

Decomposition methods for Unsupervised Learning

02582

Morten Mørup

Supervised and Unsupervised learning





Supervised Learning

- Input and output data given, task: $p(y | x)$
i.e. Linear Regression predict output y from input x .

Unsupervised learning

- Only input data given, task: $p(x)$

Today's lecture

- A Short Introduction to Unsupervised learning and Factor analysis
- Non-negative Matrix Factorization 
- Archetypal Analysis 
- Independent Component Analysis 
- Sparse Coding 

Goal of unsupervised Learning

(Ghahramani & Roweis, 1999)



- Perform dimensionality reduction
- Build topographic maps
- Find the hidden causes or sources of the data
- Model the data density
- Cluster data

Purpose of unsupervised learning


(Hinton and Sejnowski, 1999)



- Extract an efficient internal representation of the statistical structure implicit in the inputs

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson  06.23.08

THE PETABYTE AGE:

Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't

Analysis of massive amounts of data will be the main driving force of all sciences in the future!!

Decomposition Methods for Unsupervised Learning.

- **Decomposition** - the process of finding hidden internal representation of the data, i.e., to decompose the data into its internal representations.
- **Guiding Principle** - simplicity of the representation.



William of Ockham
(1288-1347)

"lex parsimoniae"/ "law of parsimony" : The explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory.

All other things being equal, the simplest solution is the best!

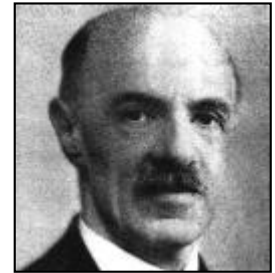
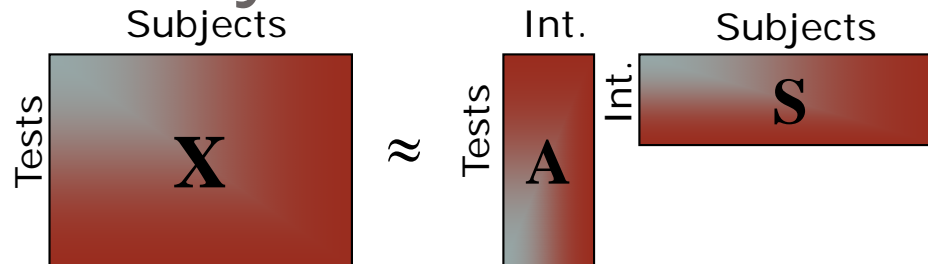
Redundancy Reduction



Horace Barlow
(1921-)

"What is a pattern? It is some kind of **regularity** or **self-similarity** in a signal or set of data. If there is no regularity, or no repetition caused by self-similarity, then surely there is no pattern. But if there is such regularity or repetition, then **this is a form of redundancy, and offers the opportunity for recoding to reduce it.** Of course the pattern element can be completely arbitrary, a sequence of randomly selected digits for example, but if repeated this element will make a pattern. Thus it seems to me **that the importance of redundancy is almost a tautology and follows simply from the nature of pattern.**"

Factor Analysis

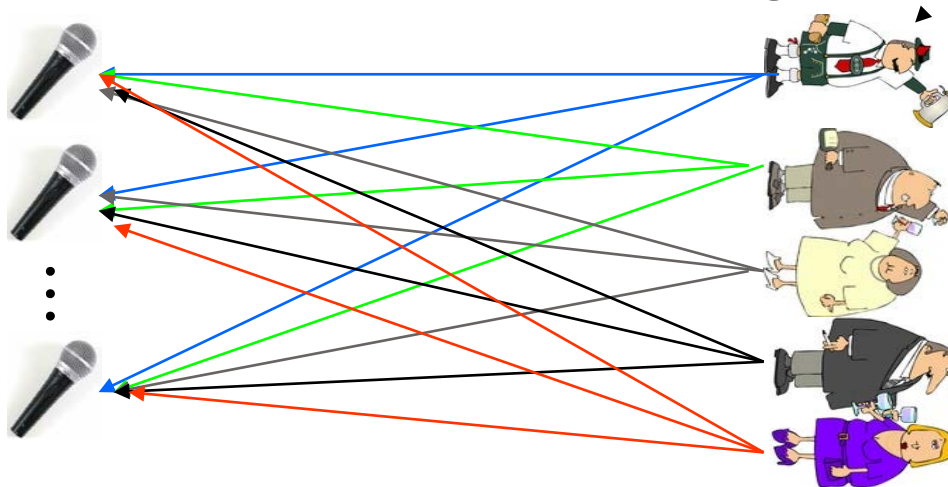


Spearman ~1900

$$\mathbf{X}_{\text{tests} \times \text{subjects}} \approx \mathbf{A}_{\text{tests} \times \text{int.}} \mathbf{S}_{\text{int.} \times \text{subject}}$$

The Cocktail Party problem

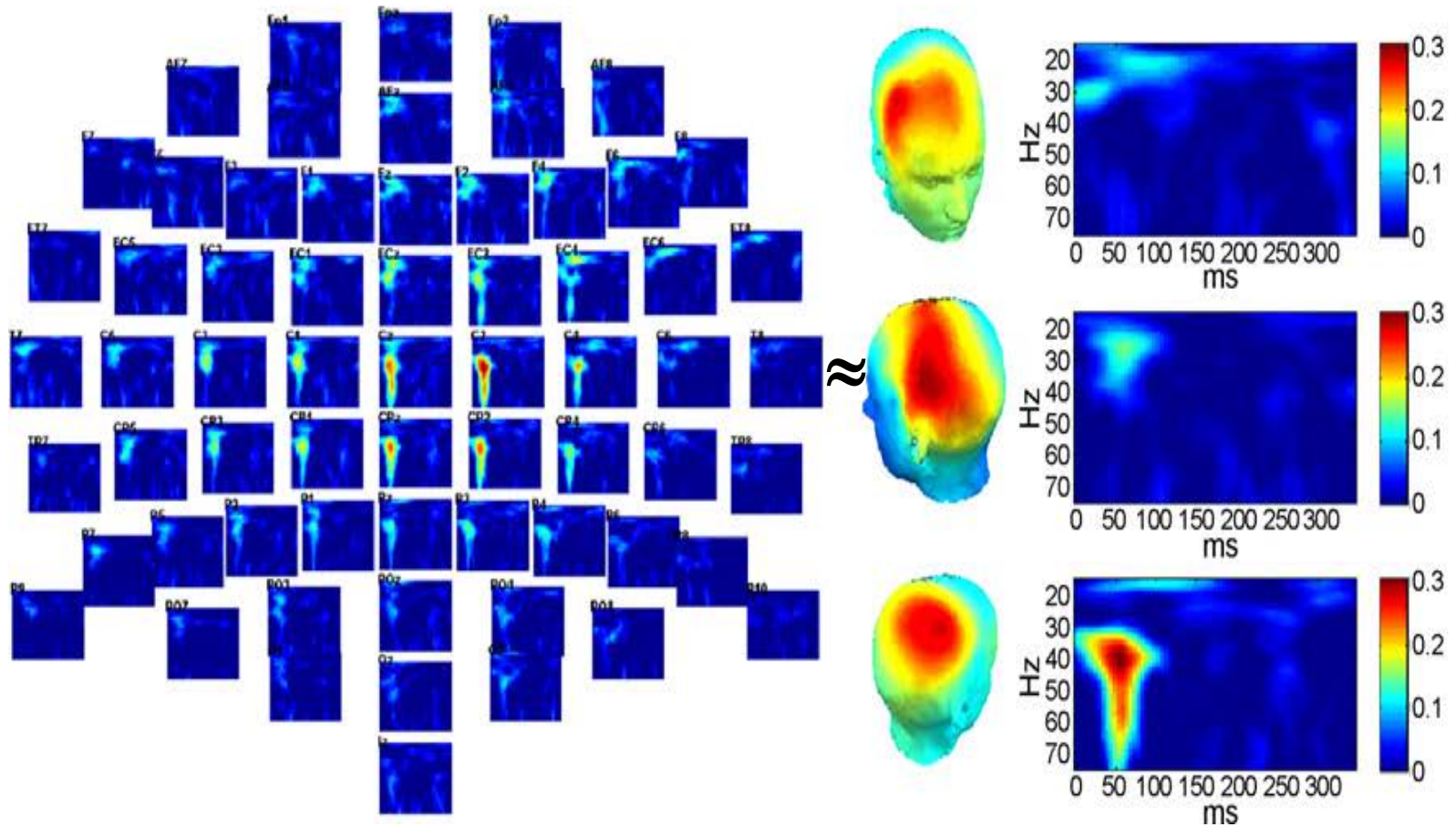
$$\mathbf{X}_{\text{microphones} \times \text{time}} \approx \mathbf{A}_{\text{microphones} \times \text{people}} \mathbf{S}_{\text{people} \times \text{time}}$$



Can't we just solve the cocktail party problem by PCA/SVD?

