Subspace models

Line Clemmensen

DTU

02582 Computational Data Analysis, 2019

Today's Lecture

- Recap
- PCA - Principal Component Analysis
  - The review to rule them all
  - The dos ans dont's
  - Applications
- PCR - Principal Component Regression (Application)
- PLS - Partial Least Squares
- CCA - Canonical Correlation Analysis

Last Week, Lagrange optimization

Original

$$\max_x f(x)$$

such that

- $g_j(x) = 0 \quad \forall j$
- $h_k(x) \geq 0 \quad \forall k$

Lagrange primal

$$\max_x \min_{\substack{\lambda \\ \mu \geq 0}} L_P(x, \lambda, \mu)$$

fulfilling

- Karush-Kuhn-Tucker

Lagrange dual

$$\min_{\substack{\lambda \\ \mu \geq 0}} \max_x L_P(x, \lambda, \mu)$$

fulfilling

- Karush-Kuhn-Tucker
- Slater

Lagrange primal function:

$$L_p(x, \lambda, \mu) = f(x) + \sum_j \lambda_j g_j(x) + \sum_k \mu_k h_k(x).$$

Where we use it,

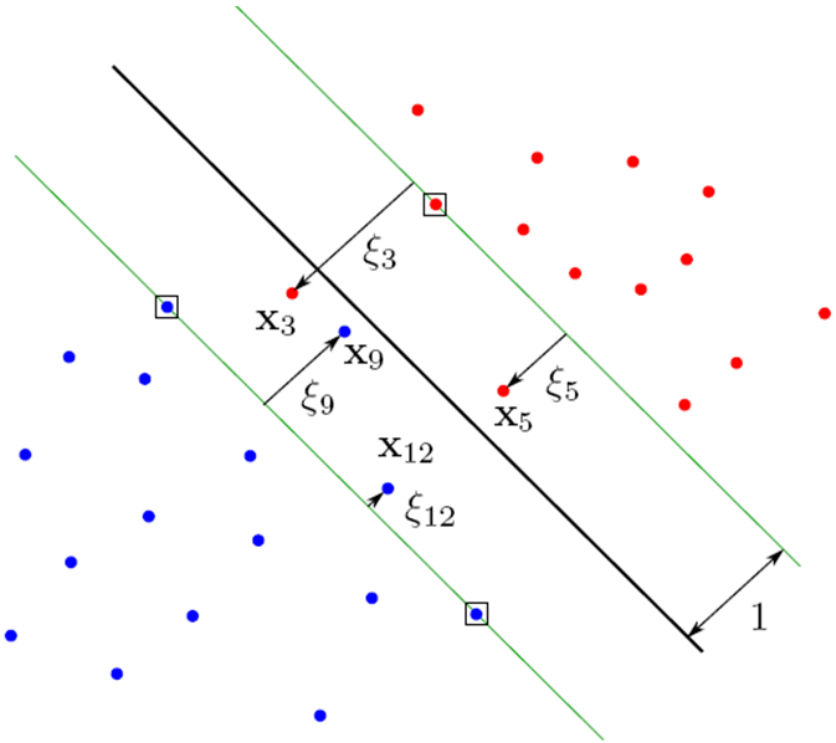
- SVM formulation for the kernel trick
- Two formulations of ridge and lasso
- Will be used to derive PCA

Last Week, The support Vector Machine

A two category classifier,

- Points that touch the boundary of the margin are support points (marked with squares)
- Budget for overlap (introduce slack variables)

$$\begin{cases} \arg\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + \lambda \sum_{i=1}^{n} \xi_i \\ \text{such that} \\ \quad y_i(x_i\beta - \beta_0) \geq 1 - \xi_i \quad \forall i \\ \quad \xi_i \geq 0 \quad \forall i \end{cases}$$



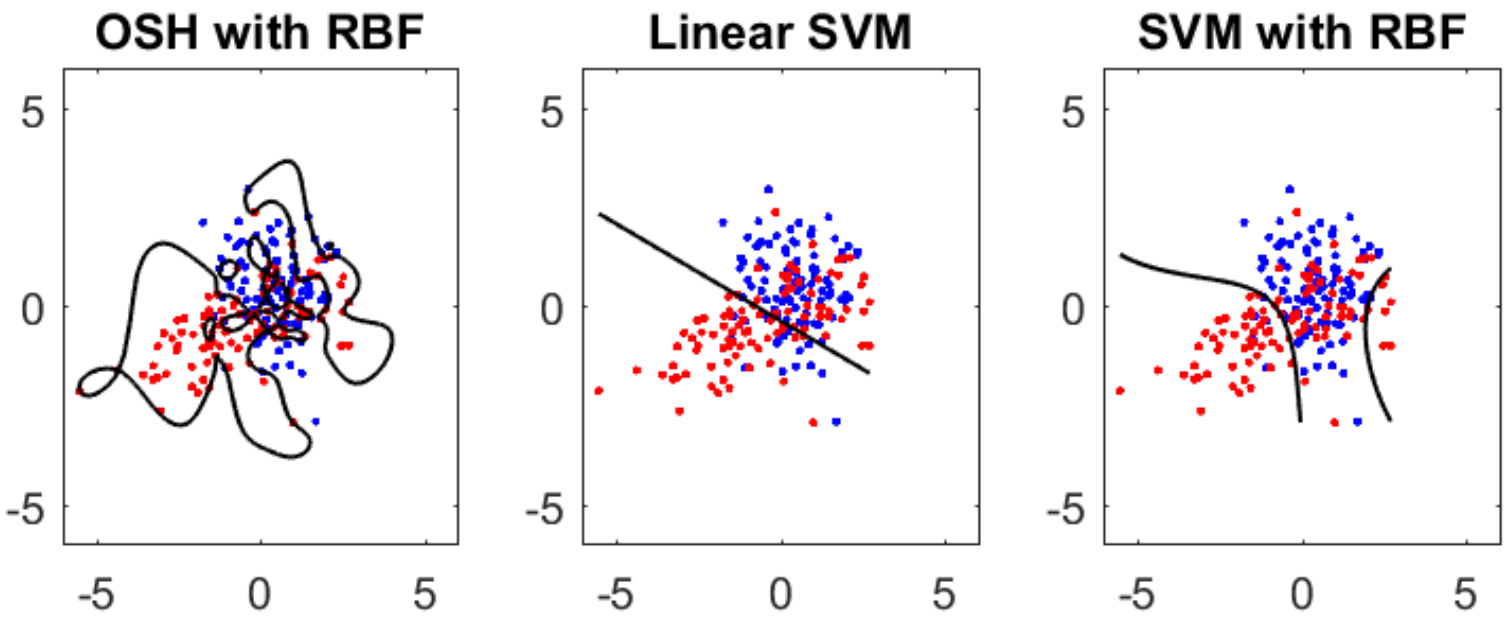
Last Week, The kernel trick

Replace XX^T with a basis expansion $h(X)h(X)^T$. The kernel trick is to implicitly define the basis expansion through $K(X) = h(X)h(X)^T$.

Most common choice is the Radial Basis Function,

$$K_{i,j} = \exp\left(-\frac{1}{c}||x_i - x_j||^2\right)$$

Last Week, Tuning the SVM



Separation can be obtained with

- non-linear boundaries,
- allowing for overlap,

or optimally by using both.

Principal Component Analysis

Principal Component Analysis

Regression and classification are confirmatory

- **Answers** to particular questions
  - Does wine-drinking influence heart disease? (regression)
  - How well can we separate between normal and abnormal ECG? (classification)
- **Supervised** - solutions are governed by the outcome variable

PCA is exploratory

- **Explore** examples of typical (common) observations based on your data set.
- **Unsupervised**, no outcome variable - let data speak for itself.
  - Structure in data
  - Outlier detection
  - Dimensionality reduction (data compression)

PCA and the Curse of Dimensionality

Say 100 observations occupy 75 % of 1-D space. Then we would need 100^{10} observations to get the same coverage in 10 dimensions.

How can a poor representation of space be useful?

- Data are clustered in space
- Data lie on a low-dimensional manifold
- Variables are correlated

PCA in parts recovers such structure

PCA - idea

Linear transformation of data: $S = XL$

Preserve relations (angels) between variables $S = XL$ subject to $L^T L = I$

- Orthogonal transformation - L is a rotation matrix.

Which rotation?

