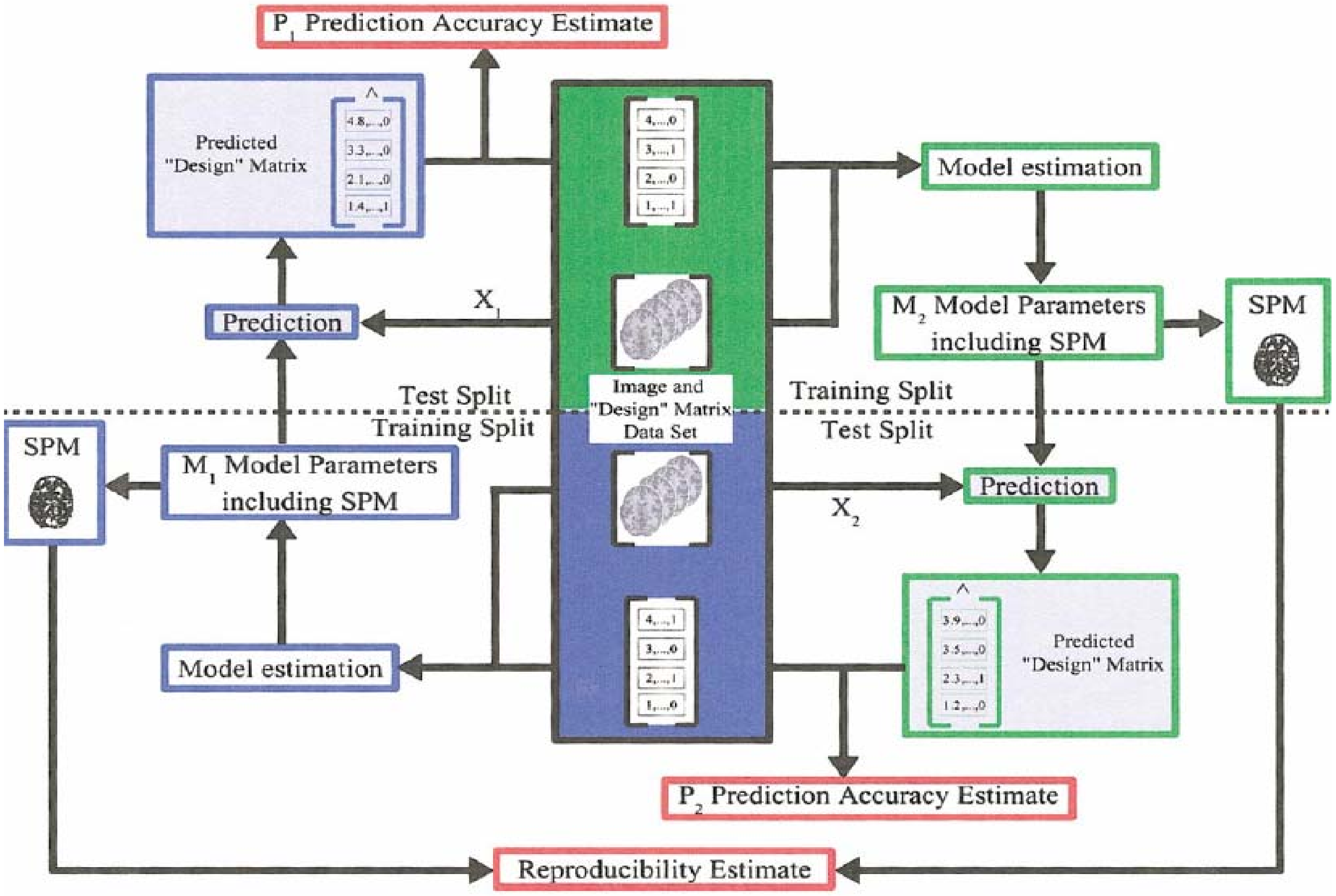
- Results can be presented as ”bias-variance trade-off curves” or ”learning curves”