Flaw: $\mathbf{X} \approx \mathbf{A}\mathbf{S} = (\mathbf{A}\mathbf{Q}^{-1})(\mathbf{Q}\mathbf{S}) = \hat{\mathbf{A}}\hat{\mathbf{S}} \Rightarrow \text{Representation not unique!}$

NeuroImaging example: Time-Frequency transformed EEG



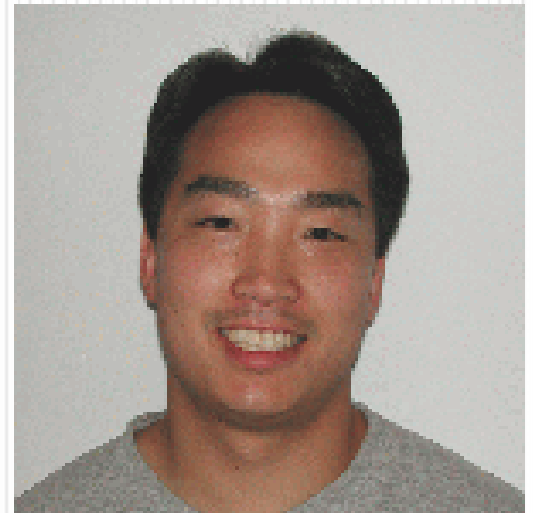
Non-negative Matrix Factorization



Pentti Paatero



Sebastian Seung



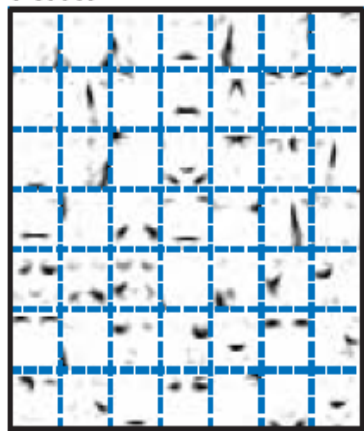
Daniel D. Lee

NMF gives part based representation

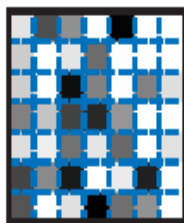
Original



NMF



×



=

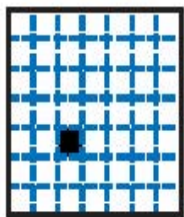


$$\mathbf{X} = \mathbf{W}\mathbf{H} + \mathbf{E},$$
$$\mathbf{W} \geq 0, \quad \mathbf{H} \geq 0$$

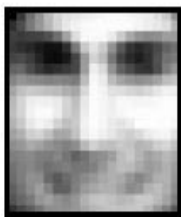
VQ (K-means)



×



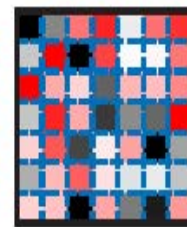
=



PCA



×

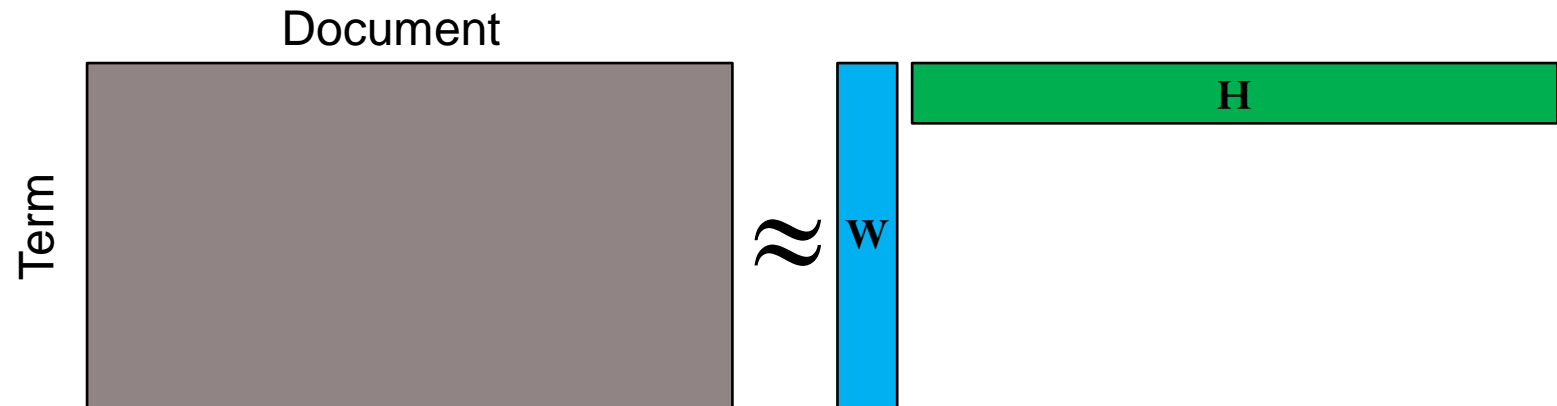



=



(Lee and Seung, Nature 1999)

NMF for text information retrieval



court government council culture supreme constitutional rights justice	president served governor secretary senate congress presidential elected	\times  \approx	Encyclopedia entry: 'Constitution of the United States' president (148) congress (124) power (120) united (104) constitution (81) amendment (71) government (57) law (49)
flowers leaves plant perennial flower plants growing annual	disease behaviour glands contact symptoms skin pain infection		

metal process method paper ... glass copper lead steel
person example time people ... rules lead leads law

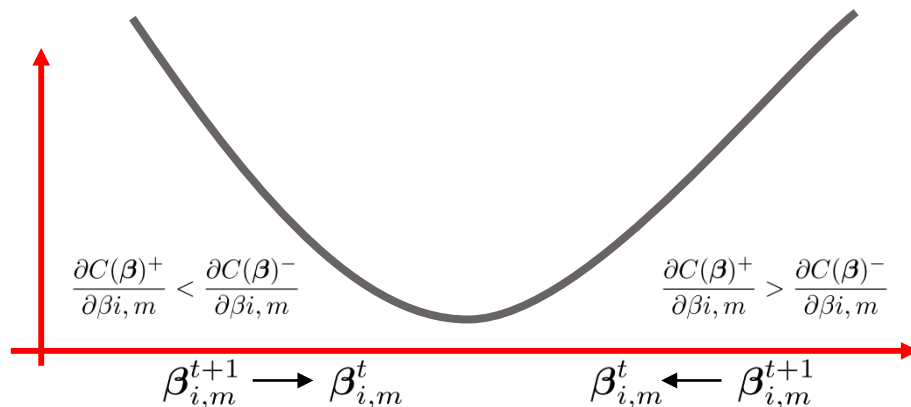
Multiplicative updates for NMF

$$\frac{\partial C(\boldsymbol{\beta})}{\partial \beta_{i,m}} = \frac{\partial C(\boldsymbol{\beta})^+}{\partial \beta_{i,m}} - \frac{\partial C(\boldsymbol{\beta})^-}{\partial \beta_{i,m}}$$

$$\beta_{i,m} = \beta_{i,m} - \mu_{i,m} \frac{\partial C(\boldsymbol{\beta})}{\partial \beta_{i,m}}, \quad \mu_{i,m} = \frac{\beta_{i,m}}{\frac{\partial C(\boldsymbol{\beta})^+}{\partial \beta_{i,m}}}$$

$$\beta_{i,m} = \beta_{i,m} - \frac{\beta_{i,m}}{\frac{\partial C(\boldsymbol{\beta})^+}{\partial \beta_{i,m}}} \left(\frac{\partial C(\boldsymbol{\beta})^+}{\partial \beta_{i,m}} - \frac{\partial C(\boldsymbol{\beta})^-}{\partial \beta_{i,m}} \right) = \beta_{i,m} \frac{\frac{\partial C(\boldsymbol{\beta})^-}{\partial \beta_{i,m}}}{\frac{\partial C(\boldsymbol{\beta})^+}{\partial \beta_{i,m}}}$$

$$\beta_{i,m}^{t+1} \leftarrow \beta_{i,m}^t \left(\frac{\frac{\partial C(\boldsymbol{\beta})^-}{\partial \beta_{i,m}^t}}{\frac{\partial C(\boldsymbol{\beta})^+}{\partial \beta_{i,m}^t}} \right)$$



Multiplicative updates for Least squares error and KL-divergence (Poisson noise)

$$X_{i,j} \geq 0 \quad , \quad w_{i,d} \geq 0 \quad \text{and} \quad h_{d,j} \geq 0$$

$$C_{LS} = \frac{1}{2} \|\mathbf{X} - \mathbf{WH}\|_F^2 = \frac{1}{2} \sum_{i,j} (X_{i,j} - (\mathbf{WH})_{i,j})^2$$



$$w_{i,d} \leftarrow w_{i,d} \frac{(\mathbf{XH}^T)_{i,d}}{(\mathbf{WHH}^T)_{i,d}}$$

$$h_{d,j} \leftarrow h_{d,j} \frac{(\mathbf{W}^T \mathbf{X})_{d,j}}{(\mathbf{W}^T \mathbf{WH})_{d,j}}$$

Can you spot any potential issues that needs to be taken care of with these beautiful updates?