- Rotate such that the projected data S has maximal variance.
- Successively maximize the variance of the principal components. Loading l_m solves

$$\arg\max_{\alpha} Var(X\alpha)$$

$$\text{subject to } ||\alpha|| = 1,\ \alpha^T l_j = 0,\ j = 1, ..., m-1$$

Principal Component Analysis

Example: 12 observations, 2 dimensions

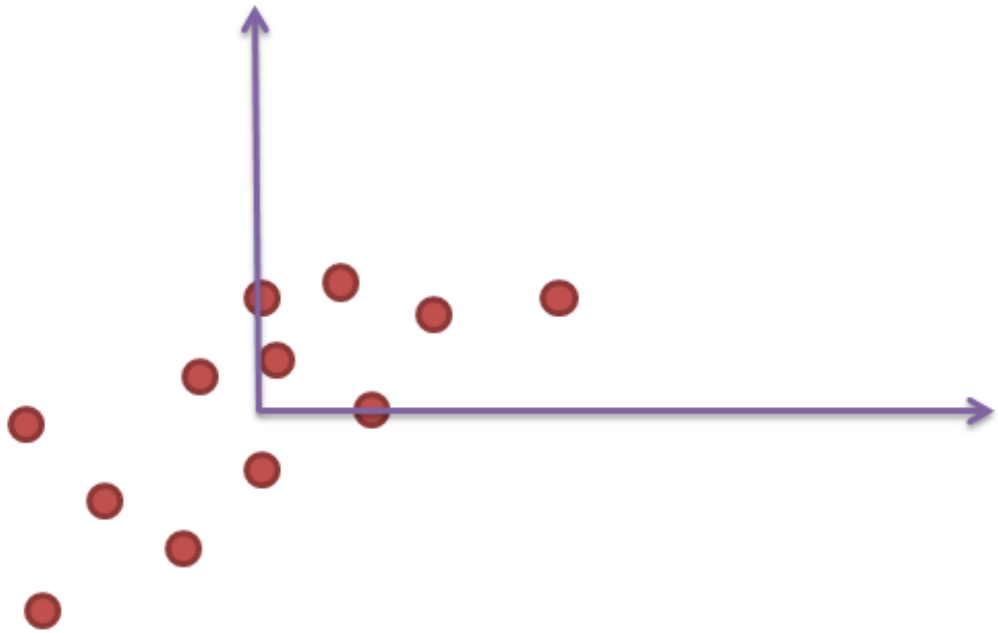
Z obs

Principal Component Analysis

Center data by removing the mean

Z obs

$X = Z - \mu_Z$



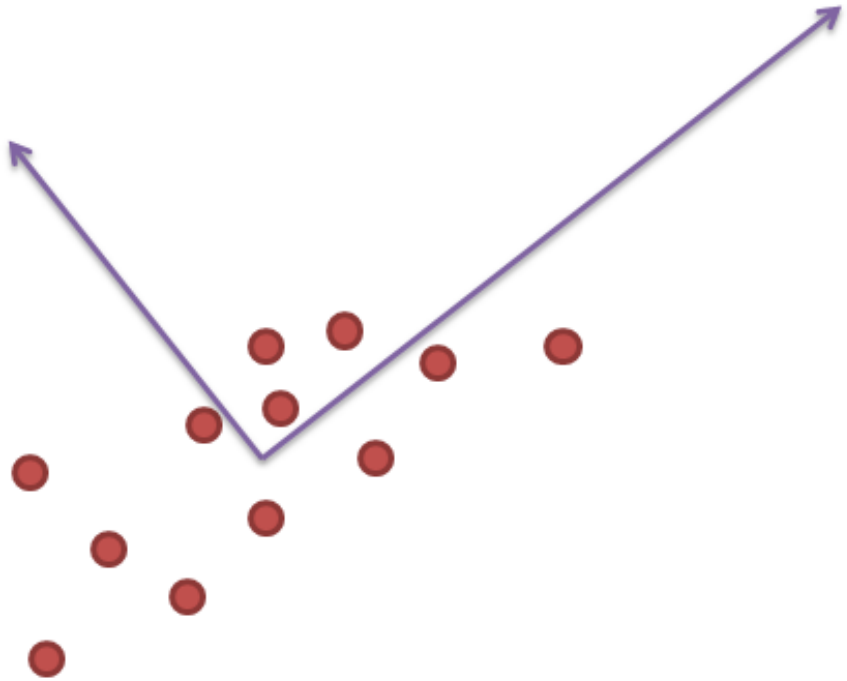
Principal Component Analysis

Rotate coordinate system. First axes in direction of maximal variance.

Z obs

$X = Z - \mu_Z$

$S = XL$



Principal Component Analysis

Observations are now given as coordinates in a new coordinate system

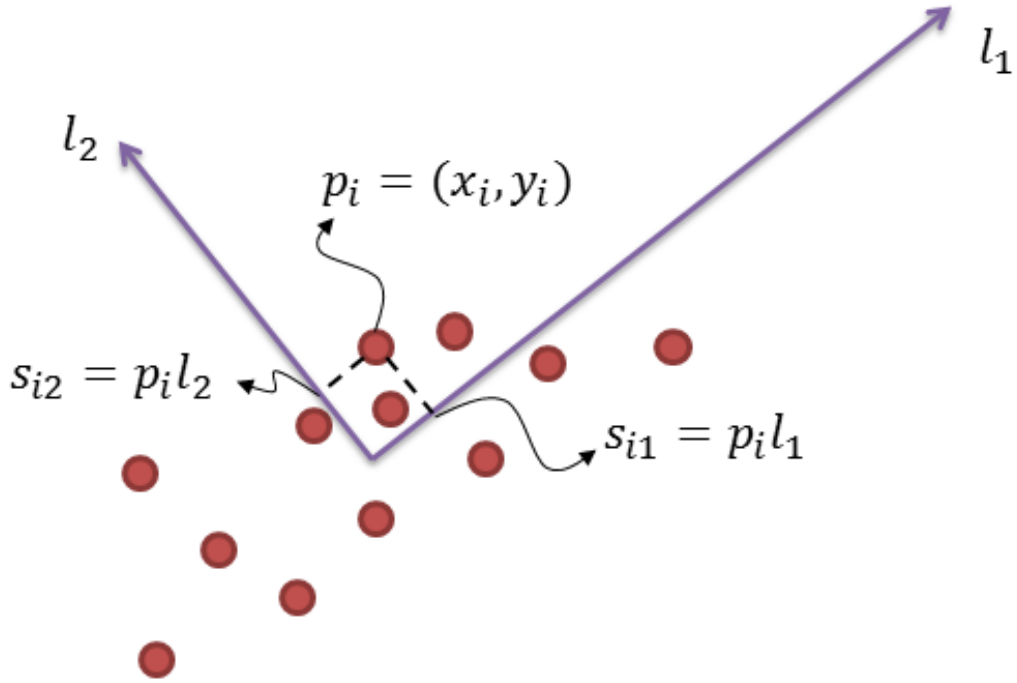
Z obs

$X = Z - \mu_Z$

$S = XL$

$s_i = p_i L$

$s_i = (s_{i1}, s_{i2})$



PCA - Derivation

Transformation,

$$S = XL, \quad L^T L = I$$

Maximize variance of projected data $\Longrightarrow$ maximize variance of each PC/columns of S

$$cov(S) = \frac{1}{n} S^T S = \frac{1}{n} L^T X^T X L = L^T \Sigma L, \quad \Sigma = cov(X)$$

First PC,

$$\arg\max_l l^T \Sigma l \quad \text{subject to } l^T l = 1$$

$$L_p = l^T \Sigma l - \lambda(l^T l - 1)$$

$$\frac{\partial L_p}{\partial l} = 2\Sigma l - 2\lambda l = 0 \Longleftrightarrow \Sigma l = \lambda l$$

Eigenvalue problem: Covariance is maximized for l equal to the eigenvector of Σ corresponding to the largest eigenvalue λ.

Remaining PCs: Orthogonalize data wrt previous components and repeat. But! L is orthogonal, so no need to orthogonalize. Eigenvectors and -values are the solutions.