Lars Kai Hansen
IMM, Technical University of Denmark

NPAIRS: Reproducibility of parameters



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)

Modeling the generalizability of SVD

(D. Hoyle, M. Rattray: Statistical mechanics of learning multiple orthogonal signals..., 2007):

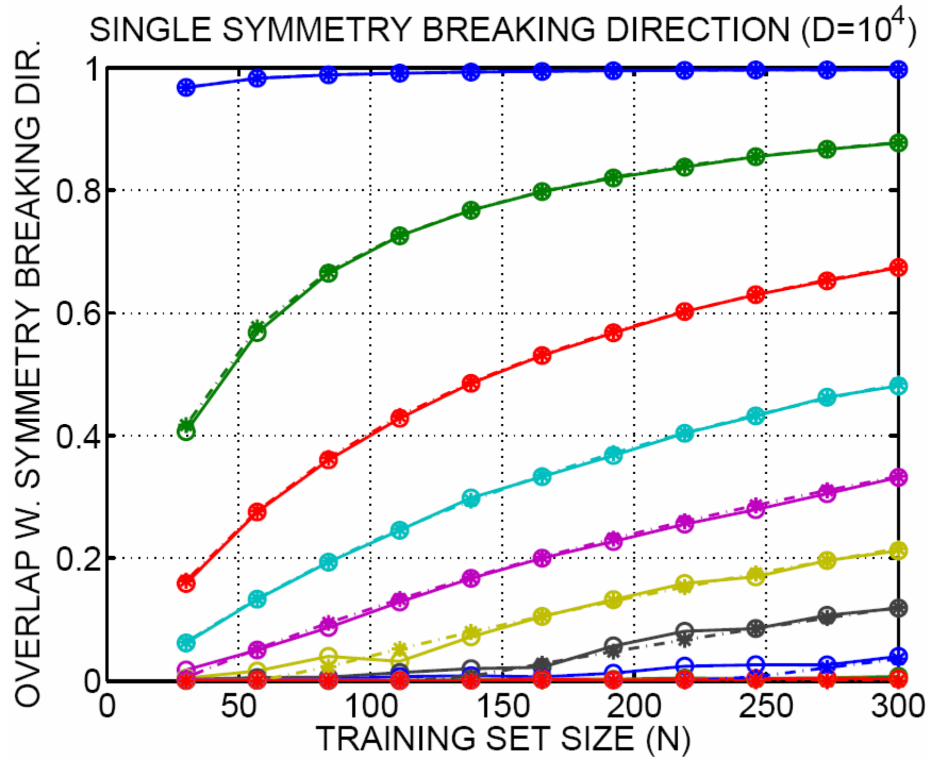
- Rich physics literature on "retarded" learning

- **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)

