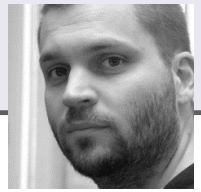
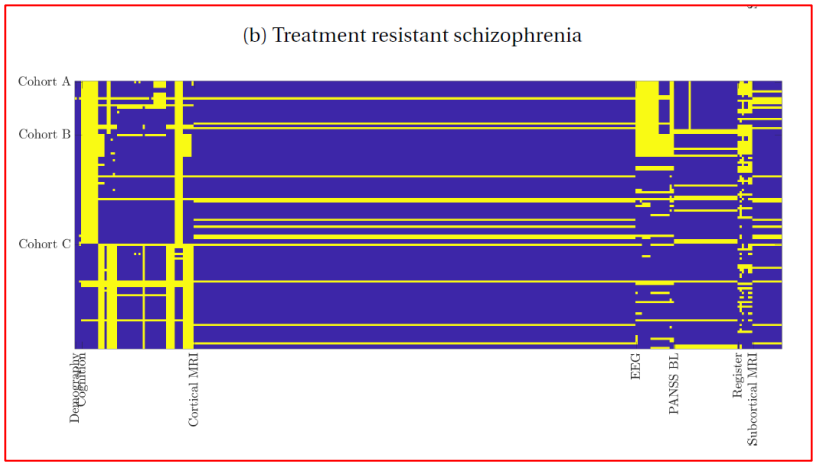
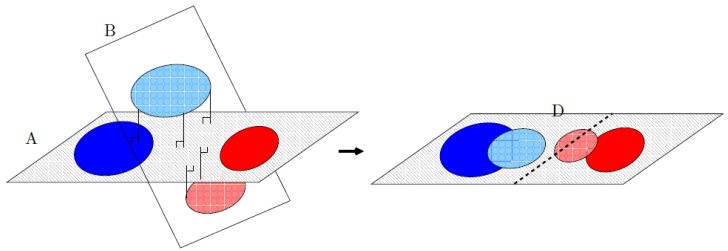
Phase transition in PCA with missing data
--Reduced signal-to-noise ratio, not sample size!

Lars Kai Hansen
DTU Compute
Technical University of Denmark

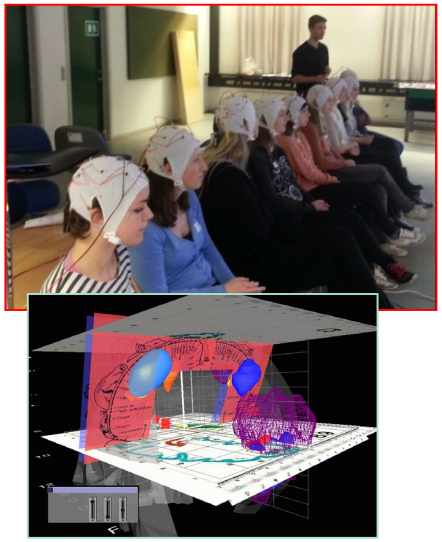




Joint work with Trine Abrahamsen
Nikolaj Bak, Niels Ipsen





Safe AI - Cognitive Systems platform at DTU Compute



Research areas

Machine learning

- Probabilistic ML, deep learning, networks, geometry, +1500 eng. students/year

Cognition

- Neurotechnology, Knowledge graphs, Hon. prof Sid Kouider (ENS),

Computational social science

- Sensible DTU, leader/follower dynamics, SODAS (KU, Social Science)

Excellence measured in top venue papers
Co-author network: Stanford, MIT, UCLA, UC London, ENS Paris, TU Berlin,
Top rated by three international review panels (2009/2014/2018)

Societal impact

Hearing Systems' high AI acceptance

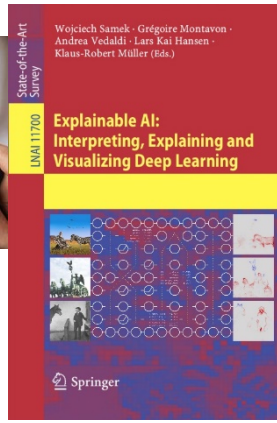
Audio ML, Personalization, neurotechnology