Scores and Loadings

$S = XL$

S - **the scores**

- Size is $n \times m$, $m = min(n, p)$
- Coordinates of data points on new axes
- Columns of S are the **principal components (PC)**
- PCs are uncorrelated - $S^T S$ is diagonal

L - **the loadings**

- Size is $p \times m$, $m = min(n, p)$
- Columns are known as **the principal axes**
- Rotation matrix $L^T L = I$
- Columns are orthogonal and of unit length

Principal components are uncorrelated

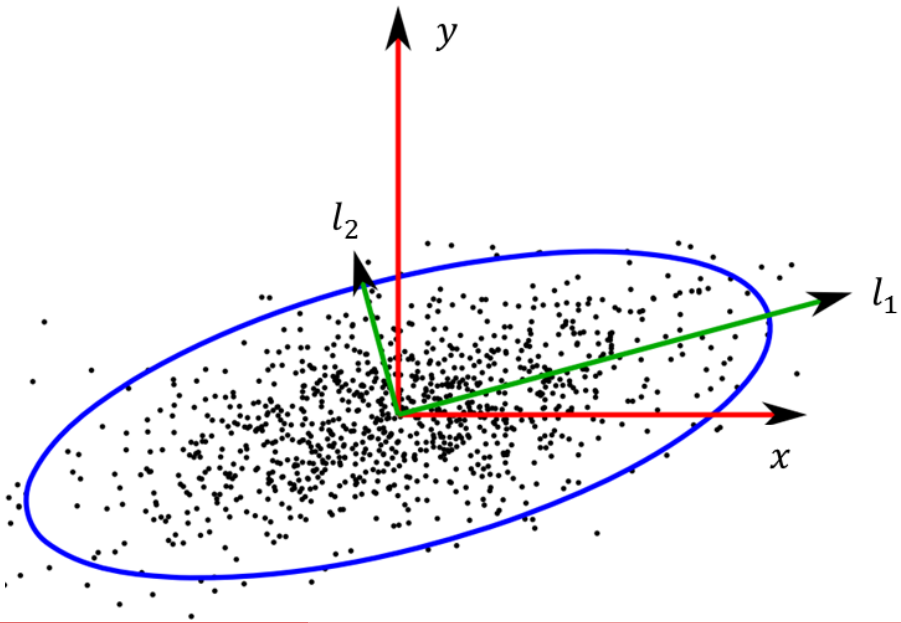
With Gaussian data we have

$$x_i \in N(\mu, \Sigma)$$

and transformed data is

$$s_i \in N(0, D) \quad \text{with} \quad D = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_k^2 \end{bmatrix}$$

where $\sigma_1^2 \geq \dots \geq \sigma_k^2$



Dimensionality reduction

The first scores explain most of the variation in data. We can use this to “compress data”.

Here is data matrix with 4 variables where we keep 3 principal components,

$$[s_1\ s_2\ s_3]_{n\times 3} = [x_1\ x_2\ x_3\ x_4]_{n\times 4}[l_1\ l_2\ l_3]_{p\times 3}$$

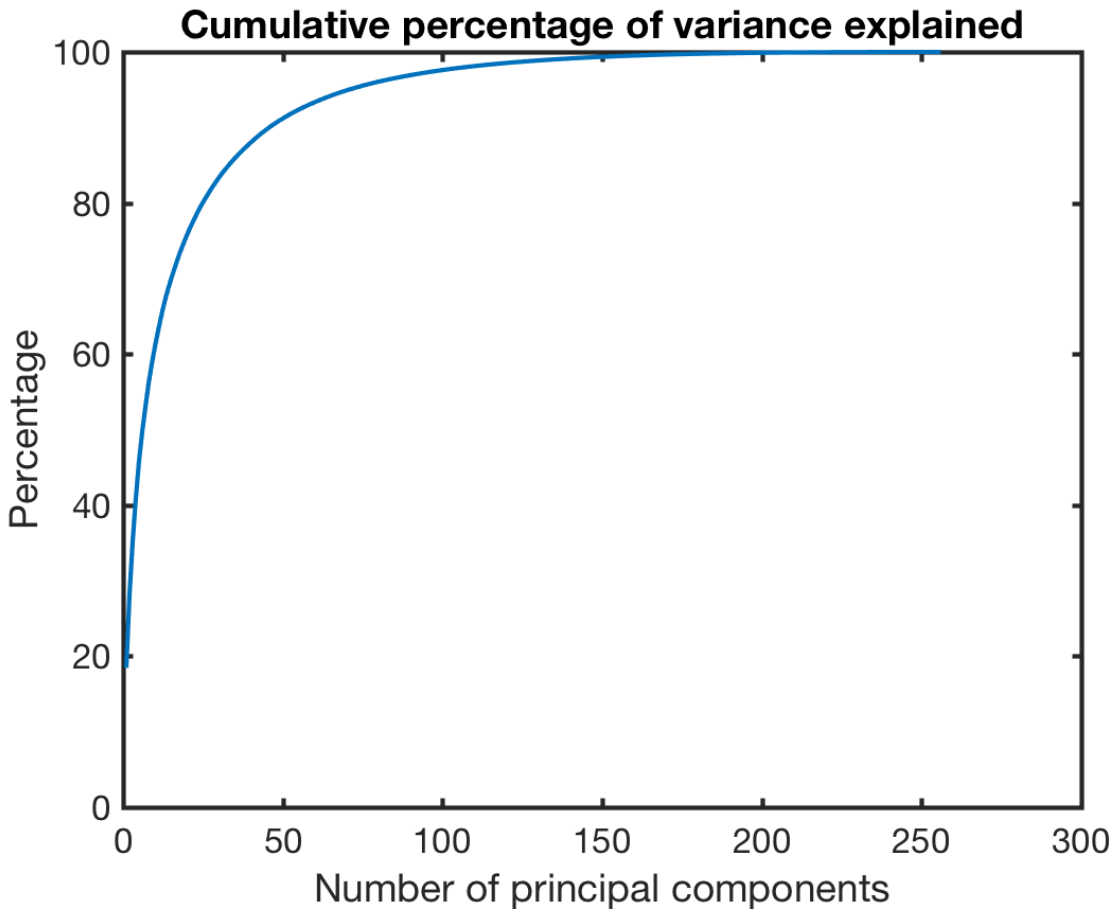
We reconstruct the observations with

$$[x_1\ x_2\ x_3\ x_4] \approx [s_1\ s_2\ s_3][l_1\ l_2\ l_3]^T$$

the error we make equals the omitted term $s_4 l_4^T$ which usually has a lower variance than previous terms.

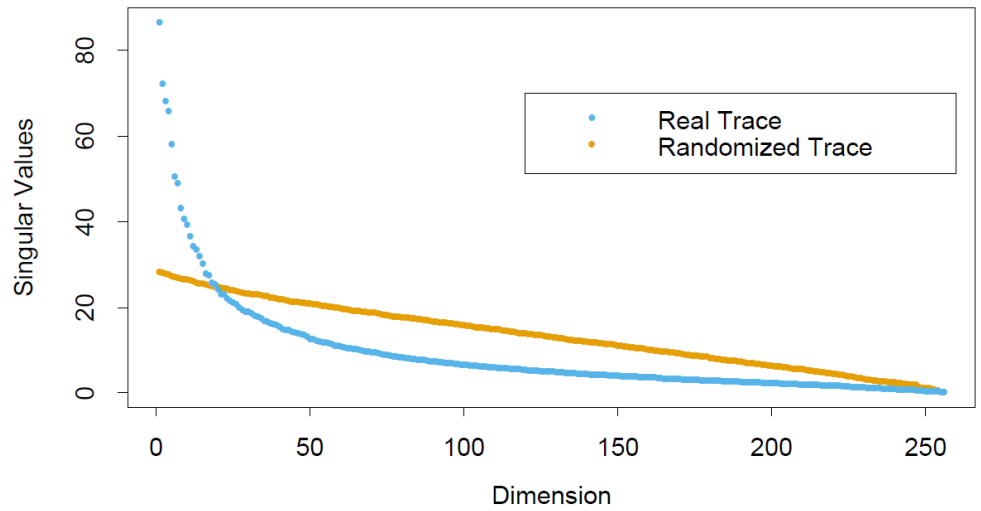
Explained variance

The fewer PCs we keep the less of the variation in data is retained.
Example: In the zip code data, the ten first PCs cover 61 % of the variance, and the first fifty PCs cover 91 % of the variance.



The number of components?

- Obtain a certain percentage of explained variance
  - Usually a bad idea
- Keep eigenvectors with eigenvalues greater than one (standardized data)
- Make a **scree plot** - compare eigenvalues to those obtained from randomized data (with same total variance).



PCA - Using Singular value Decomposition

Singular value decomposition of data matrix X: $X = US_d V^T$, where the left singular vectors are orthonormal $U^T U = I$, the right singular values are orthonormal $V^T V = I$, and S_d is a diagonal matrix with the singular values.