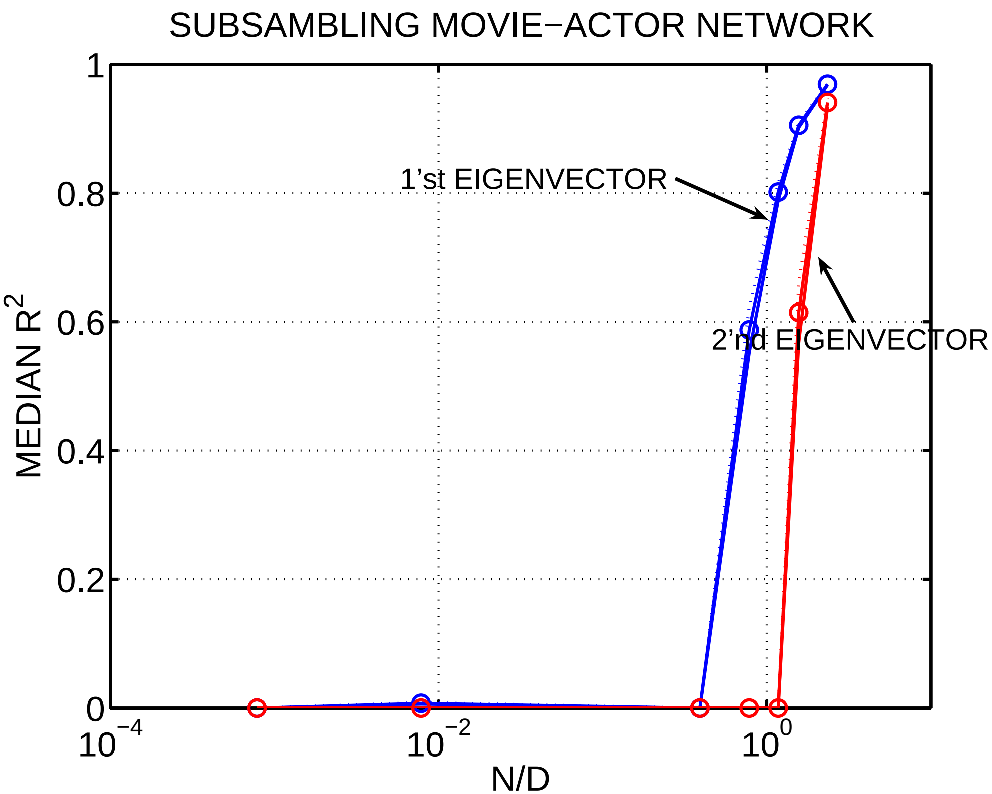


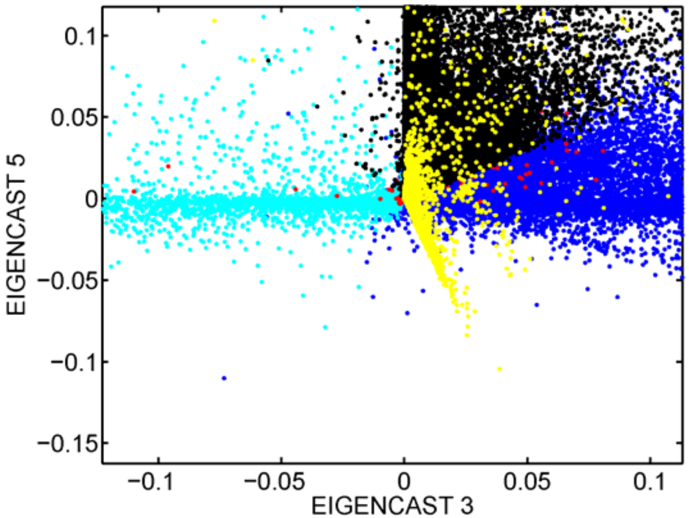
N_c = (0.0001, 0.2, 2, 9, 27, 64, 128, 234, 400, 625)

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

SVD of movie actor network -> ”eigencasts, eigengenre”