DABAI open source ML workflows + Language / Danish resources

Start-ups: Peergrade, Spektral Experience, Corti,

New AI BSc education – first class: September 2018

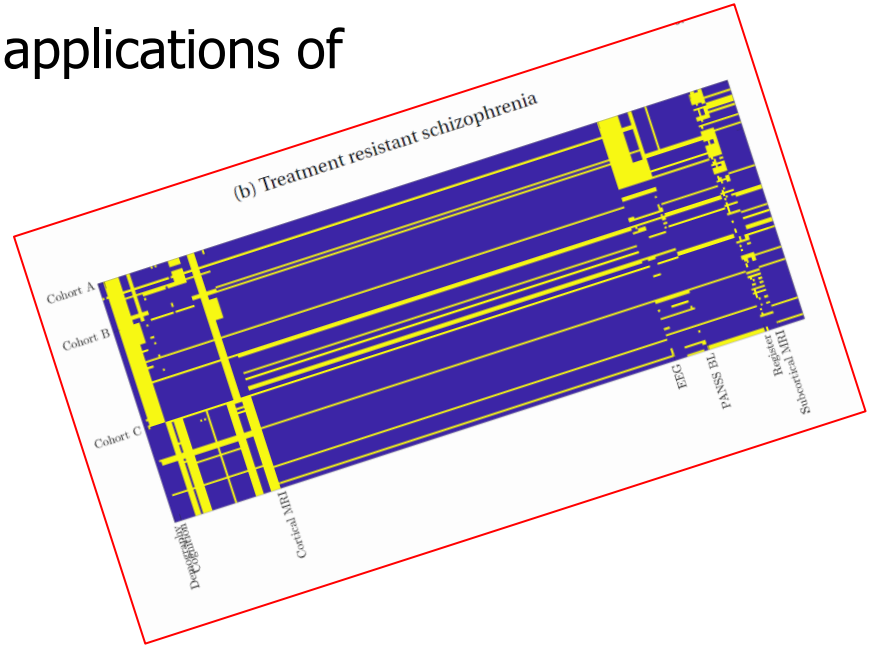
DABAI



Lars Kai Hansen
Technical University of Denmark

Missing data

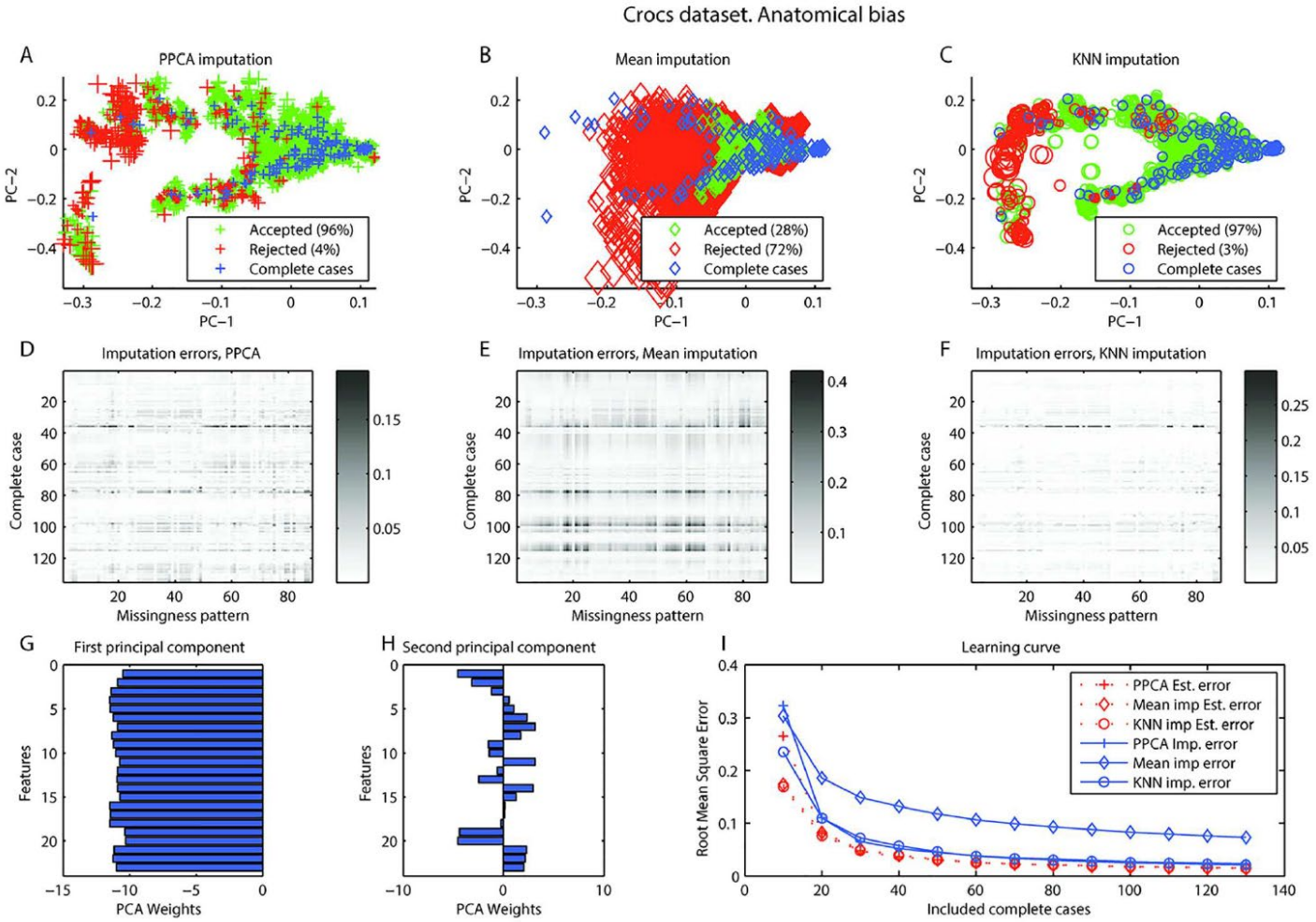
….an important problem in real life applications of Machine Learning e.g. computational psychiatry



Questions

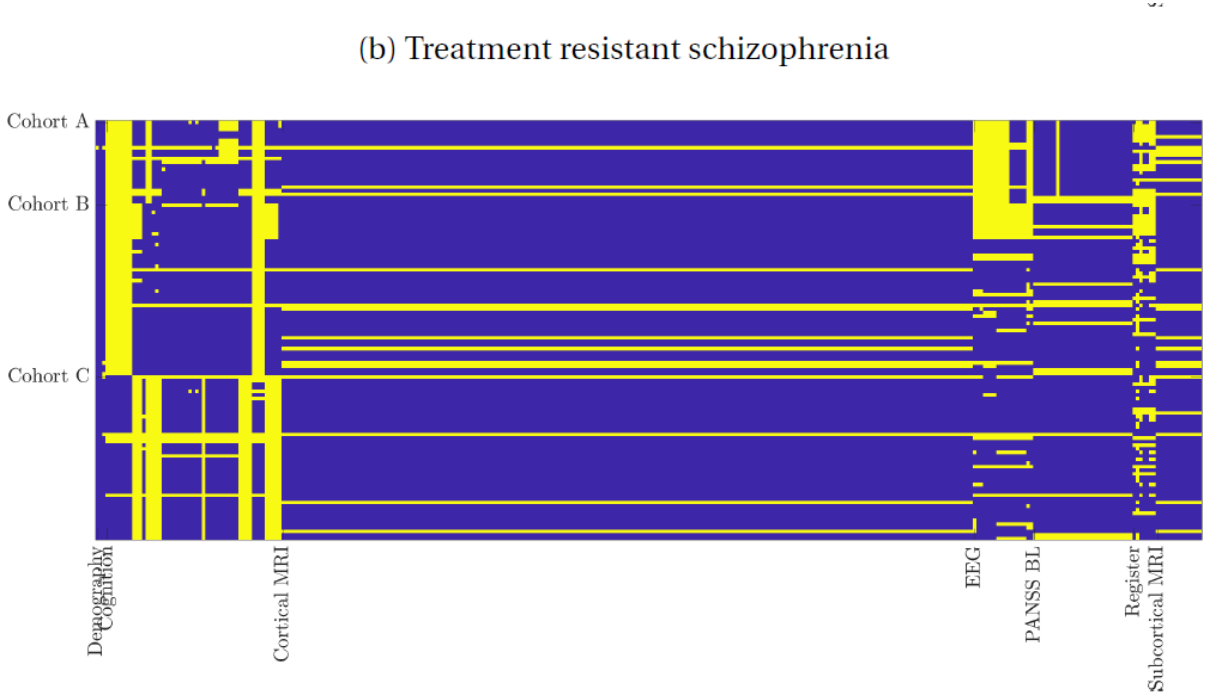
- Impute – can we trust it?
- What are the implications of missing data?
- The unexpected richness of principal component analysis

Data Driven Estimation of Imputation Error
—A Strategy for Imputation with a Reject Option



Bak, N. and Hansen, L.K., 2016.
Data driven estimation of imputation error—a strategy for imputation with a reject option.
PloS one, *11*(10), p.e0164464.

Missing data in computational psychiatry - inference based on marginal

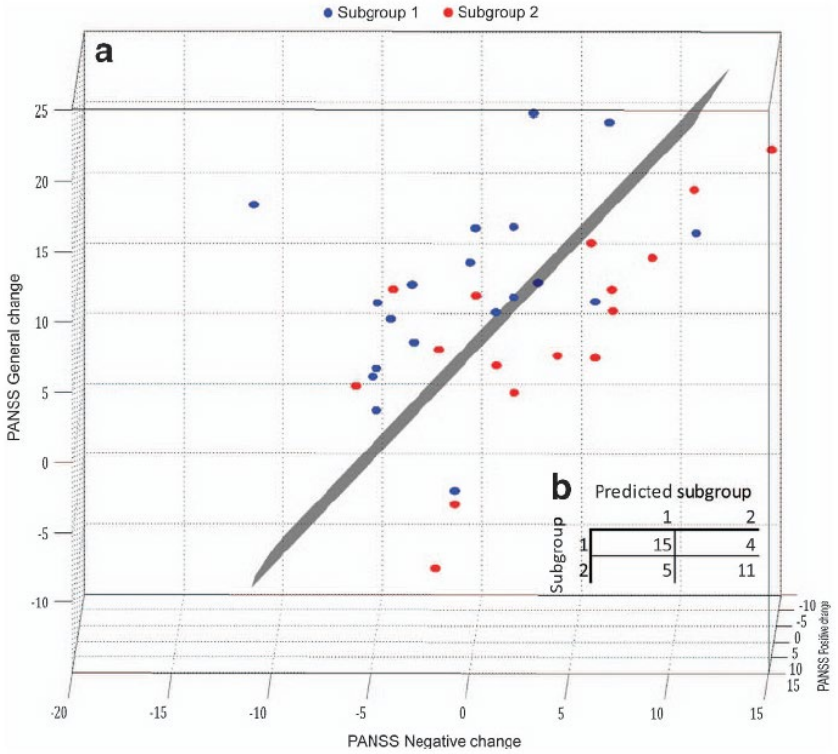


$$\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon}$$

$$\widehat{\mathbf{s}}(\mathbf{x}_m) = \left(\mathbf{A}_m^T \mathbf{A}_m + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{A}_m^T \mathbf{x}_m$$

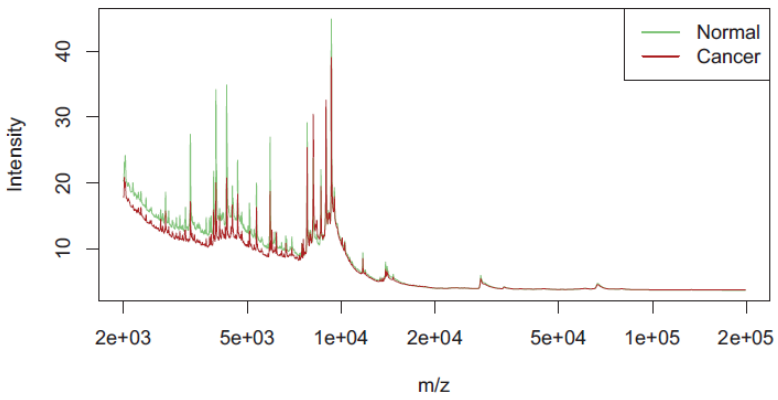
Two subgroups of antipsychotic-naive, first-episode schizophrenia patients identified with a Gaussian mixture model on cognition and electrophysiology