Note that

- $V = L$: The right singular vectors correspond to the eigenvectors, i.e. the loadings of the PCA
- $S_d = \sqrt{n\Sigma}$: The singular values are *** the eigen values, i.e. the variances corresponding to the principal components.

PCA - Computation

1. Eigen analysis of covariance matrix

   `[L D] = eig(cov(X));`

   `L` - loading matrix

   `D` - diagonal matrix of variances

   $\Sigma = LDL^T$

2. Singular value decomposition of data matrix

```
[n p] = size(X);
X = X - ones(n,1)*mean(X);
[U S_d L] = svd(X,'econ');   %X = U*S_d*L'
D = diag(diag(S_d).^2)/n);
```

   $\Sigma = \frac{1}{n} X^T X = L \left(\frac{1}{n} S_d^2 \right) L^T$

3. `[loading, score, variance] = pca(X);`

PCA captures main variation

PCA captures main variations in first principal components. This does not always mean that it captures what we are interested in

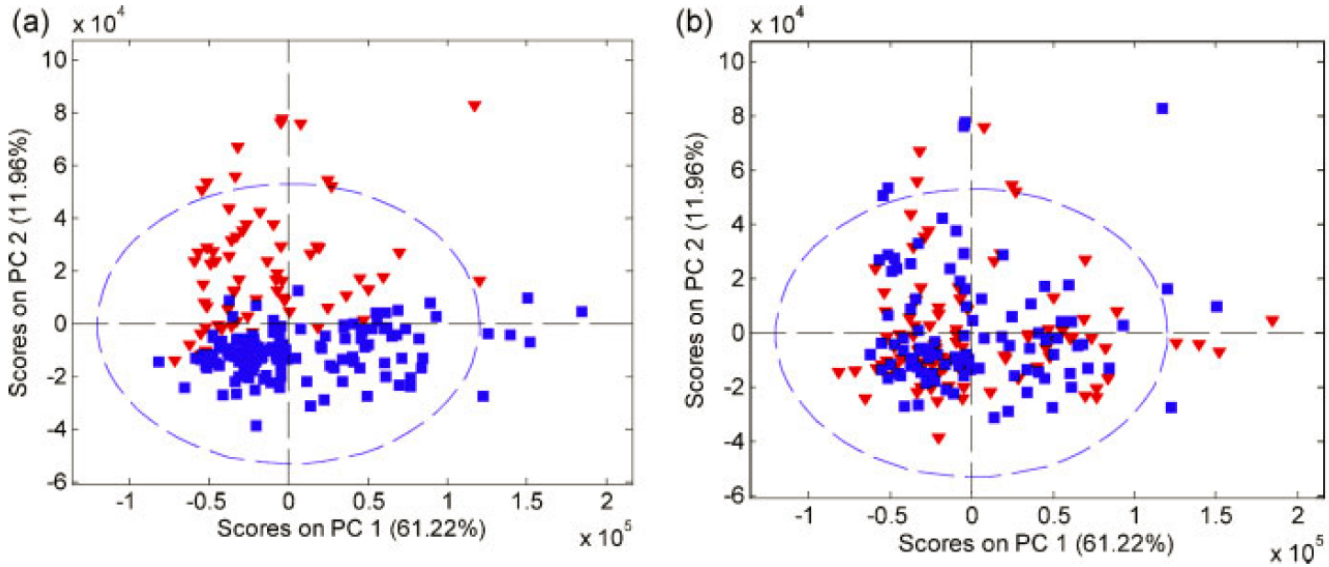


Figure 3. Two identical score plots (PC1, PC2) colored according to two different nonincluded variables: (a) male/female and (b) ill/healthy. Explained variance = 73%.

Kjeldahl & Bro, J. Chemometrics 2010; 24:558-564

Explained variance

Percentage of explained variance is relative and can not be compared

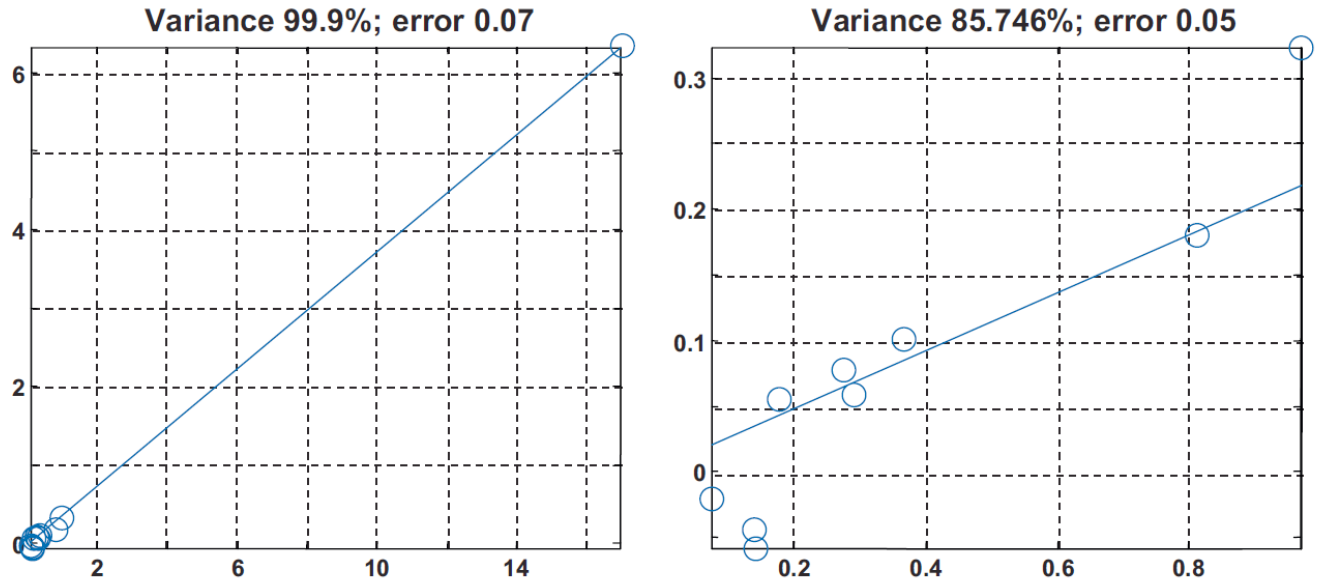


Figure 2. Example of the noncomparability of percentages across different datasets. To the left, a univariate regression describes almost 100% of the variation giving an error of 0.07. Upon removing the extreme upper-right sample, the explained variance is reduced, but so is the error.

Kjeldahl & Bro, J. Chemometrics 2010; 24:558-564

Interpreting loadings

Loadings change with scaling of variables. Use correlation loadings (standardized data) for interpretation. Variables close to each other far away from origin (0,0) are correlated. Variables close to (0,0) are not explained by the chosen principal components and not much can be concluded.

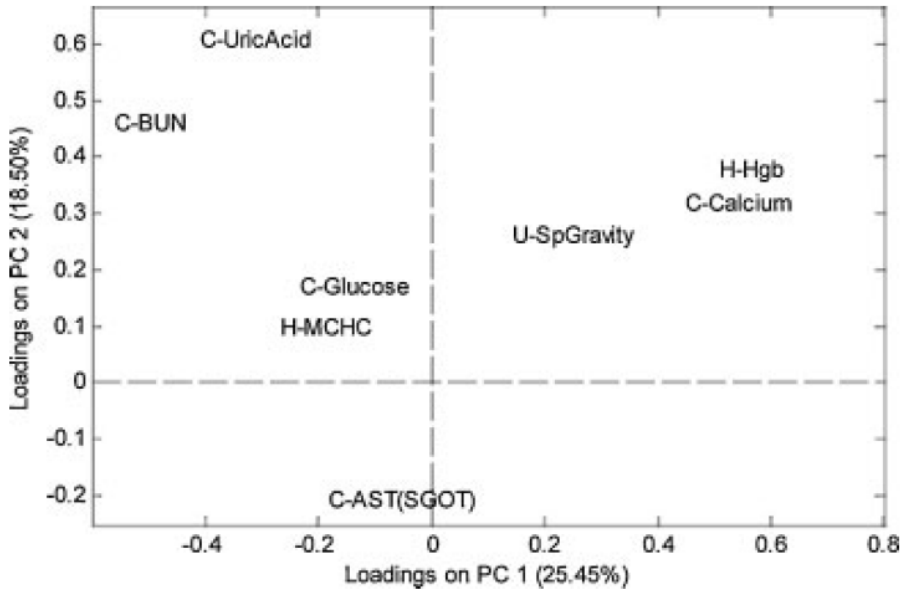


Figure 5. PCA loadings plot. H-Hgb and C-Calcium are highly correlated within the variation explained by PC1 and PC2, whereas nothing similar is certain about C-Glucose and H-MCHC regardless of their close position.