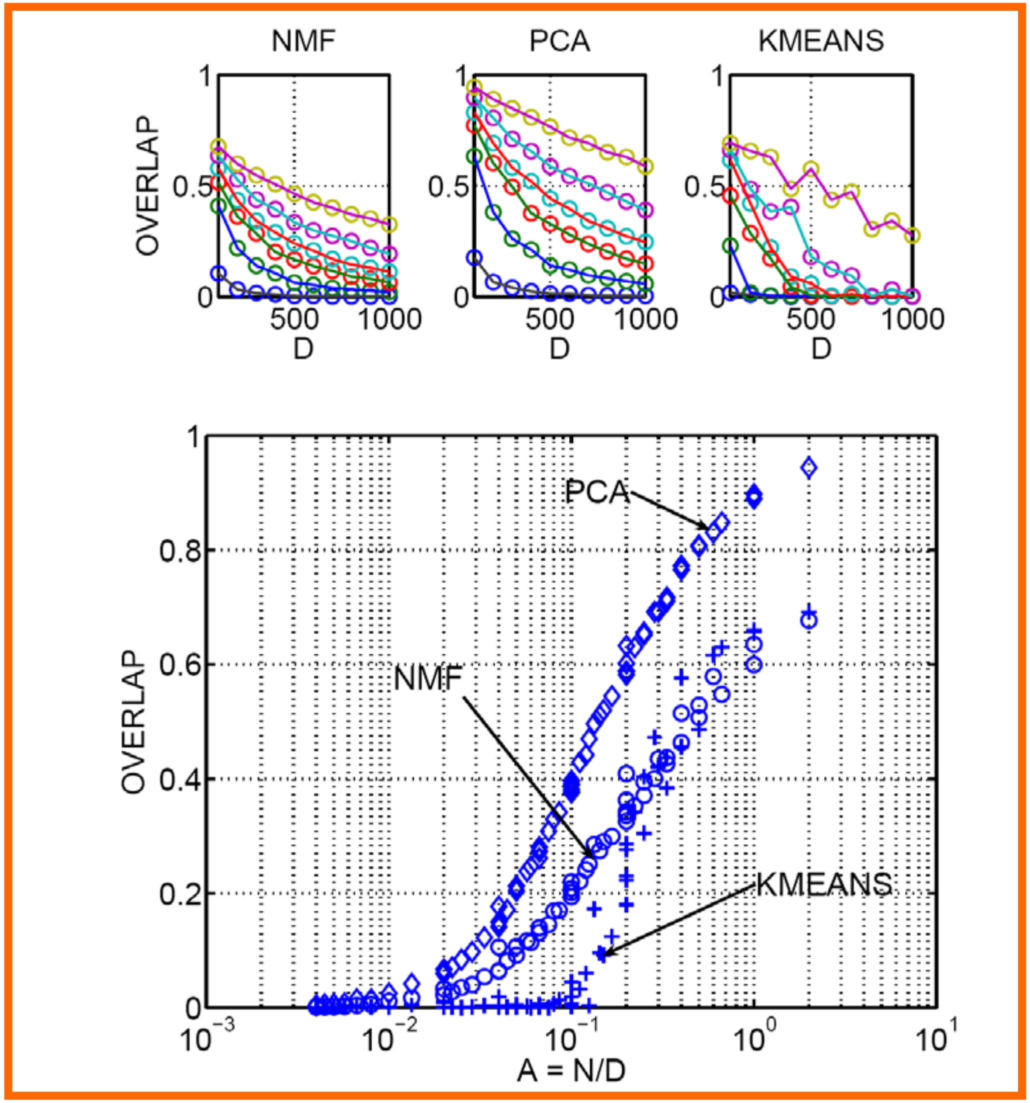
D = 128.000 movies
N = 400.000 actors





Universality in 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.