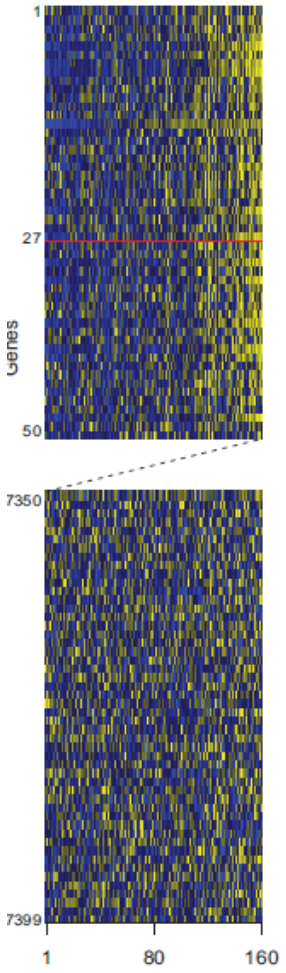
Feature	Schizophrenia 1	Schizophrenia 2	Controls	PC1	PC2	PC3	PC4
Premorbid IQ (DART)	23.65(10.37)	19.61(7.63)	23.06(6.81)				
Verbal IQ a	106.28(21.17)	98.72(13.93)	113.50(14.11)				
Performance IQ a	103.64(13.97)	97.28(13.44)	108.75(13.32)				
Full scale IQ a	105.72(18.12)	98.11(12.18)	112.50(12.55)				
List learning b	52.46(7.18)	51.06(13.37)	55.23(8.32)				
Digit Sequencing Task b	21.38(4.08)	20.11(3.92)	22.89(3.65)				
Token Motor Task b	68.23(10.81)	72.78(15.19)	76.08(13.30)				
Verbal Fluency ("supermarket") b	23.88(6.93)	26.22(9.08)	31.25(8.00)				
Verbal Fluency ("F") b	13.46(3.67)	12.22(5.41)	15.66(4.22)				
Verbal Fluency ("S") b	14.46(4.59)	12.89(5.86)	16.28(4.76)				
Symbol coding b	59.69(10.13)	54.50(15.56)	65.83(11.66)				
Tower of London b	18.73(1.93)	18.56(2.48)	19.51(2.00)				
Spatial Span c	7.19(1.41)	6.94(1.30)	7.51(1.25)				
Spatial Working Memory (strategy) c	26.69(6.42)	29.56(6.31)	24.36(5.37)				
Spatial Working Memory (total errors) c	11.54(14.92)	15.33(19.44)	7.13(8.43)				
Stockings of Cambridge (problems solved) c	9.65(1.83)	9.06(1.51)	10.21(1.35)				
Stockings of Cambridge (initial thinking time) c	9,470(8,003)	9,872(4,810)	12,472(7,545)				
Intra-Extra Dimensional Set Shift (stages) c	9.00(0.00)	7.89(0.96)	8.87(0.48)				
Intra-Extra Dimensional Set Shift (errors) c	11.42(5.09)	39.94(20.78)	14.60(13.14)				
Intra-Extra Dimensional Set Shift (EDS errors) c	3.38(3.07)	17.89(11.22)	6.25(7.68)				
Reaction Time (simple reaction) c	337.82(56.79)	322.71(35.20)	304.51(33.77)				
Reaction Time (simple movement) c	456.72(127.08)	455.81(113.01)	455.90(141.34)				
Reaction Time (choice reaction) c	410.97(78.84)	381.42(60.91)	356.42(66.47)				
Reaction Time (choice movement) c	402.50(111.79)	409.03(95.58)	413.05(120.22)				
Rapid Visual Processing (A', 3-5-7) c	0.98(0.02)	0.97(0.02)	0.99(0.01)				
Rapid Visual Processing (A', 3-5-7, 2-4-6) c	0.96(0.03)	0.95(0.04)	0.97(0.02)				
P50 c-stimulus Amplitude	1.29(0.74)	2.03(1.40)	1.70(0.99)				
P50 t-stimulus Amplitude	0.43(0.49)	0.74(0.88)	0.55(0.55)				
P50 t/c ratio	0.36(0.49)	0.40(0.43)	0.30(0.29)				
P50 c-stimulus Latency	56.72(9.78)	58.44(10.90)	58.88(8.88)				
PPI 85dB 120ms	53.50(40.66)	50.28(29.65)	58.89(37.25)				
PPI 85dB 60ms	57.00(28.00)	50.89(17.38)	58.13(33.01)				
PPI 76dB 120ms	37.77(40.11)	29.56(32.16)	48.87(34.64)				
PPI 76dB 60ms	37.04(39.90)	32.00(27.71)	36.89(48.17)				
PPI Pulse alone	106.19(73.59)	208.06(141.25)	138.28(127.10)				
PPI 85dB 120ms amplitude	40.38(35.81)	93.17(92.69)	56.72(93.12)				
PPI 85dB 60ms Amplitude	39.19(31.92)	92.28(66.52)	55.92(86.18)				
PPI 76dB 120ms Amplitude	58.19(43.75)	135.61(114.29)	74.25(111.11)				
PPI 76dB 60ms Amplitude	59.35(38.96)	153.39(129.16)	83.89(114.38)				
MMN Frequency deviant FCZ	-2.74(1.25)	-2.83(1.42)	-2.49(1.13)				
MMN Duration deviant FCZ	-4.27(2.10)	-3.75(1.40)	-4.01(1.60)				
MMN Frequency and duration deviant FCZ	-3.82(1.72)	-3.86(1.80)	-4.16(1.43)				
MMN Frequency deviant Latency	131.20(38.56)	131.78(43.07)	140.83(41.59)				
MMN Duration deviant Latency	194.33(36.10)	184.44(46.73)	187.55(34.59)				
MMN Frequency and Duration deviant Latency	142.88(37.71)	122.00(32.36)	146.49(40.72)				
Principal Component 1	0.0155(0.1007)	0.1023(0.1127)	-0.0419(0.0788)				
Principal Component 2	-0.0273(0.0613)	0.0271(0.1088)	-0.0008(0.1159)				
Principal Component 3	-0.0496(0.0868)	0.0491(0.0683)	0.0089(0.1087)				
Principal Component 4	0.0418(0.0796)	-0.0697(0.1594)	0.0072(0.0774)				



Bak, N., Ebdrup, B.H., Oranje, B., Fagerlund, B., Jensen, M.H., Düring, S.W., Nielsen, M.Ø., Glenthøj, B.Y. and Hansen, L.K., 2017. Two subgroups of antipsychotic-naive, first-episode schizophrenia patients identified with a Gaussian mixture model on cognition and electrophysiology. *Translational psychiatry*, *7*(4), p.e1087.