Kjeldahl & Bro, J. Chemometrics 2010; 24:558-564

Scatter plots

Scatter plots of scores might be used for spotting clusters and outliers.
Make sure that the scaling on the axes are comparable!

Figure 6. Two maps showing the same part of the world, either true to the surface distances (left), or with different scaling on the axes (right).

Kjeldahl & Bro, J. Chemometrics 2010; 24:558-564

Correlations might be coincidental

Don't trust calculated correlations when samples are few and variables are many).

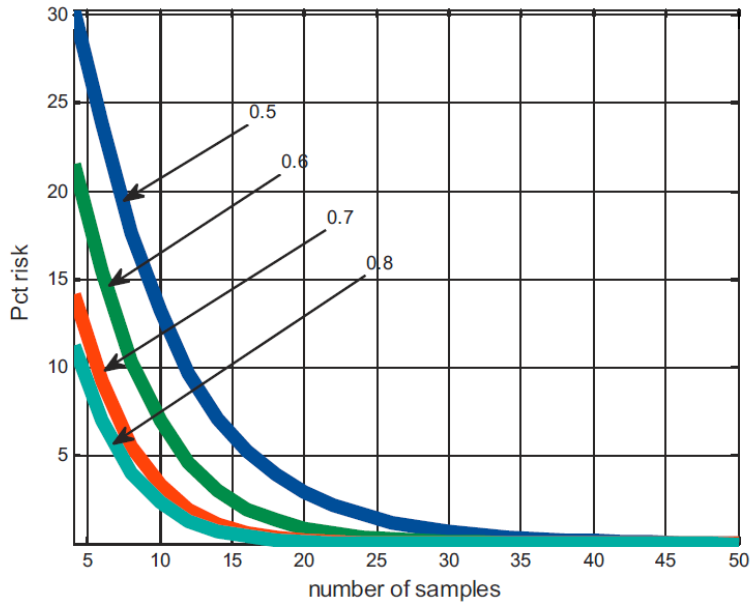


Figure 7. Effect of sample size on the risk of obtaining a Pearson's correlation r of $\pm$ 0.5–0.8 from random numbers.

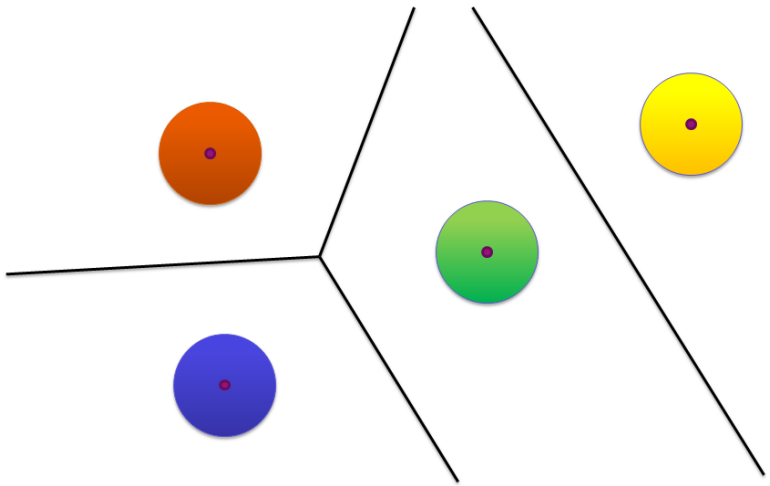
Risk of obtaining correlations of 0.5, 0.6, 0.7, and 0.8, respectively, between two random variables.

Kjeldahl & Bro, J. Chemometrics 2010; 24:558-564

Always

- Causality is not the same as an observed correlation
  - The correlation is only an observed correlation as the relation between sales of ice cream and the number of drowning accidents, where an underlying factor (the sun) is the direct causal link.
- Keep the aim of the modeling in focus
- Support choices by domain specific knowledge
- Be critical

LDA computation



Training

- Center
- Rotate to PCA directions
- Scale to unit variance
- Translate back

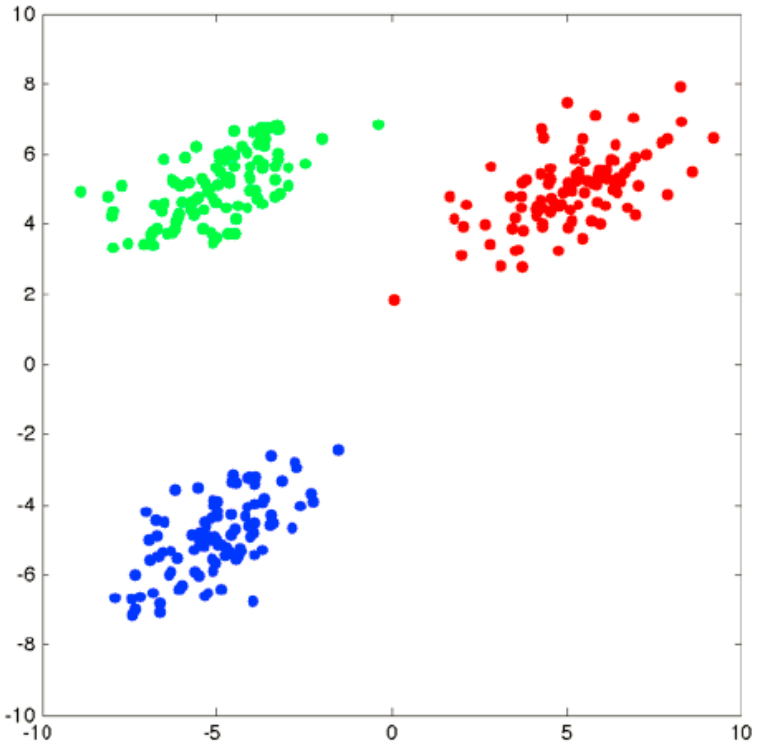
Prediction

- Transform observation into PCA-space
- Compute distance to centroids
- Assign to closest class