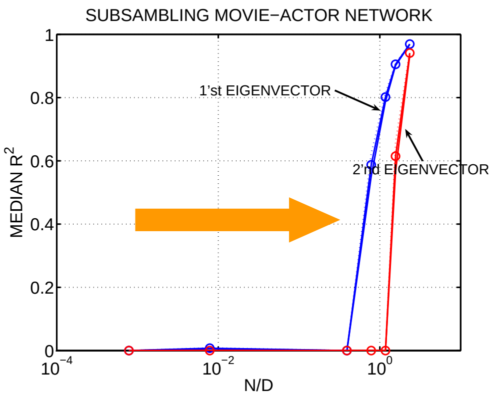
Experiment

Variable: N, D

Fixed: SNR

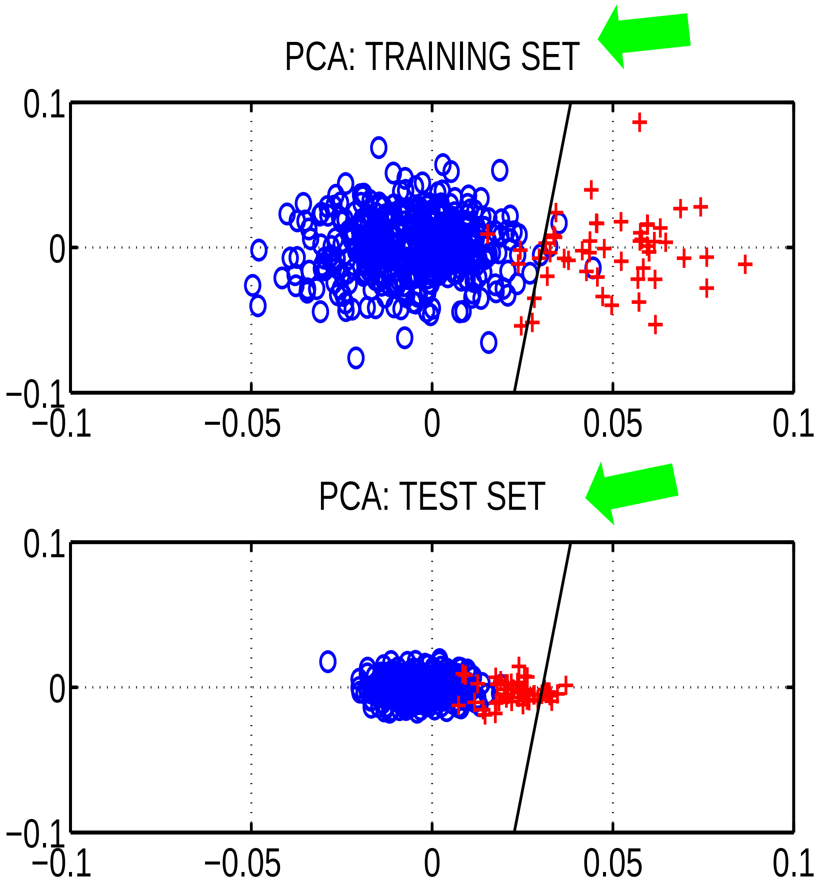
Restoring the generalizability of SVD

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

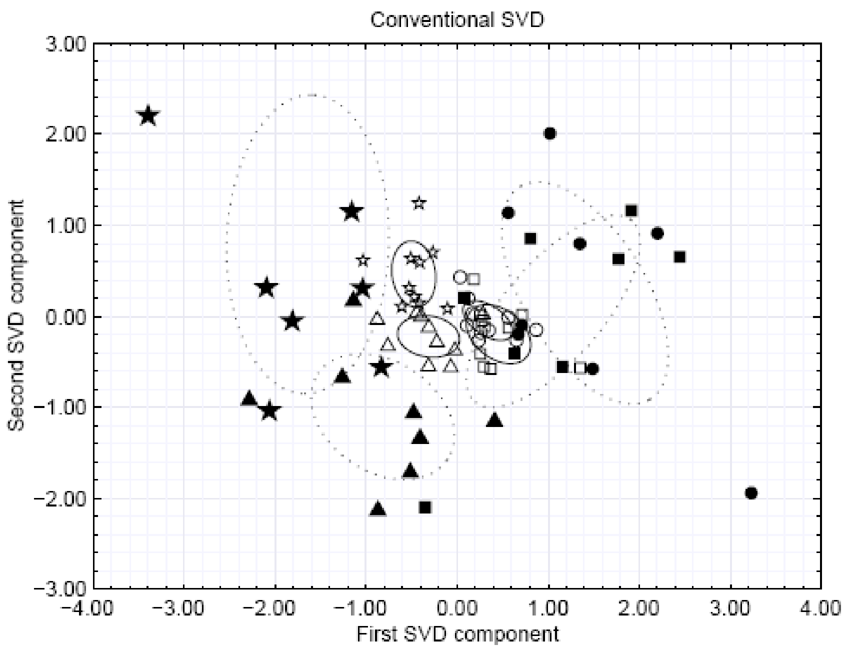