High dimensions – small samples (D >> N)

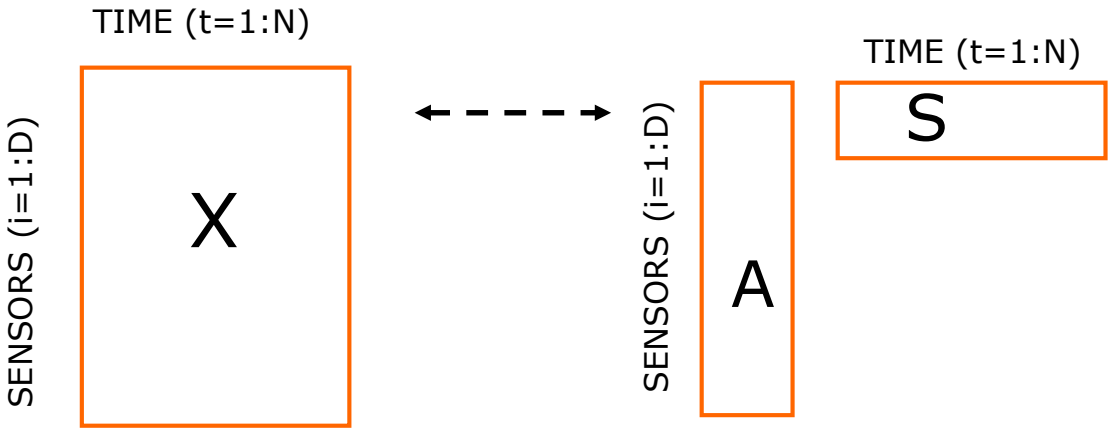




"HDLSS" high dimension, low sample size (Hall 2005, Ahn et al, 2007)
"Large p, small n" (West, 2003), "Curse of dimensionality" (Occam, 1350)
"Large underdetermined systems" (Donoho, 2001)
"Ill-posed data sets" (Kjems, Strother, LKH, 2001)

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

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

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

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

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

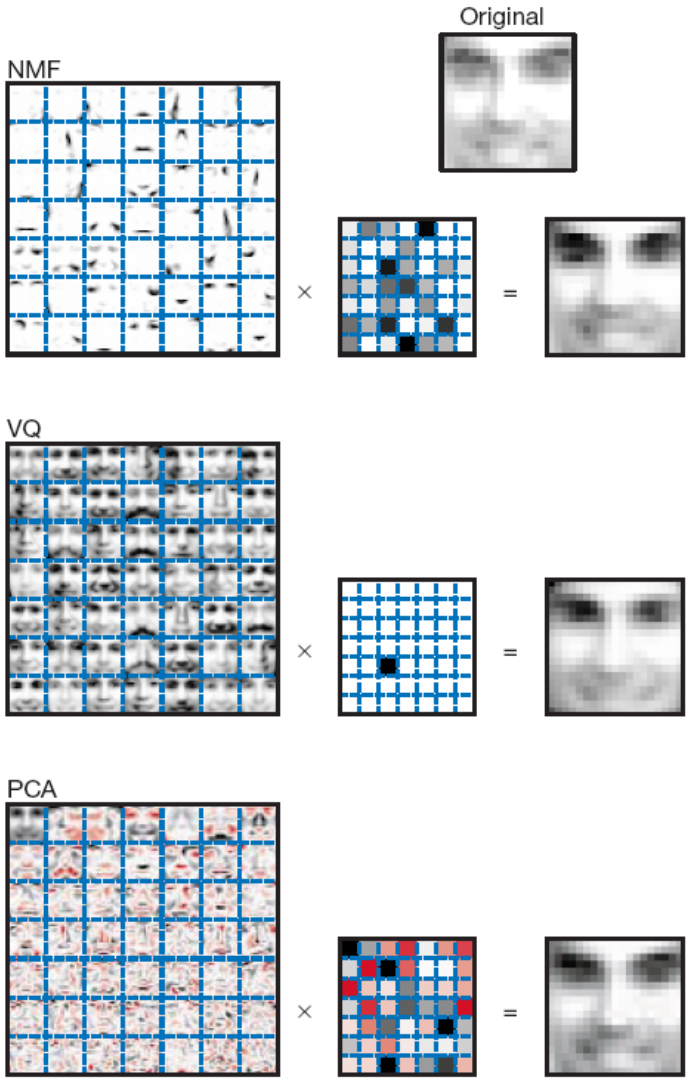


Figure 1 Non-negative matrix factorization (NMF) learns a parts-based representation of faces, whereas vector quantization (VQ) and principal components analysis (PCA) learn holistic representations. The three learning methods were applied to a database of $m = 2{,}429$ facial images, each consisting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

Modeling the generalizability of SVD

Rich physics literature on ”retarded” learning

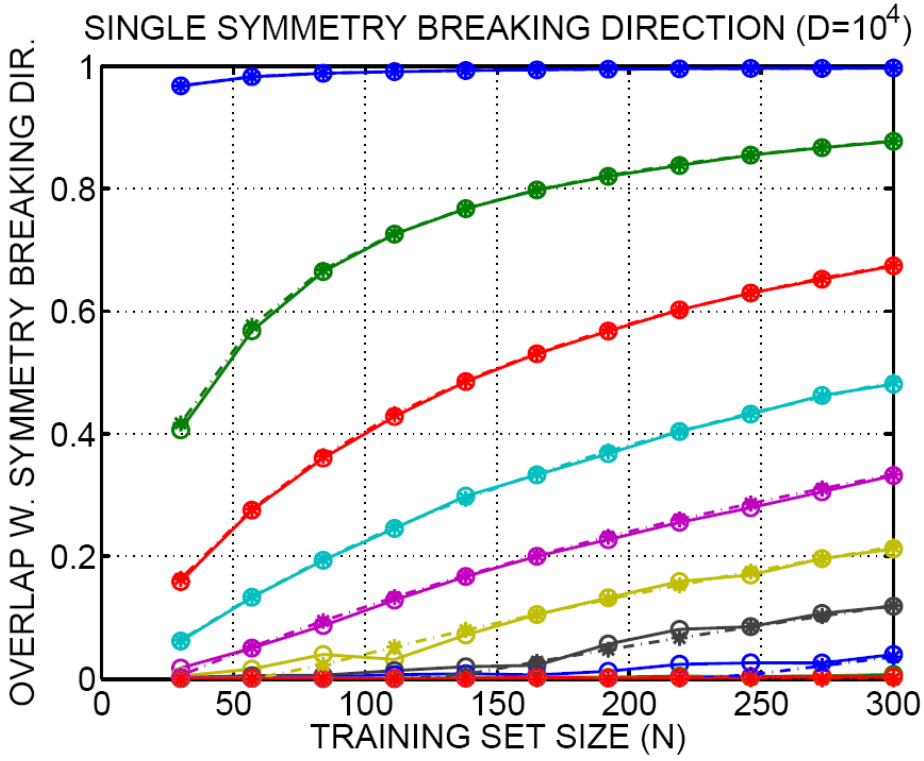
Universality

- 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)

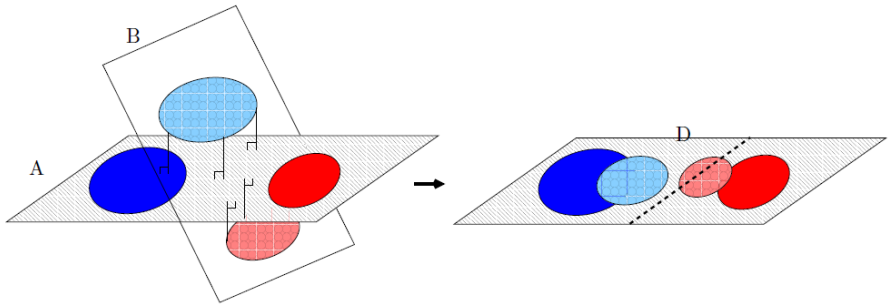
σ = (0.01, 0.06, 0.12, 0.17, 0.23, 0.28, 0.34, 0.39, 0.45, 0.5)

Variance inflation in PCA

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

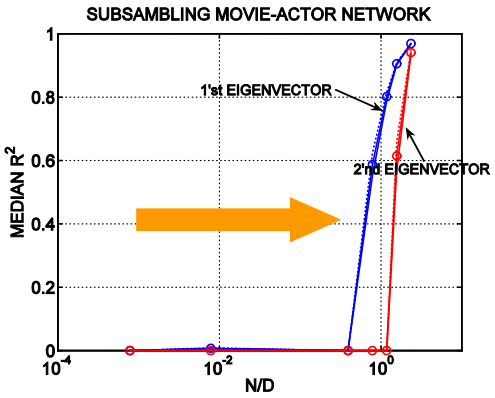
A Cure for Variance Inflation in High Dimensional Kernel Principal Component Analysis