$$C_{KL} = \sum_{i,j} x_{i,j} \log \frac{x_{i,j}}{(\mathbf{WH})_{i,j}} - x_{i,j} + (\mathbf{WH})_{i,j}$$

$$w_{i,d} \leftarrow w_{i,d} \frac{\sum_j \frac{x_{i,j}}{(\mathbf{WH})_{i,j}} h_{d,j}}{\sum_j h_{d,j}}$$

$$h_{d,j} \leftarrow h_{d,j} \frac{\sum_i w_{i,d} \frac{x_{i,j}}{(\mathbf{WH})_{i,j}}}{\sum_i w_{i,d}}$$

(Lee and Seung, Nature 1999 and NIPS 2001)

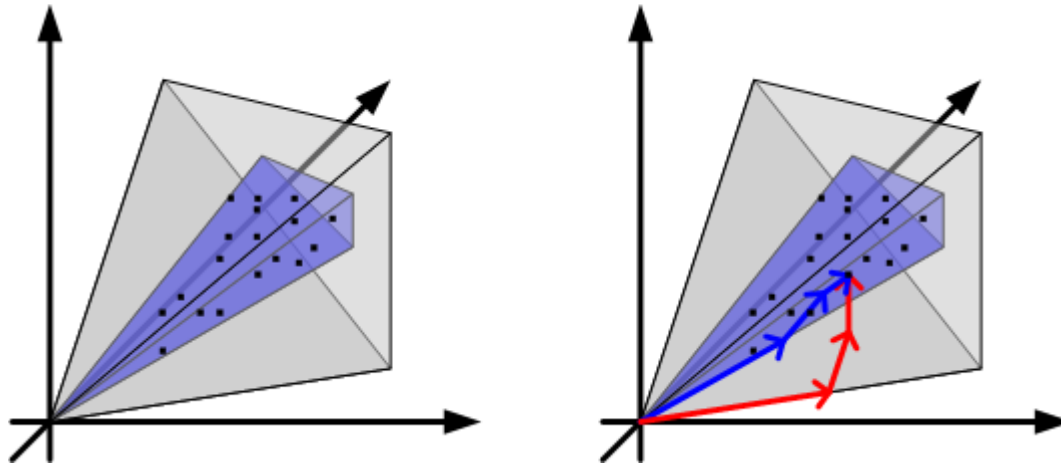
Alternative inference by projected gradient

$$\begin{aligned} \mathbf{V}_{i,j} &\geq 0 \quad , \quad \mathbf{W}_{i,d} \geq 0 \quad \text{and} \quad \mathbf{H}_{d,j} \geq 0 \\ C_{LS}(\mathbf{W}, \mathbf{H}) &= \frac{1}{2} \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2 = \frac{1}{2} \sum_{i,j} (\mathbf{V}_{i,j} - (\mathbf{W}\mathbf{H})_{i,j})^2 \\ g_{i,d}^W &= \frac{\partial C_{LS}}{\partial w_{i,d}} \\ \mathbf{W}^{new} &\leftarrow P_+(\mathbf{W}^{old} - \mu \mathbf{G}^W) \\ g_{d,j}^H &= \frac{\partial C_{LS}}{\partial h_{d,j}} \\ \mathbf{H}^{new} &\leftarrow P_+(\mathbf{H}^{old} - \mu \mathbf{G}^H) \end{aligned}$$

P_+ projects the data to the positive orthant by setting negative values to zero.
 μ_W and μ_H chosen such that $C_{LS}(\mathbf{W}^{new}, \mathbf{H}^{old}) < C_{LS}(\mathbf{W}^{old}, \mathbf{H}^{old})$
and $C_{LS}(\mathbf{W}^{new}, \mathbf{H}^{new}) < C_{LS}(\mathbf{W}^{new}, \mathbf{H}^{old})$

NMF is in general not unique

$$\mathbf{X} \approx \mathbf{W}\mathbf{H} = (\mathbf{W}\mathbf{Q}^{-1})(\mathbf{Q}\mathbf{H}) = \mathbf{W}\mathbf{H} \Rightarrow \mathbf{W} \geq \mathbf{0}, \mathbf{H} \geq \mathbf{0}$$



(Donoho & Stodden 2003, Laurberg et al. 2008)



How can we disambiguate the solutions?

Archetypal Analysis



Leo Breiman



Adele Cutler

$$\mathbf{X} \approx \mathbf{X} \mathbf{S} \mathbf{H} + \mathbf{E},$$

\approx

\mathbf{S}

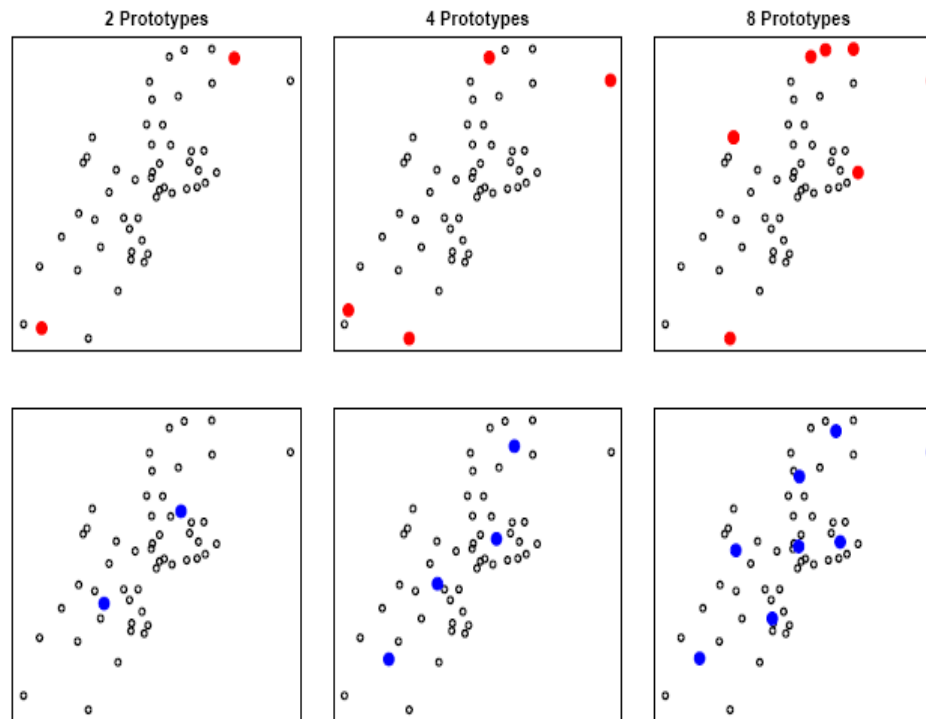
 \mathbf{H}

$$|\mathbf{S}_d|_1 = 1, \quad |\mathbf{h}_j|_1 = 1,$$

$$\mathbf{S} \geq 0, \quad \mathbf{H} \geq 0$$

Prototype

$$\mathbf{w}_d = \mathbf{X} \mathbf{s}_d$$



Blue dots: k-means prototypes, red dots: Archetypal Analysis prototypes

Inference in the AA model

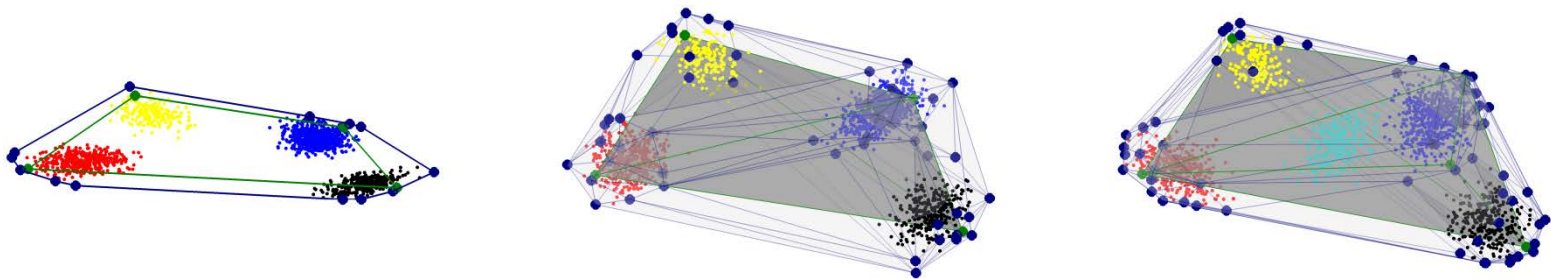
- AA can be solved by the following simple iterative procedure:

- Change of variable

$$\begin{aligned}\mathbf{X} &= \mathbf{X}\tilde{\mathbf{S}}\tilde{\mathbf{H}} + \mathbf{E}, \\ \tilde{s}_{j,d} &= \frac{s_{j,d}}{\sum_j s_{j,d}} \\ \tilde{h}_{d,j} &= \frac{h_{d,j}}{\sum_d h_{d,j}}\end{aligned}$$

- Solve alternately for \mathbf{S} and \mathbf{H} using projected gradient

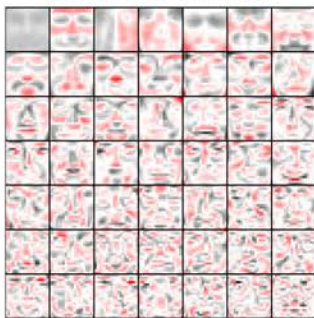
Archetypal analysis extracts the dominant convex hull of the data cloud