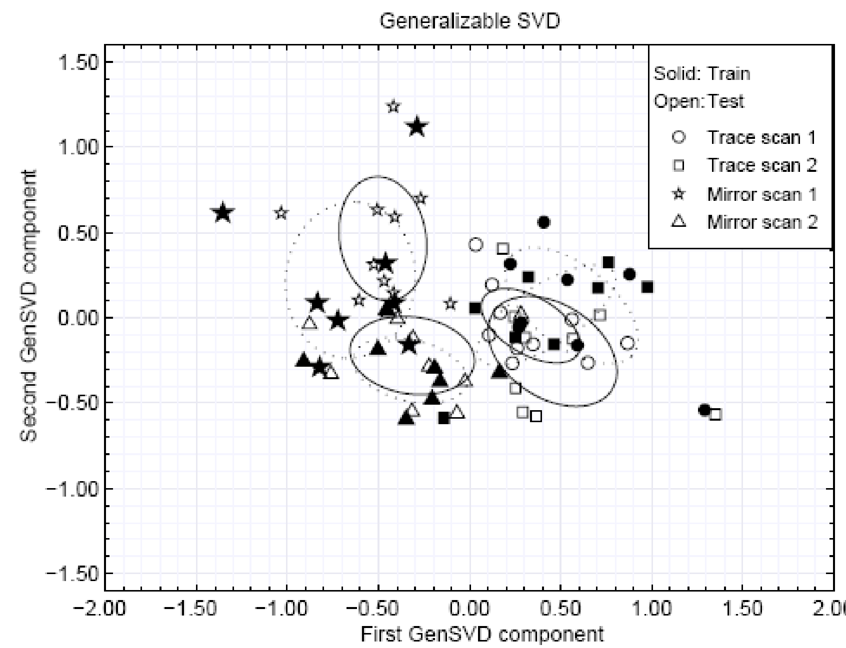
- The estimated projection is offset, hence, future projections will be too small!

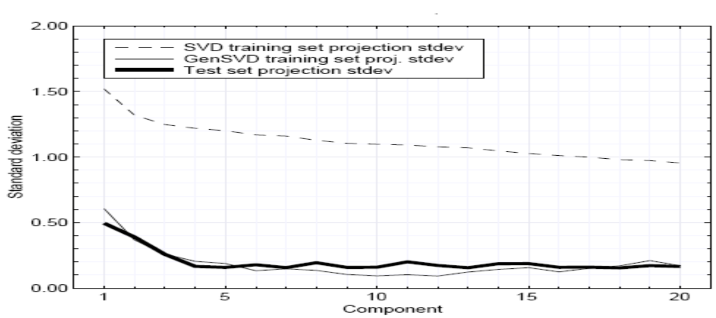
- …problem if discriminant is optimized for unbalanced classes in the training data!