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



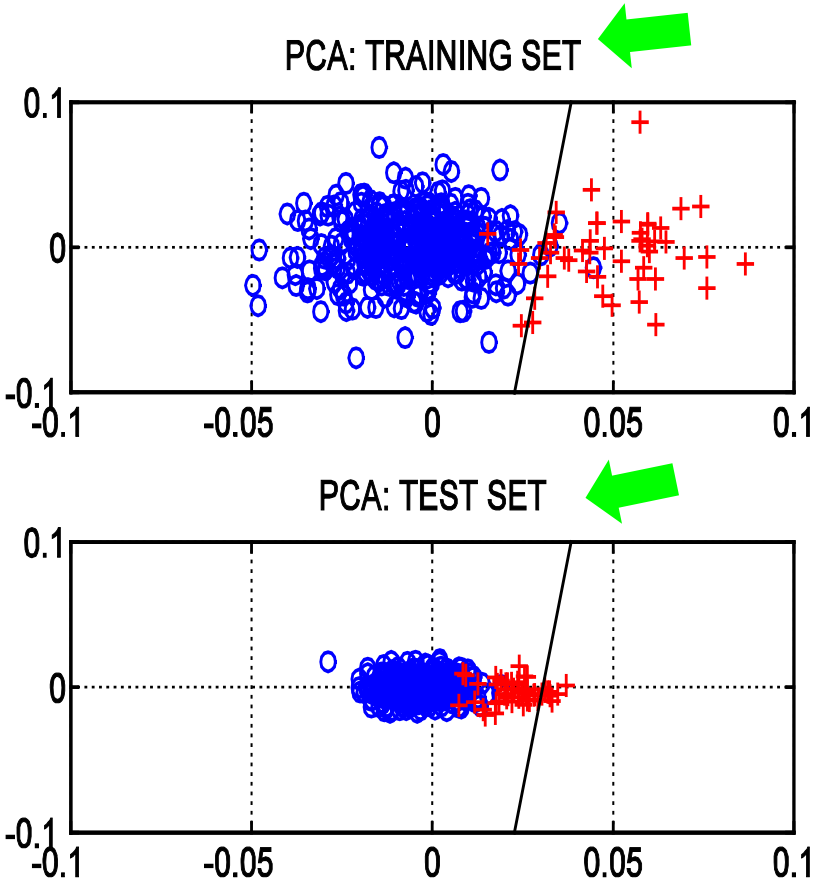
Restoring the generalizability of SVD

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

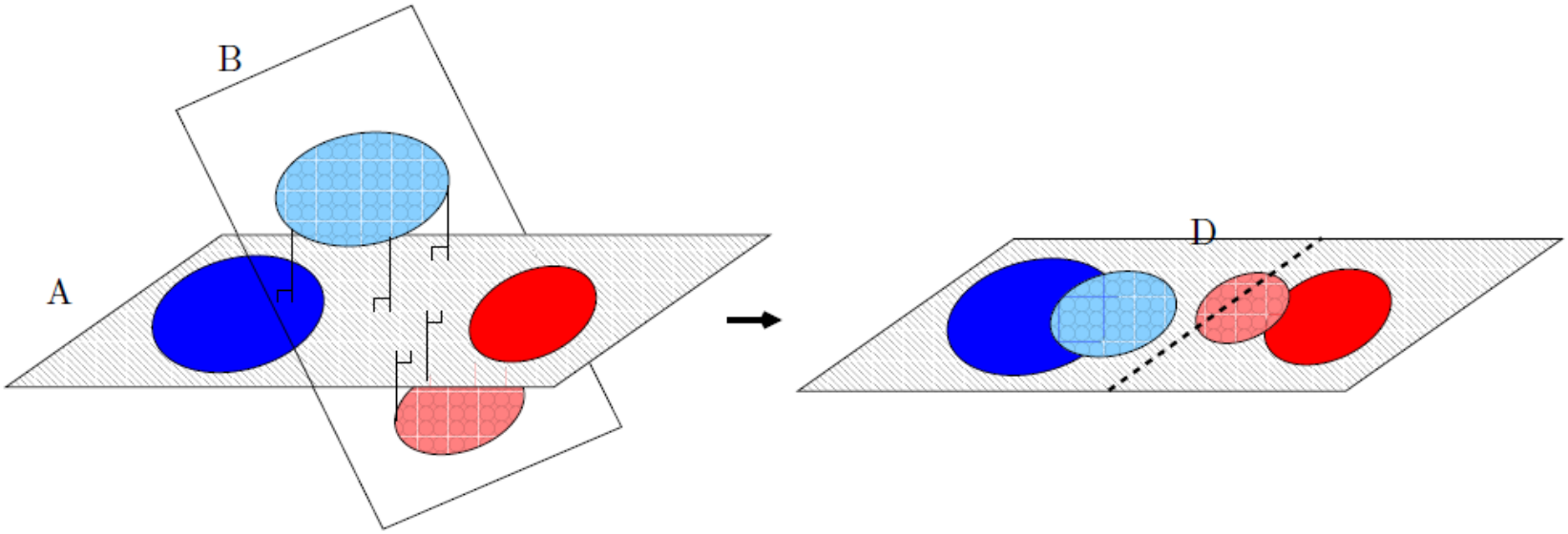


- The estimated projection is offset, hence, future projections will be too small!