Convex hull: Blue lines and light shaded region

Dominant convex hull: green lines and gray shaded region

SVD/PCA 99.44%



Low -> high frequency content

NMF 99.26+/- 0.01 %



Part based representation

AA 97.72+/- 0.02 %



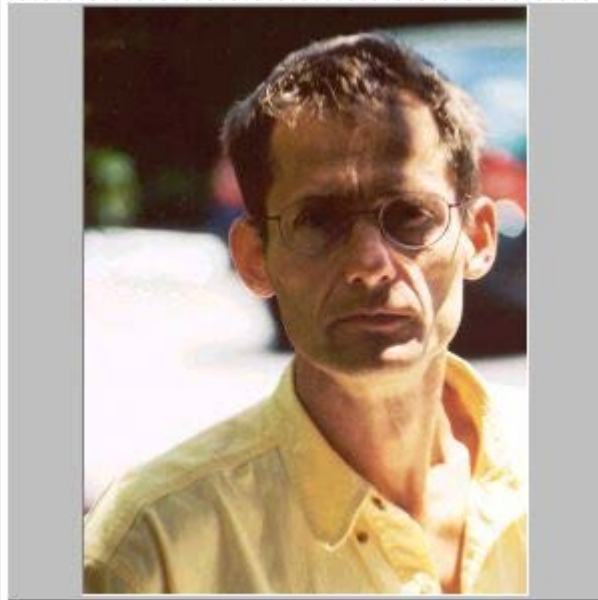
Archetypes/Freaks

k-means 96.33+/- 0.02 %



Exemplars

Independent Component Analysis



Pierre Comon

ICA is a modern approach to the rotational ambiguity (Q)

$$\mathbf{X} = \mathbf{A}\mathbf{S} = (\mathbf{A}\mathbf{Q}^{-1})(\mathbf{Q}\mathbf{S}) = \hat{\mathbf{A}}\hat{\mathbf{S}}$$

We can assume \mathbf{X} is pre-whitened such that $\mathbf{X}\mathbf{X}^T = \mathbf{I}$

if this is not the case perform SVD

$$[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^T] = \text{SVD}(\mathbf{X})$$

And perform ICA on $\mathbf{Y} = \mathbf{V}$

$$[\tilde{\mathbf{A}}, \mathbf{S}] = \text{ICA}(\mathbf{Y}) \text{ then } \mathbf{X} = \mathbf{U}\mathbf{\Sigma}\tilde{\mathbf{A}}\mathbf{S} = \mathbf{A}\mathbf{S} \text{ where } \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\tilde{\mathbf{A}}$$

Independent implies that \mathbf{S} is uncorrelated (weaker condition):

$$\mathbf{S}\mathbf{S}^T = \mathbf{I}, \text{ but as } \mathbf{Y}\mathbf{Y}^T = \tilde{\mathbf{A}}\mathbf{S}\mathbf{S}^T\tilde{\mathbf{A}}^T = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{I}$$

Thus ICA amounts to solving for an orthonormal matrix $\tilde{\mathbf{A}}$ such that $\mathbf{S} = \tilde{\mathbf{A}}^T \mathbf{Y}$ (are independent (and non-Gaussian).)



Is $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\tilde{\mathbf{A}}$ also orthogonal?

Many of the popular approaches to ICA are based on entropy. The differential entropy H of a random variable Y with density $g(y)$ is given by

$$H(Y) = - \int g(y) \log g(y) dy. \quad (14.82)$$

A well-known result in information theory says that among all random variables with equal variance, Gaussian variables have the maximum entropy. Finally, the *mutual information* $I(Y)$ between the components of the random vector Y is a natural measure of dependence:

$$I(Y) = \sum_{j=1}^p H(Y_j) - H(Y). \quad (14.83)$$

For convenience, rather than using the entropy $H(Y_j)$, Hyvärinen and Oja (2000) use the *negentropy* measure $J(Y_j)$ defined by

$$J(Y_j) = H(Z_j) - H(Y_j), \quad (14.86)$$

where Z_j is a Gaussian random variable with the same variance as Y_j . Negentropy is non-negative, and measures the departure of Y_j from Gaussianity. They propose simple approximations to negentropy which can be computed and optimized on data. The ICA solutions shown in Figures 14.37–14.39 use the approximation

$$J(Y_j) \approx [EG(Y_j) - EG(Z_j)]^2, \quad (14.87)$$

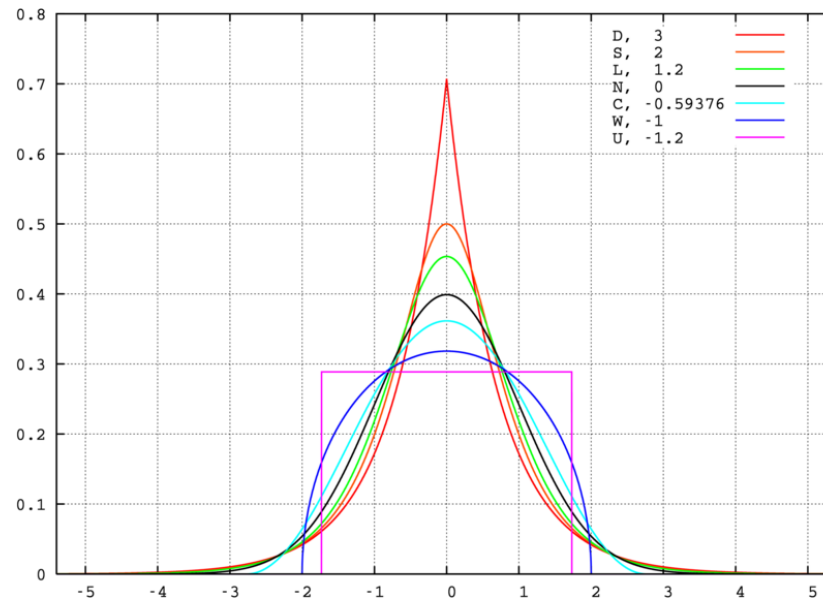
where $G(u) = \frac{1}{a} \log \cosh(au)$ for $1 \leq a \leq 2$. When applied to a sample of x_i , the expectations are replaced by data averages. This is one of the options in the **FastICA** software provided by these authors. More classical (and less robust) measures are based on fourth moments, and hence look for departures from the Gaussian via kurtosis. See Hyvärinen and Oja (2000)

In summary, ICA applied to multivariate data looks for a sequence of orthogonal projections such that the projected data look as far from Gaussian as possible. With pre-whitened data, this amounts to looking for components that are as independent as possible.

Kurtosis

Kurtosis is defined as the fourth cumulant divided by the square of the second cumulant, which is equal to the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3,

$$\gamma_2 = \frac{\mu^4}{\sigma^4} - 3 = \frac{\frac{1}{n} \sum_i^n (x_n - \bar{x})^4}{\left(\frac{1}{n} \sum_i^n (x_n - \bar{x})^2\right)^2} - 3$$



D: Laplace distribution, a.k.a. double exponential distribution, red curve, excess kurtosis = 3

S: hyperbolic secant distribution, orange curve, excess kurtosis = 2

L: logistic distribution, green curve, excess kurtosis = 1.2

N: normal distribution, black curve, excess kurtosis = 0

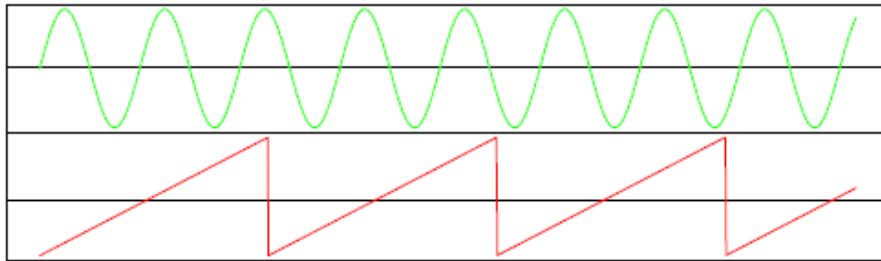
C: raised cosine distribution, cyan curve, excess kurtosis = -0.593762...

W: Wigner semicircle distribution, blue curve, excess kurtosis = -1

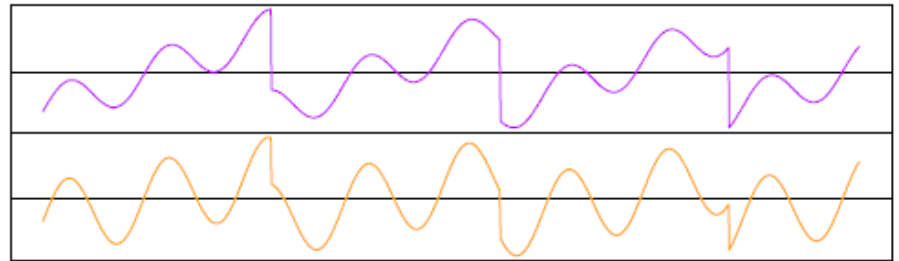
U: uniform distribution, magenta, excess kurtosis = -1.2.