LDA training

Observed data

$\mathbf{x}_1$



$\mathbf{x}_2$

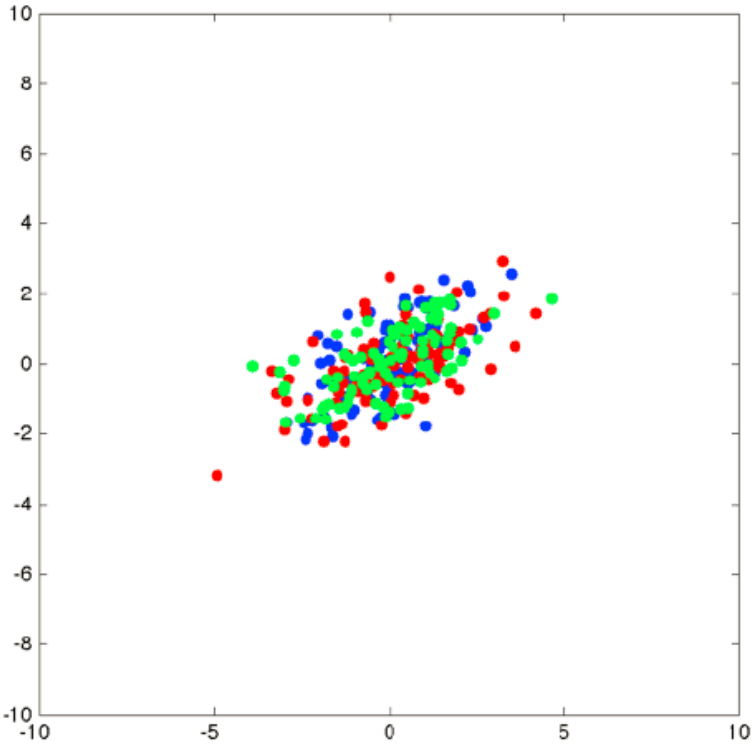
$\mathbf{x}_3$

LDA training

Center data

$\mathbf{X}_1 - \boldsymbol{\mu}_1$

$\mathbf{X}_2 - \boldsymbol{\mu}_2$



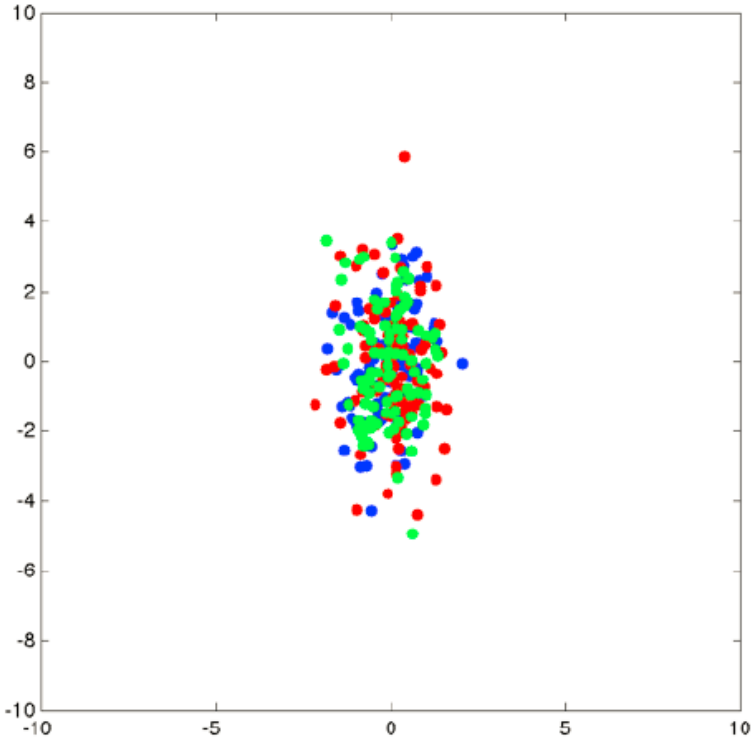
$\mathbf{X}_3 - \boldsymbol{\mu}_3$

LDA training

Rotate to principal components

$(\mathbf{X}_1 - \boldsymbol{\mu}_1)\mathbf{L}$

$(\mathbf{X}_2 - \boldsymbol{\mu}_2)\mathbf{L}$



$(\mathbf{X}_3 - \boldsymbol{\mu}_3)\mathbf{L}$

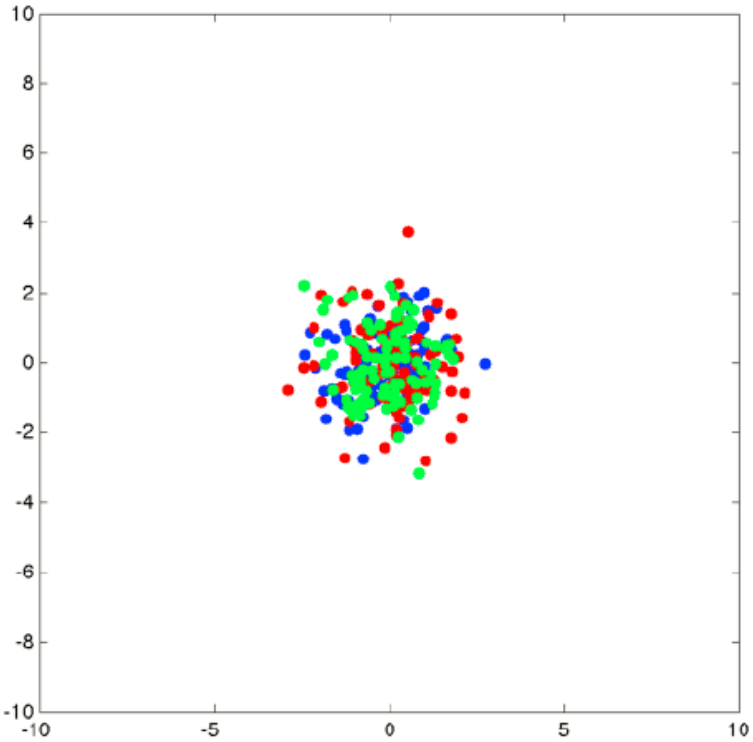
LDA training

Scale to unit variance

$(\mathbf{X}_1 - \boldsymbol{\mu}_1)\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$

$(\mathbf{X}_2 - \boldsymbol{\mu}_2)\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$

$(\mathbf{X}_3 - \boldsymbol{\mu}_3)\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$

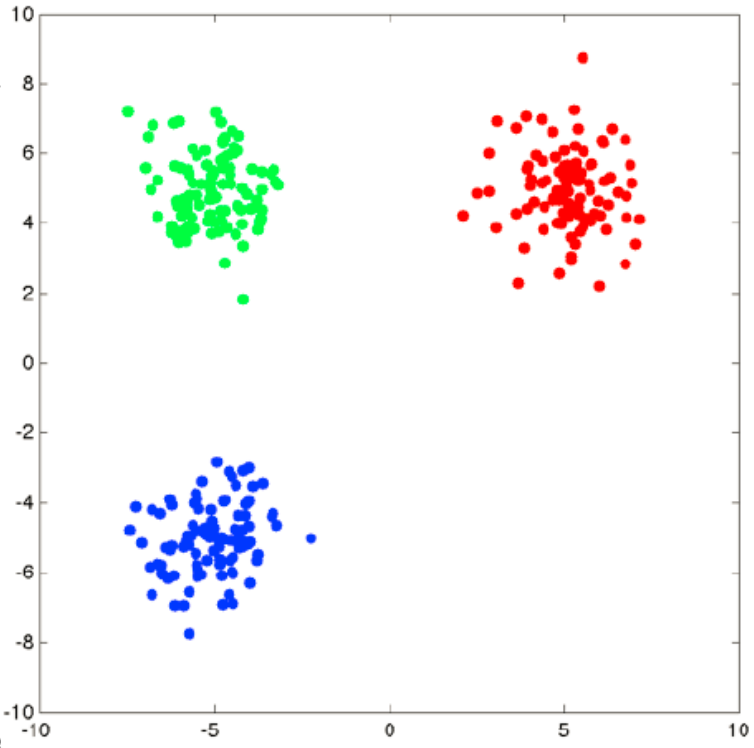


LDA training

Translate back to original mean

$(\mathbf{X}_1 - \boldsymbol{\mu}_1)\mathbf{L}\mathbf{D}^{-\frac{1}{2}} + \boldsymbol{\mu}_1$

$(\mathbf{X}_2 - \boldsymbol{\mu}_2)\mathbf{L}\mathbf{D}^{-\frac{1}{2}} + \boldsymbol{\mu}_2$



$(\mathbf{X}_3 - \boldsymbol{\mu}_3)\mathbf{L}\mathbf{D}^{-\frac{1}{2}} + \boldsymbol{\mu}_3$

LDA prediction

Calculate distance to means

$$\begin{aligned} d_k &= ||(x - \mu_k)LD^{-\frac{1}{2}} + \mu_k - \mu_k||^2 \\ &= ... \\ &= (x - \mu_k)\Sigma^{-1}(x - \mu_k)^T \end{aligned}$$

- Select smallest $(x - \mu_k)\Sigma^{-1}(x - \mu_k)^T$
- Previously, select largest $x\Sigma^{-1}\mu_k^T - \frac{1}{2}\mu_k\Sigma^{-1}\mu_k^T$
- Equivalent!

Reduced Rank LDA: Project data on first two principal components of centroids.

Today's data

- Biological shape data
- Morphometry - the quantification of shape and shape changes.
- Started in zoology - classification of species.

Sir D'Arcy Wentworth Thompson

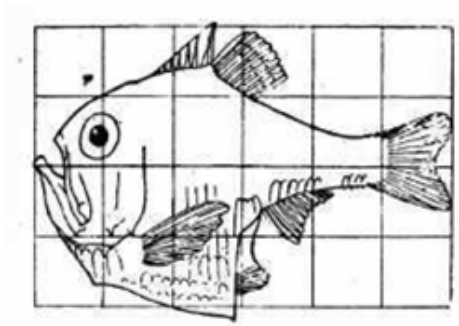


Fig. 517. *Argyropelecus Olfersi.*

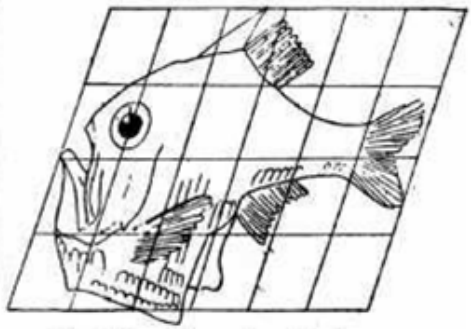


Fig. 518. *Sternoptyx diaphana.*

Shape data

Today we will work with 2D shape data

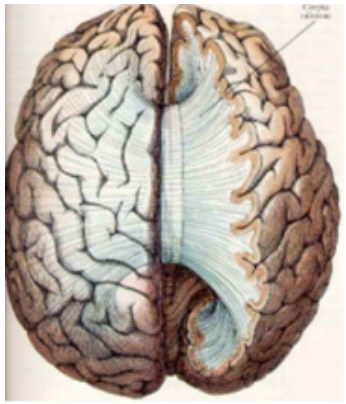
- 3D is conceptually the same, but tends to be more complex in practice.