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



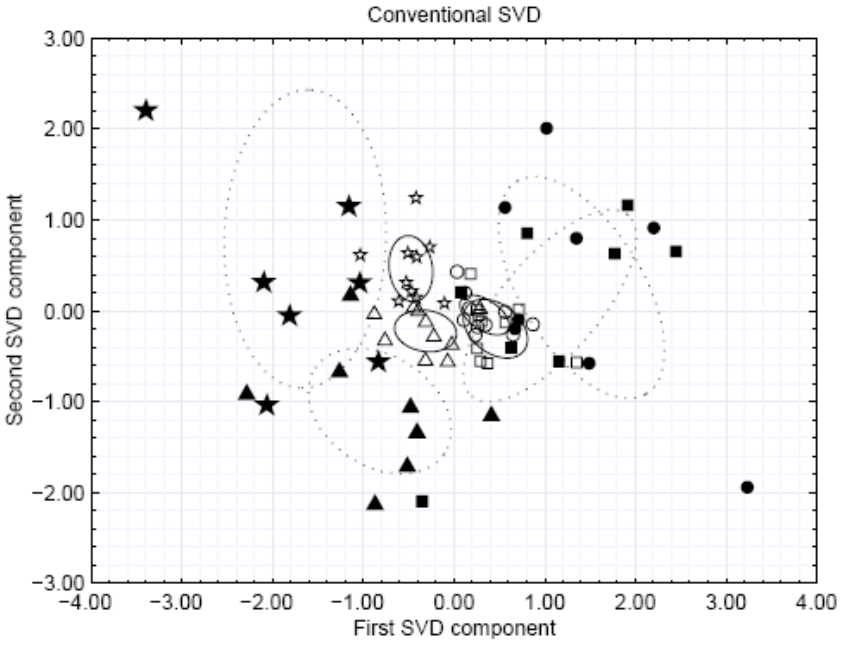
Variance inflation in PCA

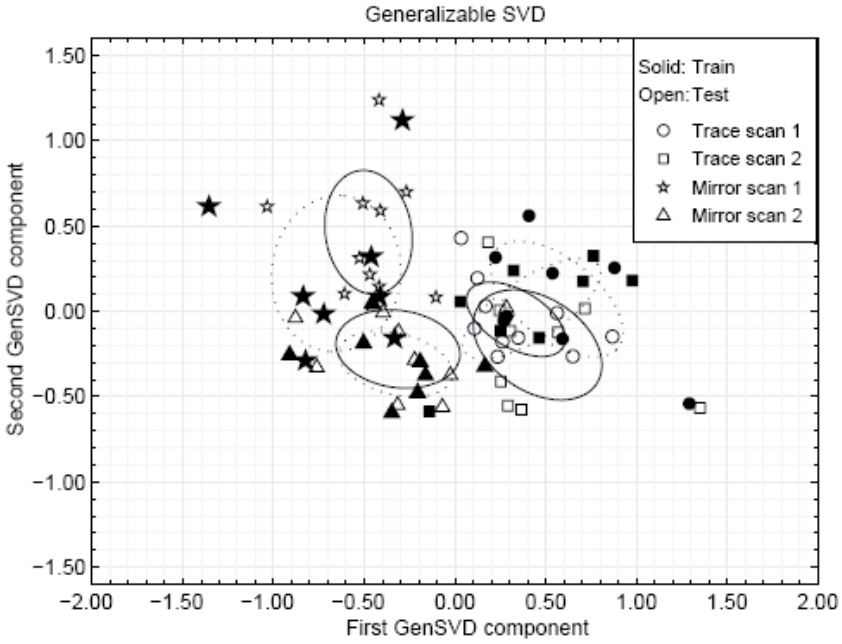


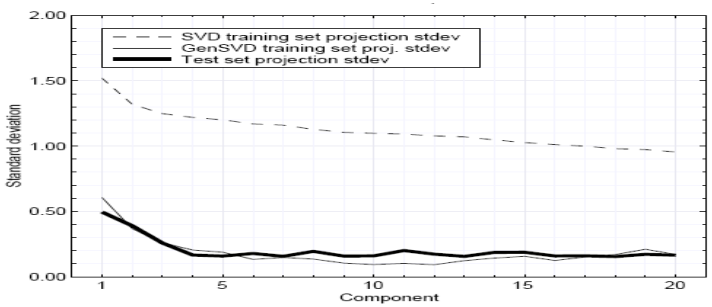
Who shrunk the test set?

Lars Kai Hansen
Technical University of Denmark

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





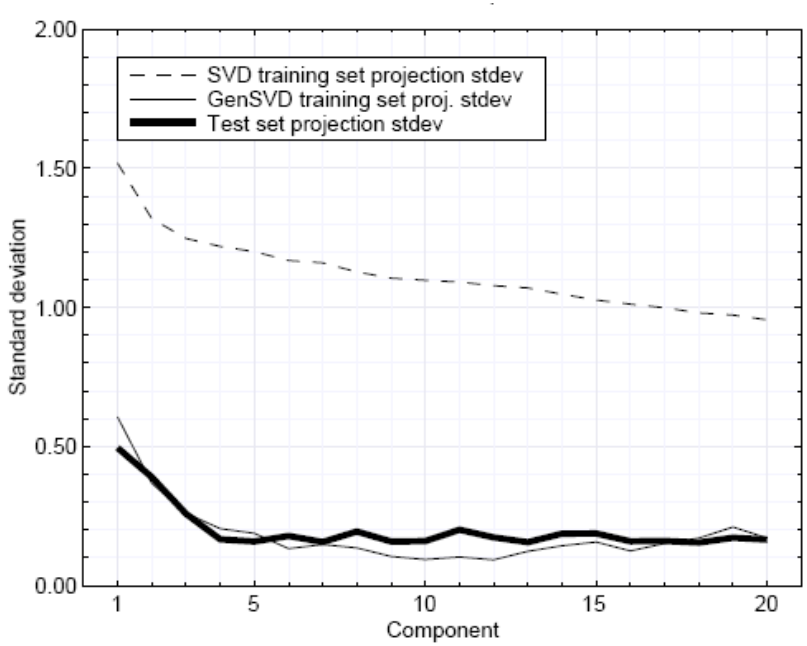


N=72, D=2.5 10^4

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



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

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

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

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).

Two approximations

Adjusting for the mean overlap

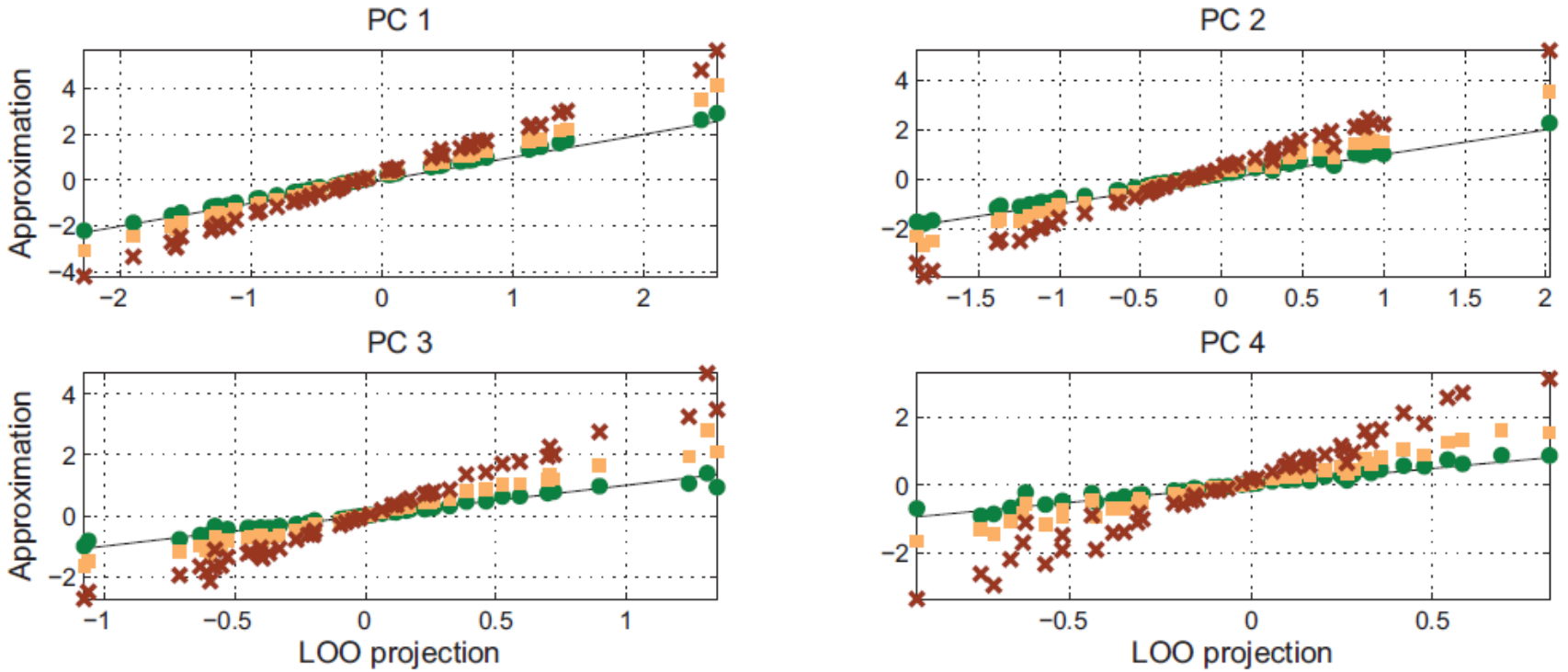
$$R^2 = \begin{cases} (\alpha S^2 - 1)/S(1+\alpha S) & \alpha > 1/S^2 \\ 0 & \alpha \le 1/S^2 \end{cases}$$