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





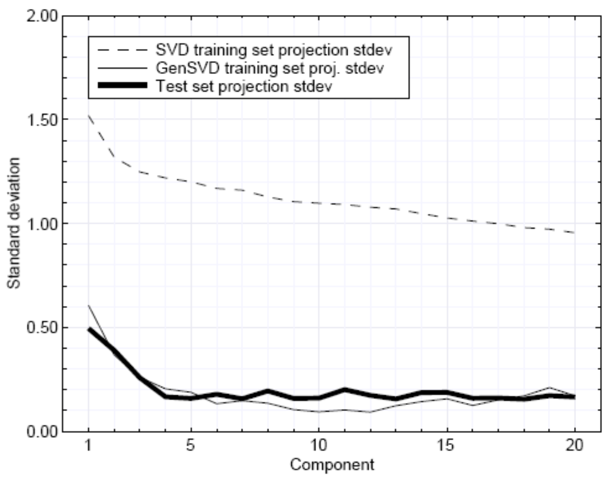


$N=72$, $D=2.5\ 10^4$

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

Re-scaling the component variances

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



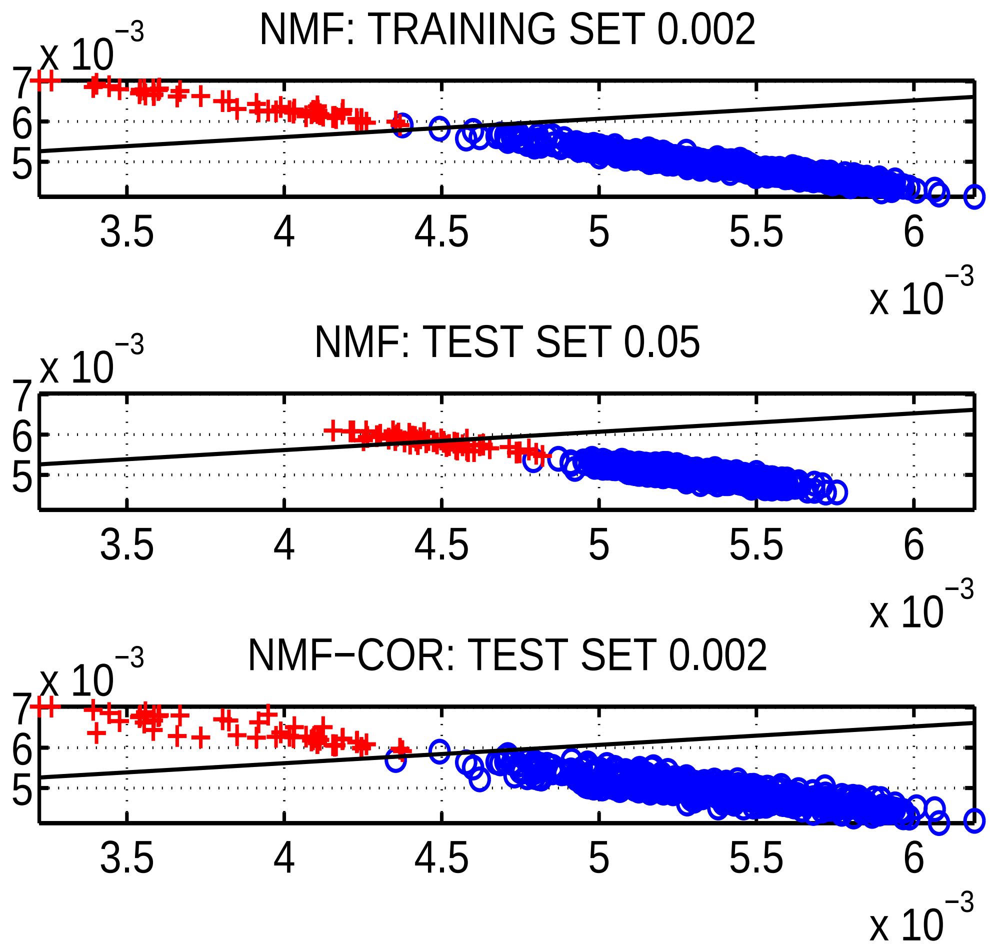
$$
\begin{aligned}
&\text{Compute } \boldsymbol{U}_0 \boldsymbol{\Lambda}_0 \boldsymbol{V}_0^\top = \mathrm{svd}(X) \text{ and } \boldsymbol{Q}_0 = \left[\, \boldsymbol{q}_j \,\right] = \boldsymbol{\Lambda}_0 \boldsymbol{V}_0^\top \\
&\text{foreach } j = 1 \ldots N \\
&\qquad \bar{\boldsymbol{q}}_{-j} = \tfrac{1}{N-1} \textstyle\sum_{j' \neq j} \boldsymbol{q}_{j'} \\
&\qquad \text{Compute } \boldsymbol{B}_{-j} \boldsymbol{\Lambda}_{-j} \boldsymbol{V}_{-j}^\top = \mathrm{svd}(\boldsymbol{Q}_{-j} - \bar{\boldsymbol{Q}}_{-j}) \\
&\qquad \boldsymbol{z}_j = \boldsymbol{B}_{-j} \boldsymbol{B}_{-j}^\top (\boldsymbol{q}_j - \bar{\boldsymbol{q}}_{-j}) \\
&\hat{\lambda}_i^2 = \tfrac{1}{N-1} \textstyle\sum_j z_{ij}^2
\end{aligned}
$$