ICA example

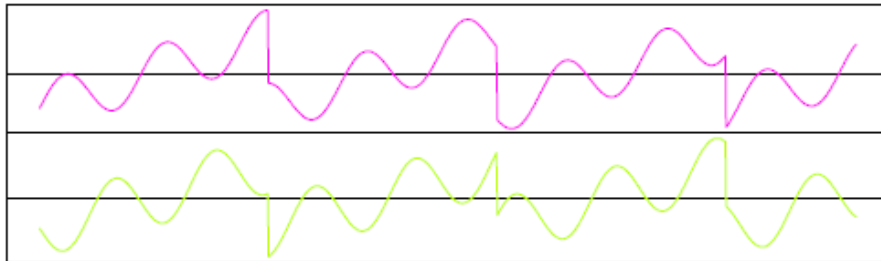
Source Signals



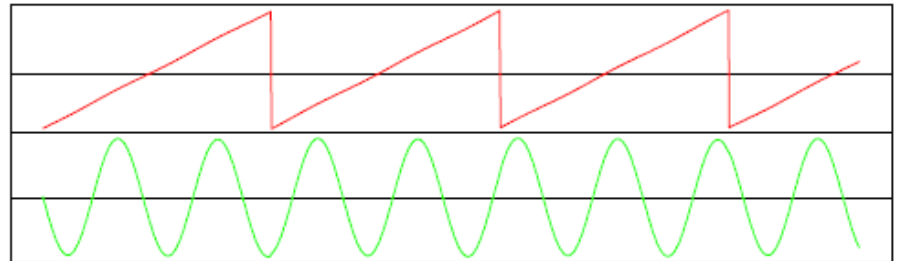
Measured Signals



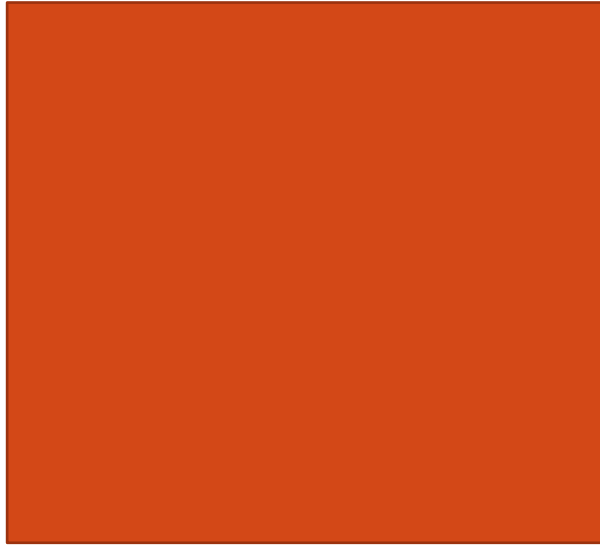
PCA Solution



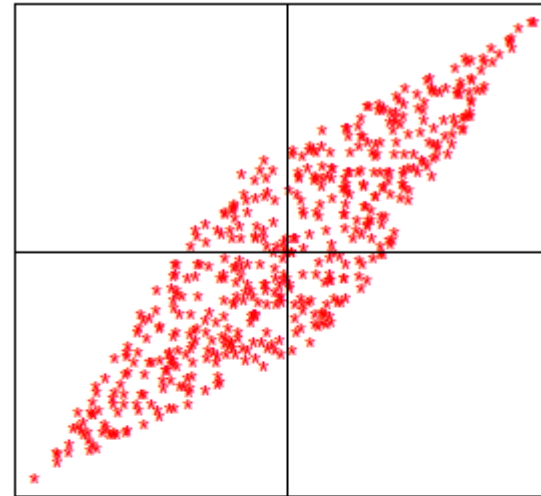
ICA Solution



Source S



Data X



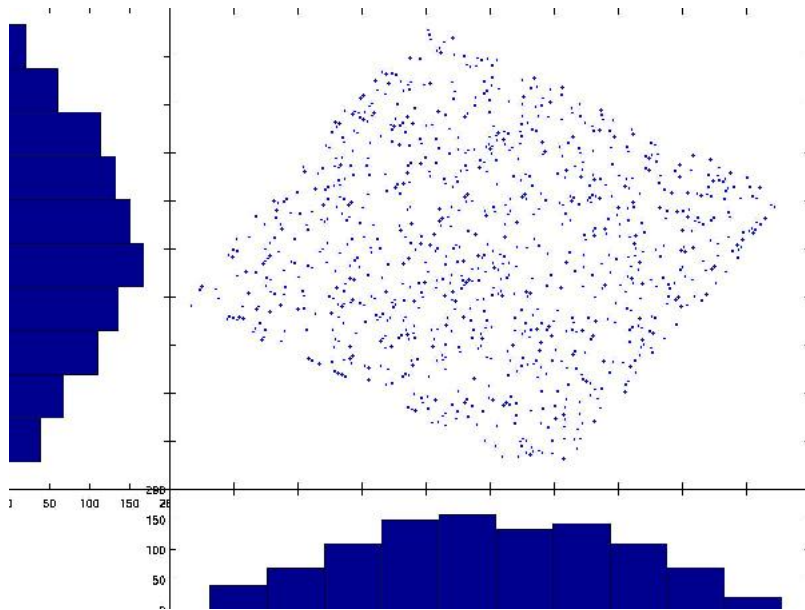
PCA Solution



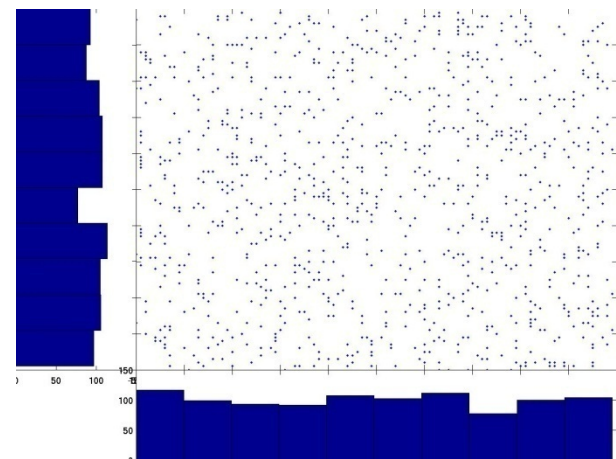
ICA Solution



What are the PCA and ICA directions



Somewhat gaussian looking



Strongly deviating from Gaussian distribution

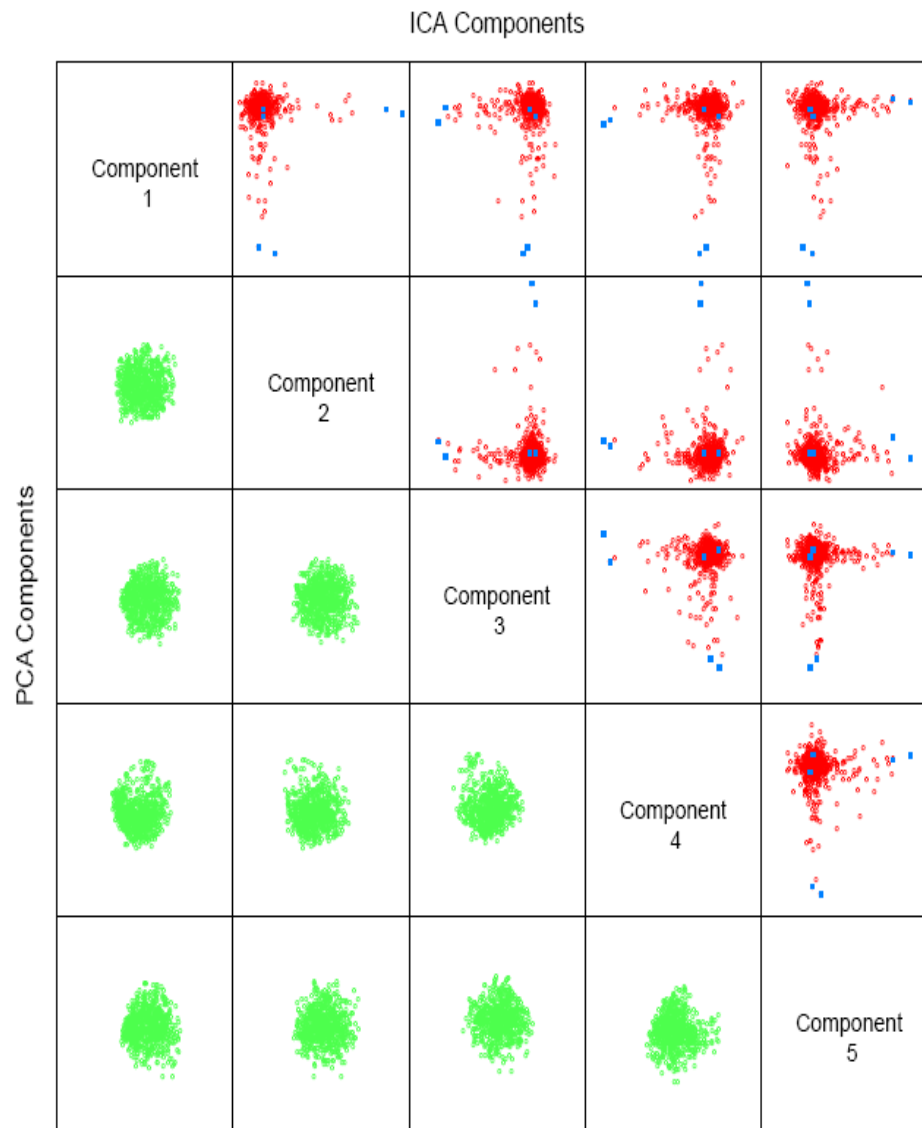
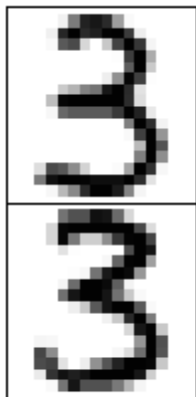
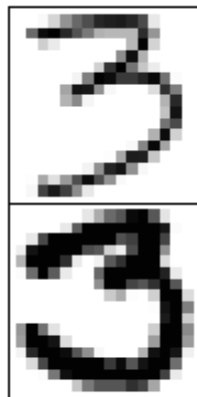


FIGURE 14.39. A comparison of the first five ICA components computed using FastICA (above diagonal) with the first five PCA components (below diagonal). Each component is standardized to have unit variance.