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

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

Adjusting for lost projection

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

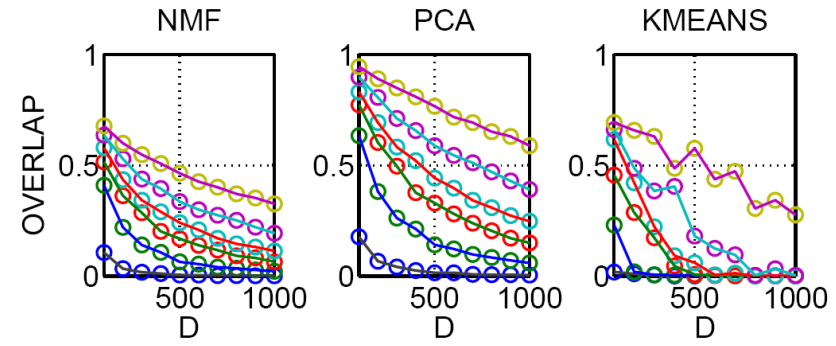


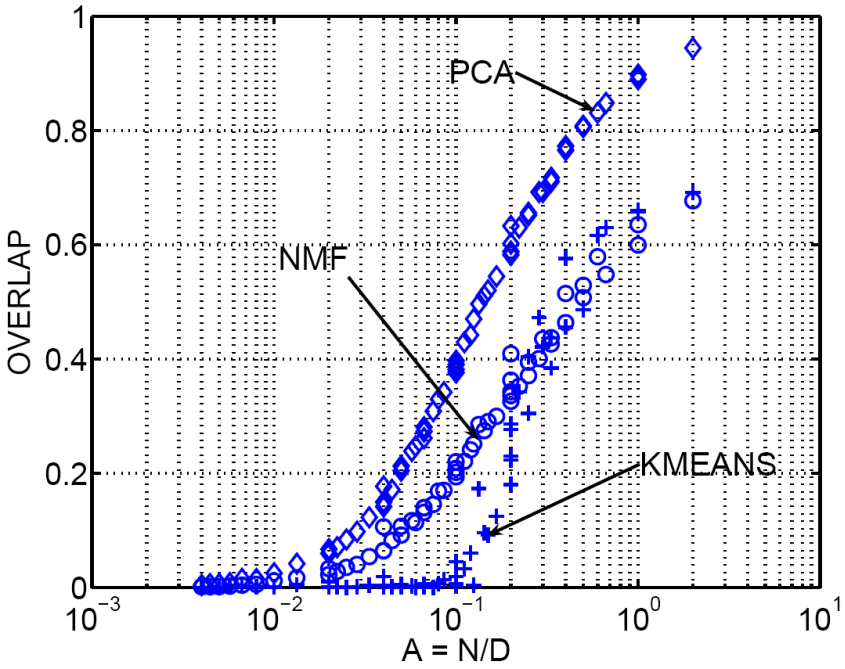
Approximating the leave-one-out (LOO) procedure. Here we simulate data with four normal independent signal components, $x = \Sigma_{k=1}^{4} \eta_k u_k + \epsilon$ of strengths $(1.4, 1.2, 1.0, 0.8,$ embedded in i.i.d. normal noise $\epsilon \sim N(0, \sigma^2 1)$, with $\sigma = 0.2$. The dimension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

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





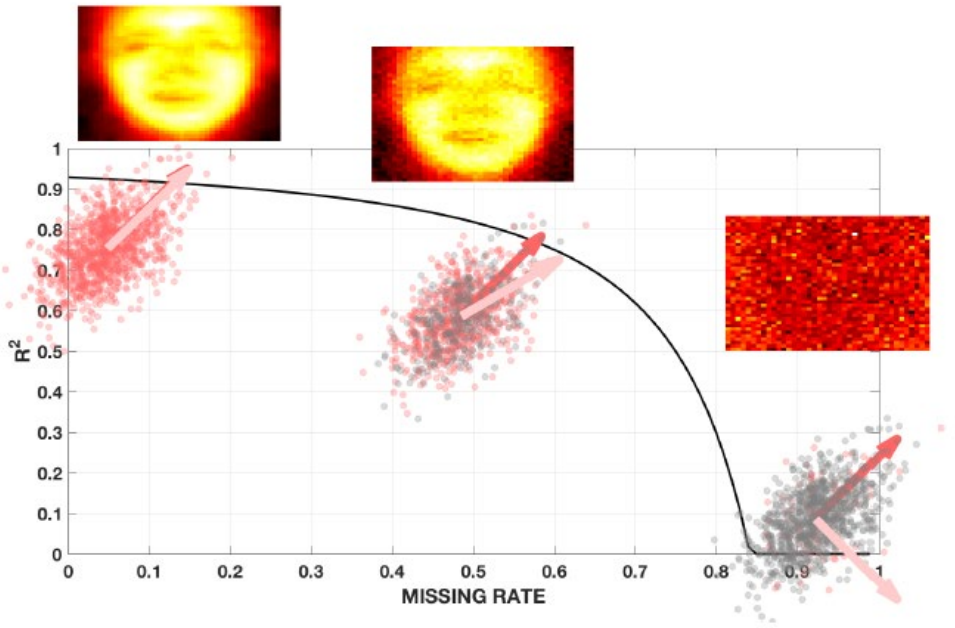
Lars Kai Hansen
Technical University of Denmark

Factor models for handling missing data

Phase transition in PCA with missing data:
Reduced signal-to-noise ratio, not sample size!

Niels Bruun Ipsen [1] **Lars Kai Hansen** [1]

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.



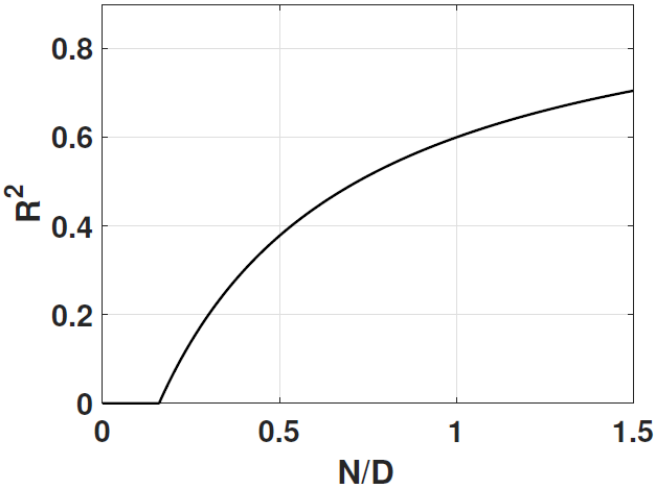
Schafer and Graham (2002): *'In missing data problems the sample may have to be larger than usual, because missing values effectively reduce the sample size*'.

Modeling – replica method

Universal learning curves for $R^2 = \langle (\mathbf{a}^T * \mathbf{a}_0)^2 \rangle$;

S is signal to noise, $\alpha = N/D$

$$\langle R^2 \rangle_{\mathcal{X}} = \begin{cases} 0 & \alpha S^2 < 1, \\ \frac{\alpha S^2 - 1}{S + \alpha S^2} & \alpha S^2 \geq 1 \end{cases}$$



Proof strategy:

1) $R^2 = \langle (\mathbf{a}^T * \mathbf{a}_0)^2 \rangle$ is obtained from a generating function (log Z)
2) Compute average of log Z, via moments using $\langle \log Z \rangle = \lim_{n \to 0} (\langle Z^n \rangle - 1)/n$
3) The average $\langle Z^n \rangle$ can be compute as N, D -> ∞, with α=N/D finite
4) Assume replica symmetry - all $\langle (\mathbf{a}_j^T * \mathbf{a}_k)^2 \rangle$ and $\langle (\mathbf{a}_j^T * \mathbf{a}_0)^2 \rangle$ all identical

Hoyle, D. C., & Rattray, M. (2007)
Statistical mechanics of learning multiple orthogonal signals: Asymptotic theory and fluctuation effects.
Physical Review E 75.1 (2007): 016101.

Biehl, M., & Mietzner, A. (1993)
Statistical mechanics of unsupervised learning.
EPL (Europhysics Letters) 24.5 (1993): 421.