Kjems, Hansen, Strother: NIPS (2001)

Re-scaling for other factorizations: NMF?

- Test projections are obtained by running the factorization alg with W fixed
- NMF suffers from the same distributional problem as SVD
- Simple scaling can fail because of non-normal distributions
- Use histogram equalization for re-mapping the densities of the factors

- Implicit hypothesis:
  - NMF factors are approx independent



Conclusion & Perspectives

- Evidence of universality in SVD/PCA, NMF, Kmeans,
- Evidence for ”phase transition”-like learning curves in high-dimensional unsupervised learning
- Working heuristic for re-scaling of projection on test set
  - Linear scaling in SVD/PCA
  - Non-linear scaling in NMF
- More formal investigation of NMF, Kmeans, higher order factorizations, etc – how universal are the learning curves
- Structured/sparse matrices?
  - How do priors shift the phase transition?
  - Multiple order parameters: Sequence of phase transitions ala fractal structure in disordered systems

Thanks and … a little add placement

Sponsors

Lundbeck foundation

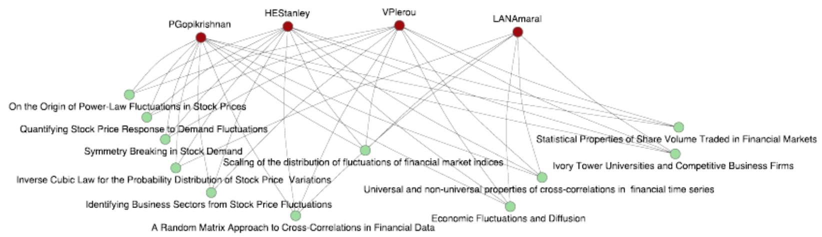
Danish Research Councils

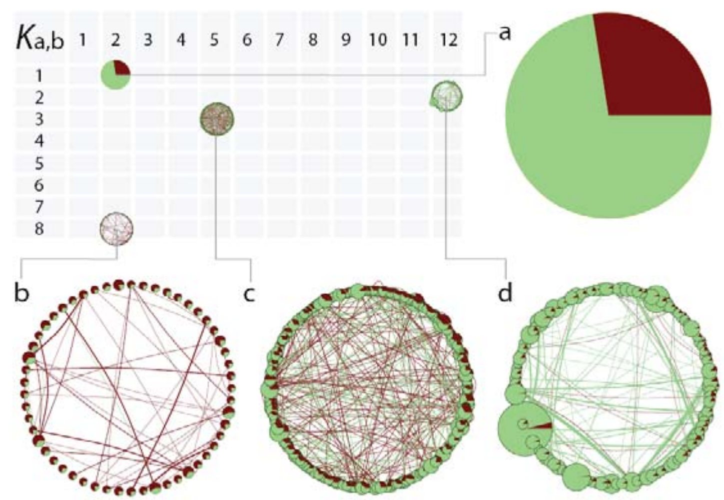
NIH

EU Commission

BCFinder: Tool for bi-clique community detection

http://www2.imm.dtu.dk/~mhs/bcfinder/





Lars Kai Hansen

IMM, Technical University of Denmark