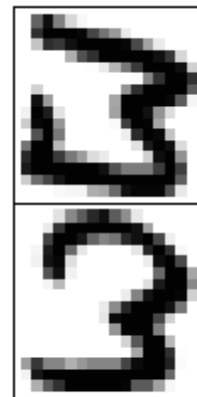
Mean



ICA 1



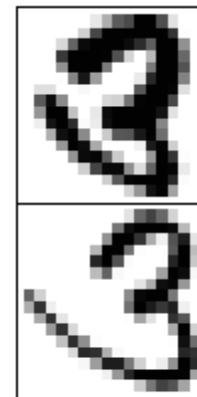
ICA 2



ICA 3



ICA 4



ICA 5

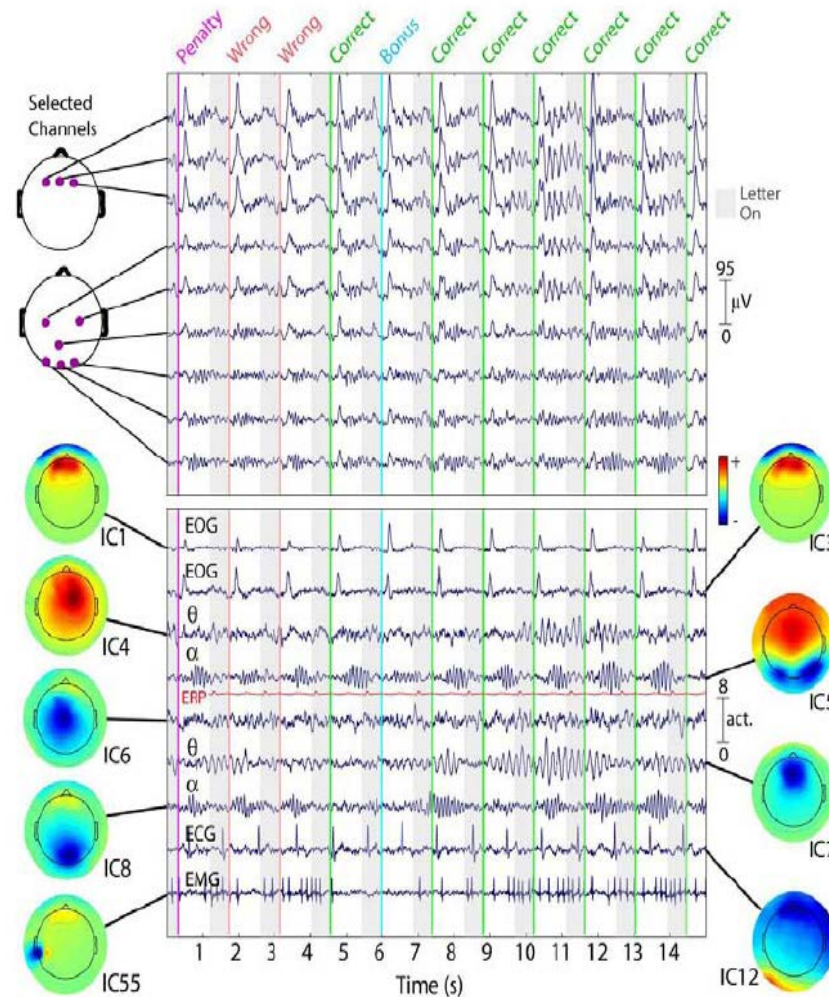


FIGURE 14.41. Fifteen seconds of EEG data (of 1917 seconds) at nine (of 100) scalp channels (top panel), as well as nine ICA components (lower panel). While nearby electrodes record nearly identical mixtures of brain and non-brain activity, ICA components are temporally distinct. The colored scalps represent the ICA unmixing coefficients \hat{a}_j as a heatmap, showing brain or scalp location of the source.

Sparse Coding (Dictionary Learning)



Bruno A. Olshausen



David J. Field

Sparse Coding

$$\operatorname{argmin}_{\mathbf{A}, \mathbf{S}} \underbrace{D(\mathbf{X}, \mathbf{AS})}_{\text{Preserve Information}} + \underbrace{\lambda \operatorname{sp}(\mathbf{S})}_{\text{Preserve Sparsity (Simplicity)}}$$

Tradeoff parameter

$$\operatorname{sp}(\mathbf{s}) = |\mathbf{s}|_\gamma = \sum_d |s_d|^\gamma$$

Ultimate measure of sparsity given by ℓ_0 norm (i.e., directly minimizes the number of non-zero entries). However, this results in NP-hard optimization! Instead the ℓ_1 norm is commonly invoked.

$$\operatorname{argmin}_{\mathbf{A}, \mathbf{S}} \frac{1}{2} \|\mathbf{X} - \mathbf{AS}\|_F^2 + \lambda |\mathbf{S}|_1$$

Solving \mathbf{S} for fixed \mathbf{A} correspond to a series of sparse regression problems

$$\operatorname{argmin}_{\mathbf{s}_j} \frac{1}{2} \|\mathbf{x}_j - \mathbf{A}\mathbf{s}_j\|_F^2 + \lambda |\mathbf{s}_j|_1$$

c.f. Lasso in Lecture 5, solvable for instance by LARS, QP or BPD

Why use the ℓ_1 norm and not for instance $\ell_{0.5}$ or $\ell_{0.0000001}$?



ℓ_1 convex proxy for ℓ_0 !

$$\operatorname{argmin}_{\mathbf{s}} \frac{1}{2} \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_2^2 + \lambda |\mathbf{s}|_\gamma, \quad |\mathbf{s}|_\gamma = \sum_d |s_d|^\gamma$$

$$\mathbf{g}_{\frac{1}{2} \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_2^2} = -\mathbf{A}^\top (\mathbf{x} - \mathbf{A}\mathbf{s})$$

$$\mathbf{H}_{\frac{1}{2} \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_2^2} = \mathbf{A}^\top \mathbf{A}$$

$$\mathbf{g}_{|\mathbf{s}|_\gamma} = \gamma |\mathbf{s}|_{\gamma-1} \bullet \operatorname{sign}(\mathbf{s})$$

$$\operatorname{diag}(\mathbf{H}_{|\mathbf{s}|_\gamma}) = (\gamma - 1)\gamma |\mathbf{s}|_{\gamma-2}$$

Note: Hessian not in general defined at $s_j=0$

$\gamma=1$



$\gamma=1.1$



$\gamma=2$



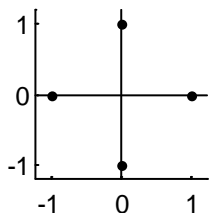
$\gamma=4$



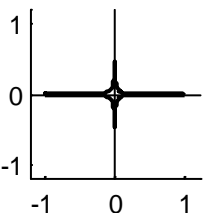
$\gamma=\infty$



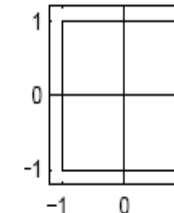
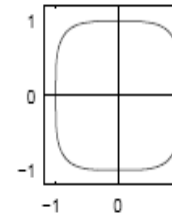
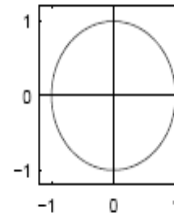
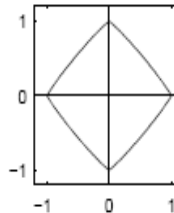
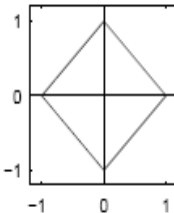
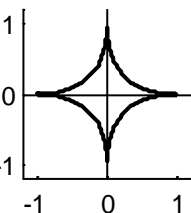
$\gamma=0$



$\gamma=0.25$



$\gamma=0.5$



Is sparse coding unique?

$$\mathbf{X} \approx \mathbf{A}\mathbf{S} = (\mathbf{A}\mathbf{Q}^{-1})(\mathbf{Q}\mathbf{S}) = \hat{\mathbf{A}}\hat{\mathbf{S}}$$

$$|\mathbf{S}|_1 = |\tilde{\mathbf{S}}'|_1$$

This trivially holds if for instance \mathbf{S} is non-negative and \mathbf{Q} is a Markov matrix, i.e.