Modeling – replica method

Universal learning curves for $R^2 = \langle(\mathbf{a}^T * \mathbf{a}_0)^2\rangle$;
S is signal to noise, $\alpha = N/D$

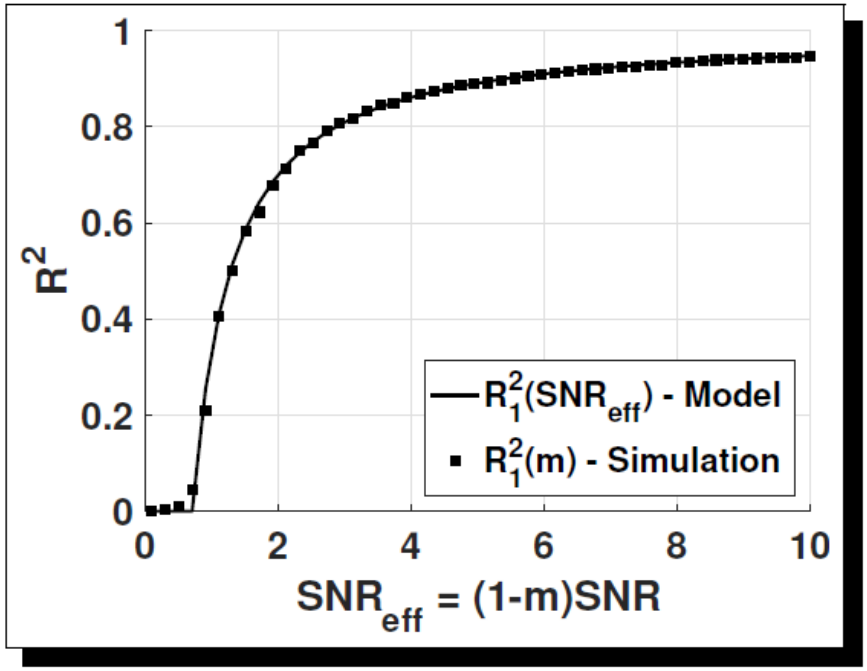
$$\langle R^2 \rangle_{\mathcal{X}} = \begin{cases} 0 & \alpha S^2 < 1, \\ \frac{\alpha S^2 - 1}{S + \alpha S^2} & \alpha S^2 \geq 1 \end{cases}$$

Define effective signal to noise $S(m) = (1 - m)S$

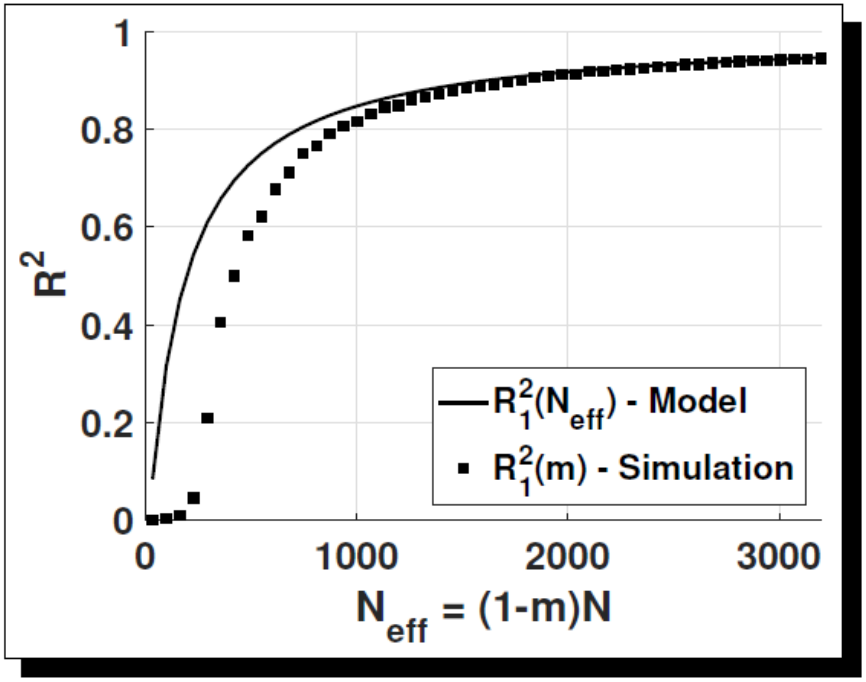
$$\langle R^2 \rangle_{\mathcal{X}} = \begin{cases} 0 & \alpha S(m)^2 < 1, \\ \frac{\alpha S(m)^2 - 1}{S(m) + \alpha S(m)^2} & \alpha S(m)^2 \geq 1. \end{cases}$$

Modeling the effect of missing data

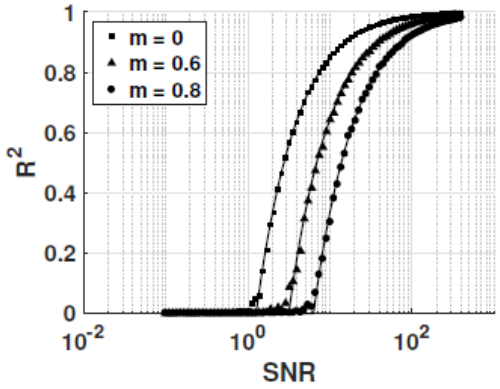
Effective SNR



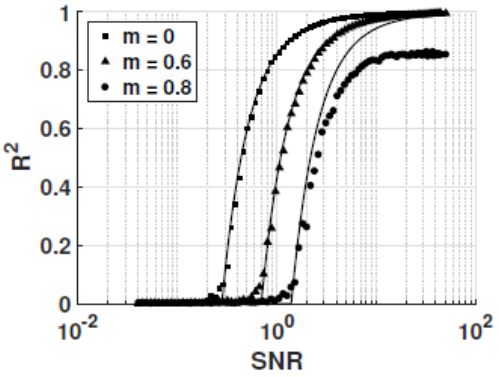
Effective sample size?



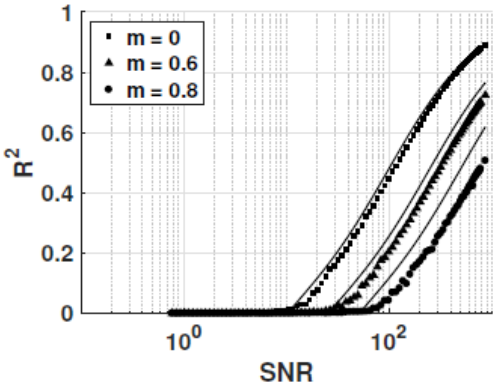
Real world data



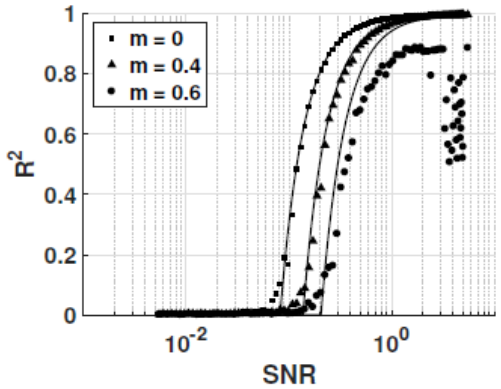
(a) Wild Faces



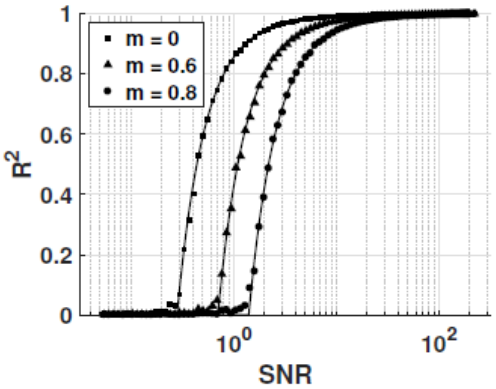
(b) MNIST



(c) NCI60



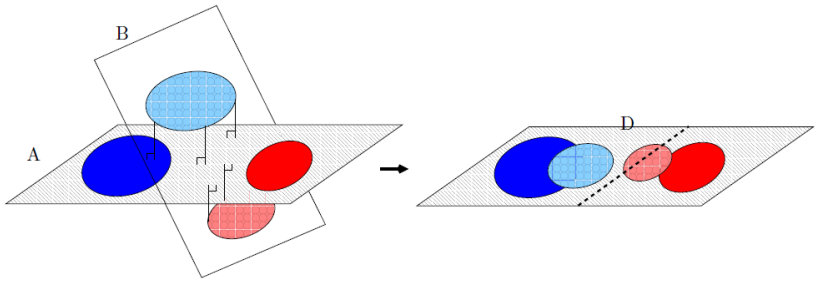
(d) Food Pairing



(e) Fashion

Conclusions

PCA is tricky
- variance inflation



Missing data can be handled with PPCA
Phase transition is found – similar to PCA learning curve
very accurate model of simulated data,
& reasonable agreement for real data
Missing data reduces SNR not samples size

Acknowledgement
Innovation Foundation Denmark -DABAI
Danish Research Councils