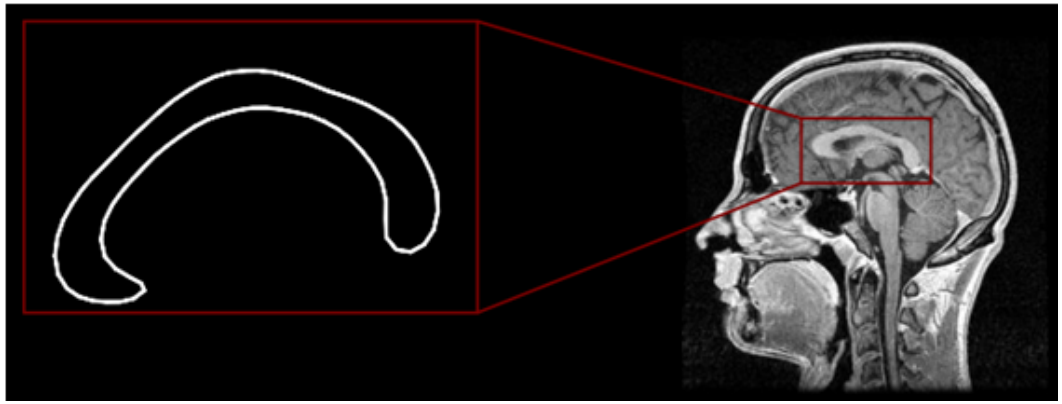
Place **landmarks** along the outline of a structure

- True landmarks, e.g. the tip of the nose
- Pseudolandmarks, e.g. at the point of maximal curvature
- Semilandmarks, e.g. equally spaced between other landmarks

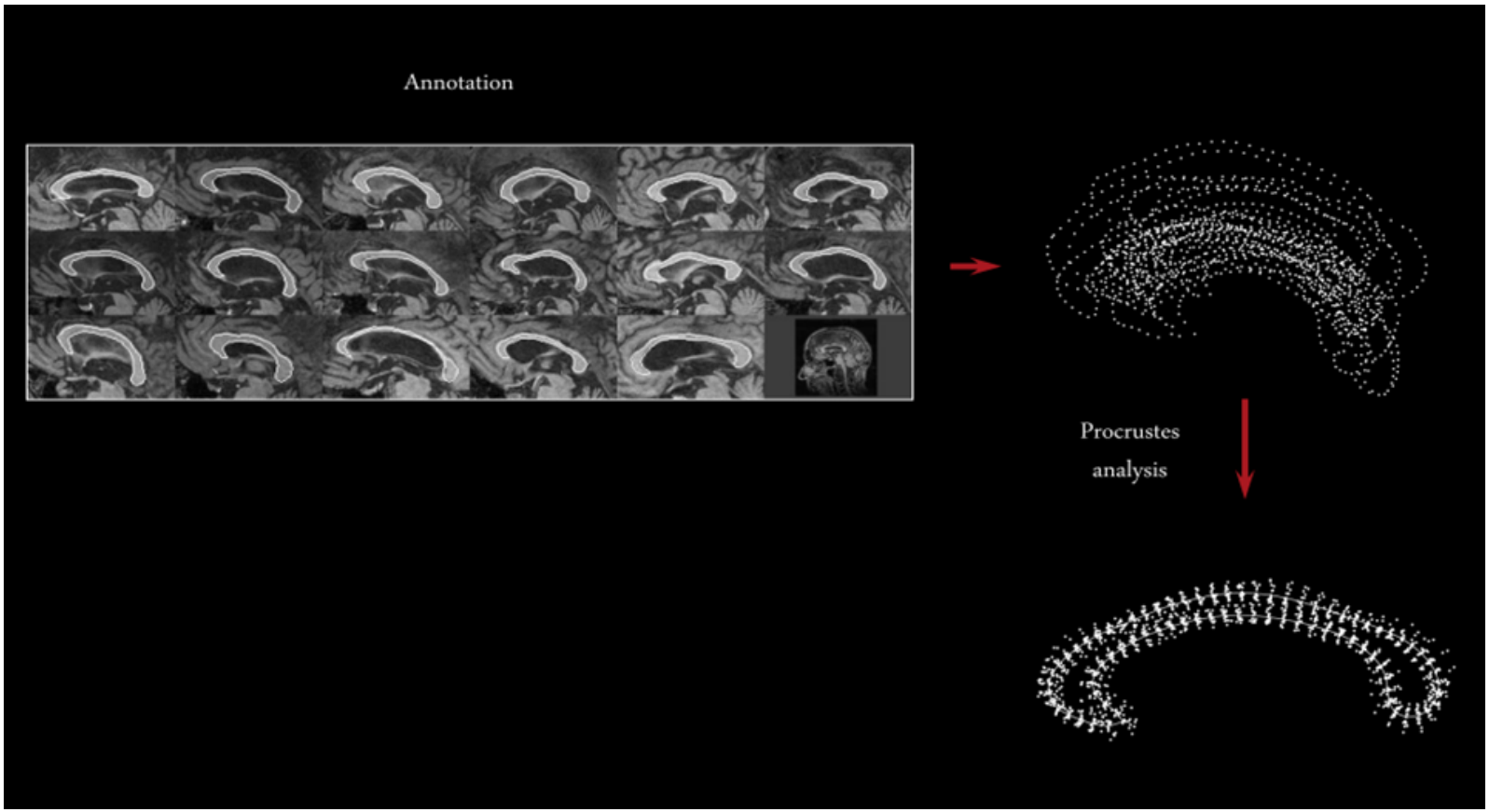
Shape data

- x and y coordinates are assumed independent
- One observation = one shape
- Coordinates are arranged as $[x_1\ x_2\ x_3\ y_1\ y_2\ y_3]$ (3 landmarks)
- p = 2 x No. of landmarks, n = No. of shapes
- Example: The corpus callosum