$$\mathbf{Q} \geq \mathbf{0}, \quad \sum_i q_{i,j} = 1$$

Furthermore, trivial sparse solution given by $|\mathbf{A}|_1 \rightarrow \infty, |\mathbf{S}|_1 \rightarrow 0$

Solution: \mathbf{A} constrained to have fixed norm, i.e.
or \mathbf{A} equivalently ℓ_2 regularized, i.e.

$$\|\mathbf{a}_d\|_F = 1$$
$$\mathbf{A} = \mathbf{X}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top + \mathbf{\Lambda})^{-1}$$

Solving the Sparse Coding problem

- Commonly solved alternatingly solving for **A** and **S**
- **S** update: Sparse Regression problem

$$C(\mathbf{s}) = \frac{1}{2} \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_F^2 + \lambda |\mathbf{s}|_1$$

- **A** update: Norm constrained least squares problem

$$C(\mathbf{A}) = \frac{1}{2} \|\mathbf{X} - \mathbf{A}\mathbf{S}\|_F^2 \text{ s.t. } \|\mathbf{a}_d\|_F = 1$$

Solving for \mathbf{S} :

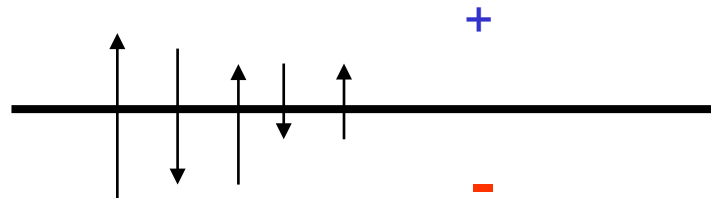
Simple gradient descent does not work!

$$C(\mathbf{s}) = \frac{1}{2} \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_F^2 + \lambda |\mathbf{s}|_1$$

$$\mathbf{g}_{C(\mathbf{s})} = -\mathbf{A}^\top (\mathbf{x} - \mathbf{A}\mathbf{s}) + \lambda \text{sign}(s)$$

$$\text{Gradient descent: } \mathbf{s} \leftarrow \mathbf{s} - \mu \mathbf{g}_{C(\mathbf{s})}$$

Problem: Coefficients that ideally should be zero will oscillate around zero.

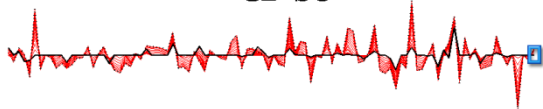


Reconstruction penalty

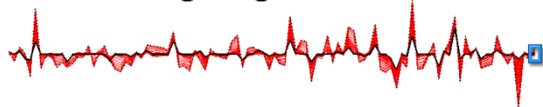
Sparsity penalty

$$C(\mathbf{s}) = \frac{1}{2} \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_F^2 + \lambda \|\mathbf{s}\|_1$$

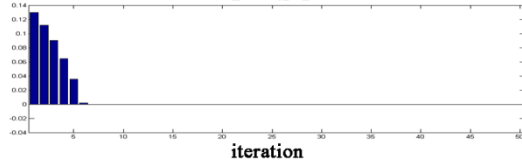
GB-SC



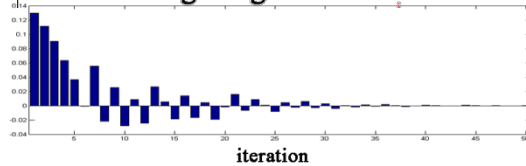
Regular gradient descent



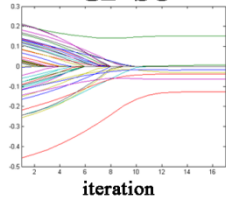
GB-SC



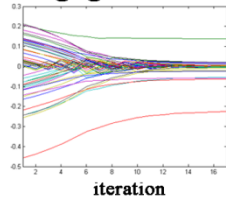
Regular gradient descent



GB-SC



Reg. grad. desc.



Gradient Based Sparse Coding (GB-SC)

- 1: **repeat**
- 2: Update \mathbf{s} according to reconstruction penalty
- 3: $\mathbf{s}^{new} = \mathbf{s} - \mu(\mathbf{A}^\top(\mathbf{A}\mathbf{s} - \mathbf{x}))$
- 4: Update \mathbf{s}^{new} according to the sparsity penalty such that element crossing zero are set to zero
- 5:
$$\mathbf{s}_j^{new} = \begin{cases} 0 & \text{if } |\mathbf{s}_j^{new}| < \mu\lambda \\ \mathbf{s}_j^{new} - \mu\lambda \text{sign}(\mathbf{s}_j^{new}) & \text{otherwise} \end{cases}$$
- 6: **if** $C(\mathbf{s}^{new}) < C(\mathbf{s})$ **then**
- 7: $\mu = 1.2\mu$
- 8: $\mathbf{s} = \mathbf{s}^{new}$
- 9: **else**
- 10: $\mu = \mu/2$
- 11: **end if**
- 12: **until** convergence

Solving for \mathbf{A} :

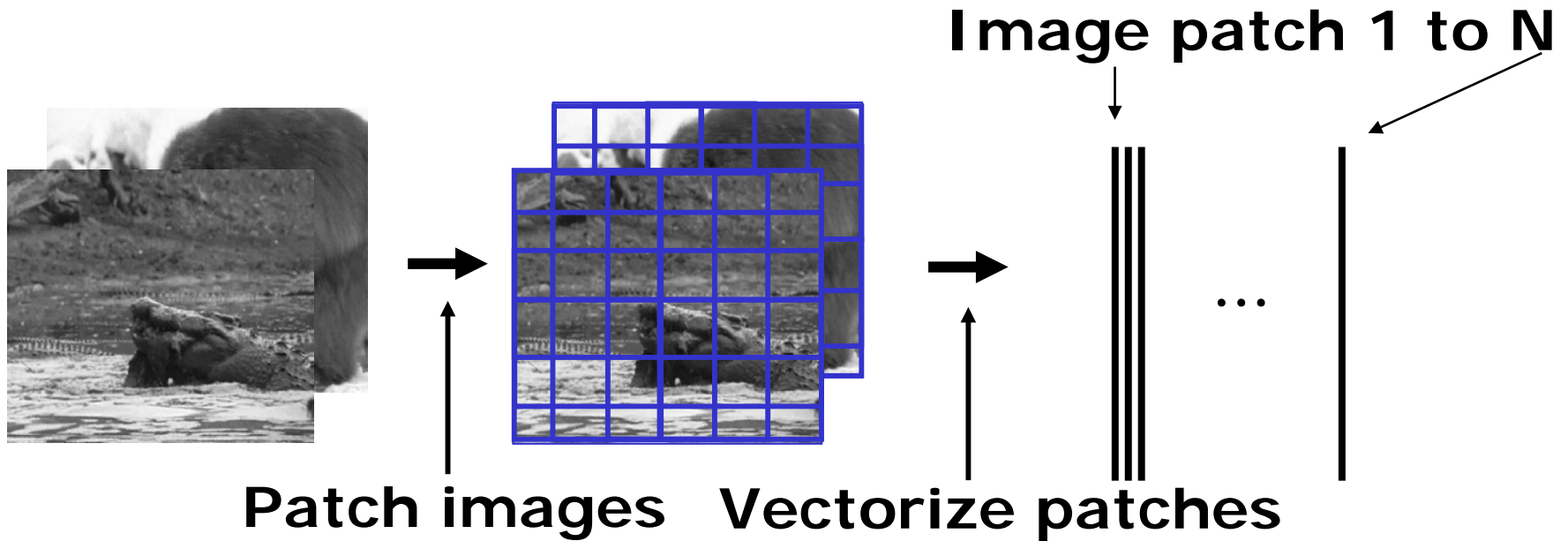
Change of variable

$$\begin{aligned}\mathbf{X} &= \tilde{\mathbf{A}}\mathbf{S} + \mathbf{E}, \\ \tilde{a}_{i,d} &= \frac{a_{i,d}}{\sqrt{\sum_i a_{i,d}^2}}\end{aligned}$$

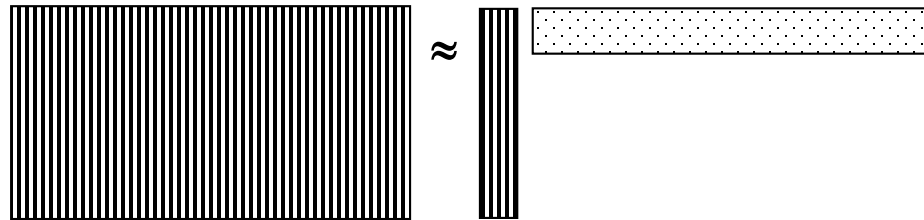
Minimize by gradient descent

$$\frac{1}{2}\|\mathbf{X} - \tilde{\mathbf{A}}\mathbf{S}\|_F^2$$

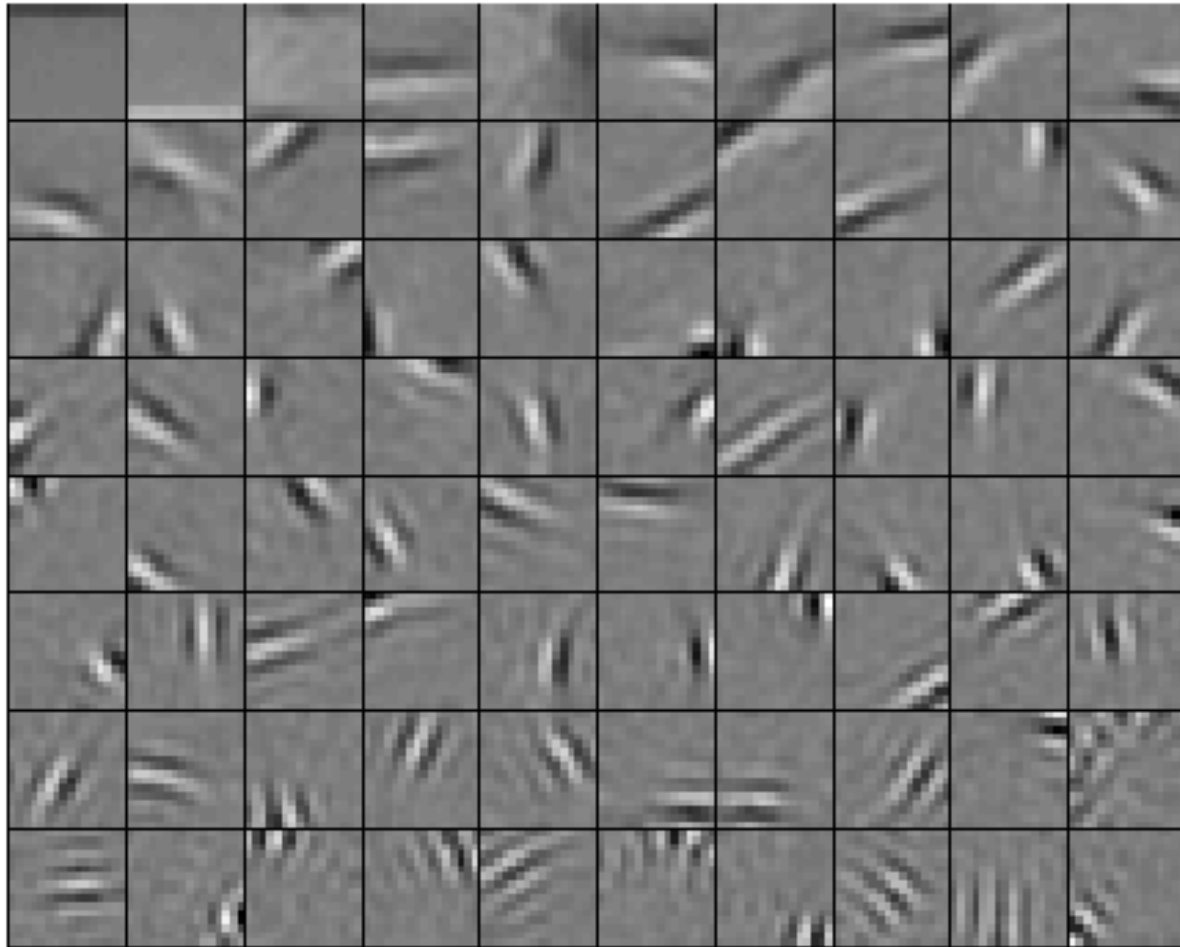
Sparse Coding (cont.)



$$C(\mathbf{A}, \mathbf{S}) = \frac{1}{2} \|\mathbf{X}^{pixels \times patches} - \mathbf{A}^{pixels \times feat.} \mathbf{S}^{feat. \times patches}\|_F^2 + \lambda |\mathbf{S}|_1$$



A corresponds to Gabor like features resembling simple cell behavior in V1 of the brain!!





NMF



Archet. Anal



ICA



Sparse Coding



William of Ockham
(1288-1347)

Principle of parsimony

Great starting point for exploratory analysis

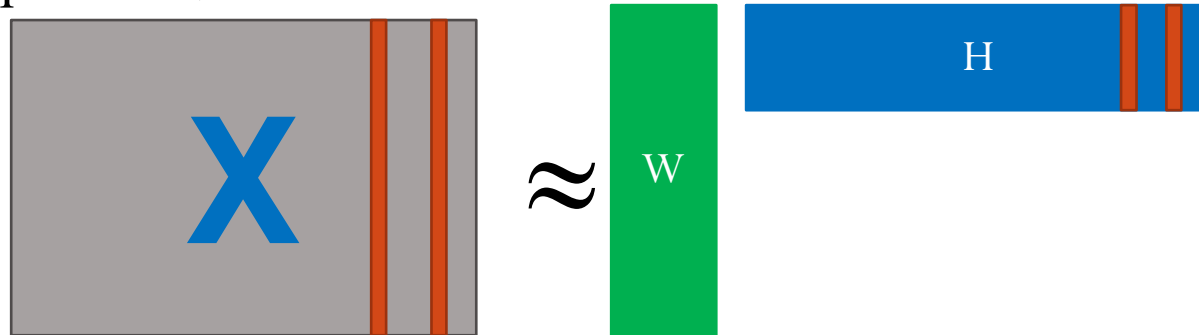


Finding the number of components/Degree of sparsity etc. in Unsupervised Learning

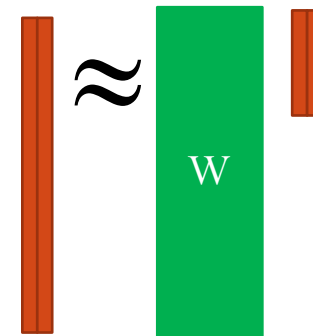
- Cross-validation
- NPAIRS (compare the consistency of the extracted components across data splits (Strother et al., NeuroImage 2002))
- Bayesian frameworks approximating generalization error by the model evidence (c.f. Bishop chapter 12)

How can we use Cross-Validation?

- Simple idea, remove some column of \mathbf{X} for validation



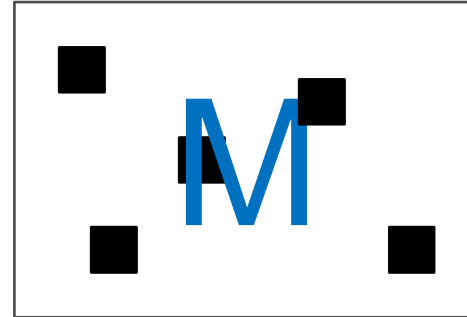
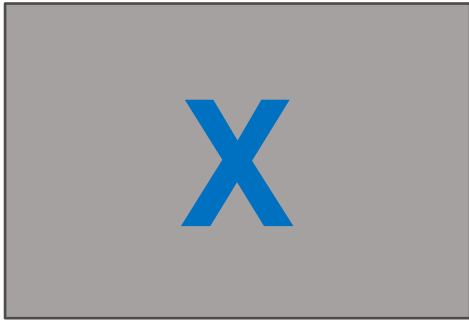
Then fit these columns based on the estimated \mathbf{W} , i.e.



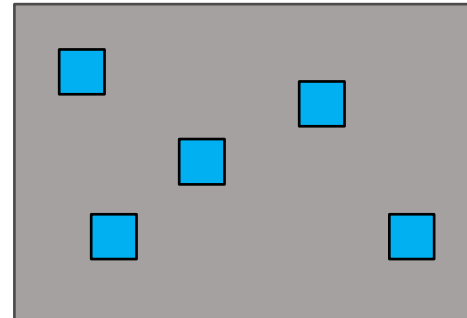
- Problem: More features will always improve the fit!

Cross-validation

- Treat part of the data as missing during the estimation process and validate the model on this test-data set.



$$\mathbf{WH} =$$



Least squares objective function:

$$\frac{1}{2} \|(\mathbf{1} - \mathbf{M}) * (\mathbf{X} - \mathbf{WH})\|_F^2 = \frac{1}{2} \sum_{i,j} (1 - m_{i,j}) (x_{i,j} - (\mathbf{WH})_{i,j})^2$$

Prediction error

$$\frac{1}{2} \|(\mathbf{M} * (\mathbf{X} - \mathbf{WH}))\|_F^2 = \frac{1}{2} \sum_{i,j} m_{i,j} (x_{i,j} - (\mathbf{WH})_{i,j})^2$$

Problem: Test and training data not necessarily independent

- Presented basic very simple 1st order methods (i.e. Gradient methods) for solving the various models. Note however, that more specialized algorithms exist for each of the described problems that may have better convergence properties!

Application of Unsupervised Learning


- Collaborative Filtering
- Bio-informatics
- NeuroImaging
- Web-mining
- Chemo-informatics
- Computer Vision

To mention but a few

Useful in general for multi-variate data analysis! In particular when facing a high degree of correlation and redundancy in the data.

Exploratory analysis can drive novel hypothesis!

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson  06.23.08



THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't

Analysis of massive amounts of data will be the main driving force of all sciences in the future!!

Today's exercise

- ICA method based on fastICA
- NMF, Archetypal Analysis and Sparse Coding implemented by the gradient based methods described in the slides.