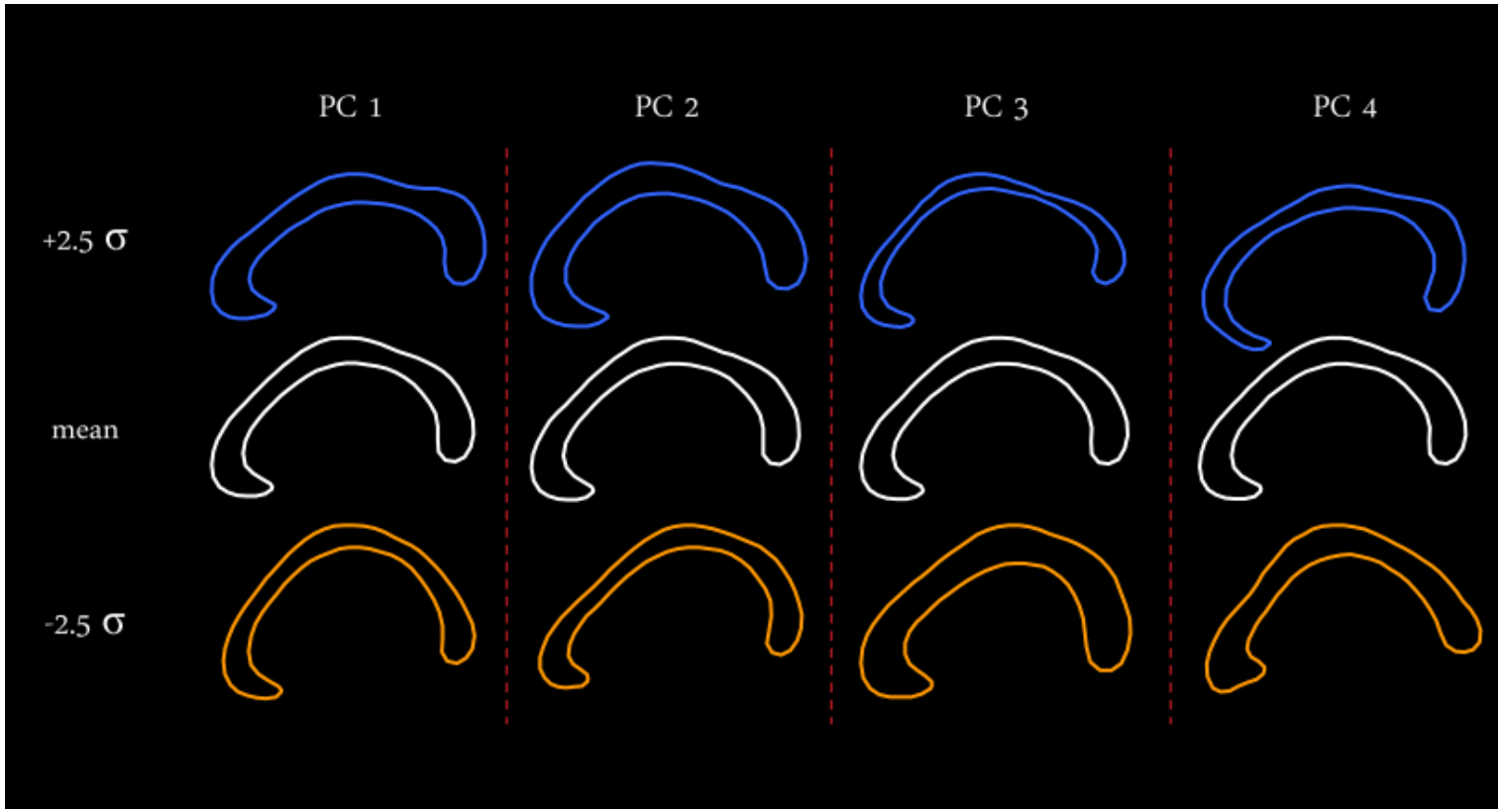




Shape analysis pipeline



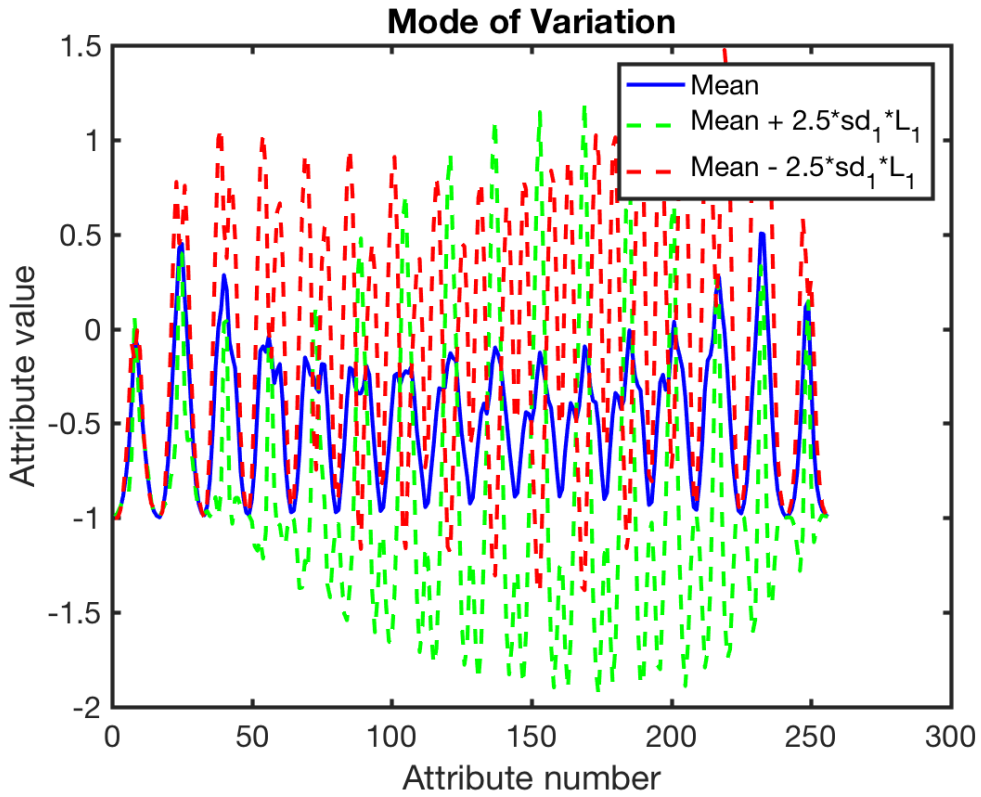
Corpus callosum decomposition - PCA



Mode of variation in PCA - Example

The mode of variation illustrates how much each principal component varies from the mean by for example using 2.5 times the standard deviation (square root of variance) for the given PC.

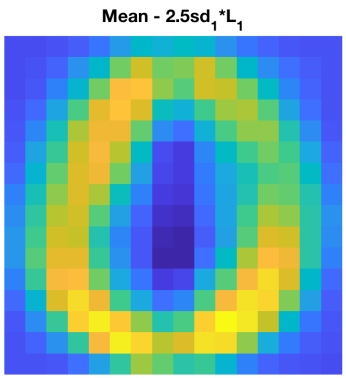
Example: Mode of variation for first principal component of zip data.

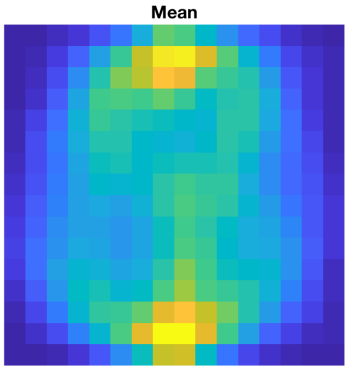


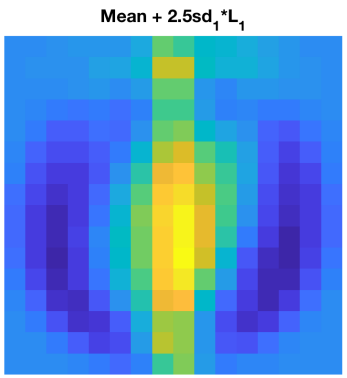
Mode of variation in PCA - Example

The mode of variation plottes in original image space for ease of interpretation (as with shape data).

Example: Mode of variation for first principal component of zip data.







Sparse PCA

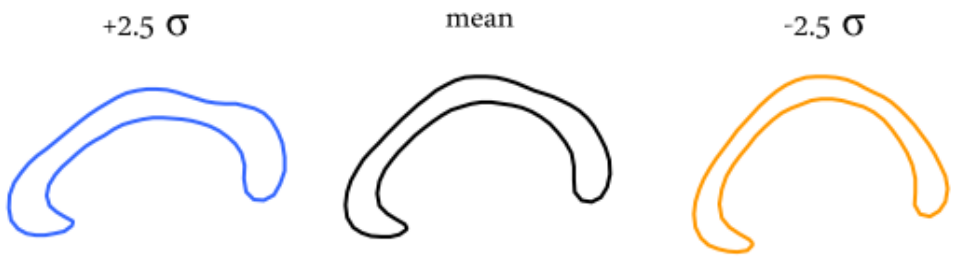
- Thresholding of loadings
- Varimax rotation of loadings
- Estimation using the Elastic Net

PCA benefits

- The **only** linear transform with independent loading vectors and uncorrelated scores.
- The linear transformation that gives the most compact data representation
- Easy and quick to calculate
  - Also for $p > n$

PCA drawbacks

- For understanding data, PCA is not optimal
- In many analyses we need to understand what is going on (in terms of the original variables)
- Example
  - What is the clinical explanation to this shape change?



Sparse PCA

- Each PC is a linear combination of **all** variables
  - $s_1 = l_{11}x_1 + l_{21}x_2 + l_{31}x_3 + l_{41}x_4$
- We have learned about sparse regression methods
  - Approximate s_1 but drive some coefficients to zero.
- Sparse PCA
  - Aim for maximization, independent loadings and uncorrelated scores, but drive some loadings to zero
  - This will not be strictly possible. Why?

Sparse PCA

- When no sparsity is imposed, we would like to have the regular PCA.
- For maximal sparsity, $L = 0$

Sparse PCA - How to calculate?

SPCA seems like a good idea

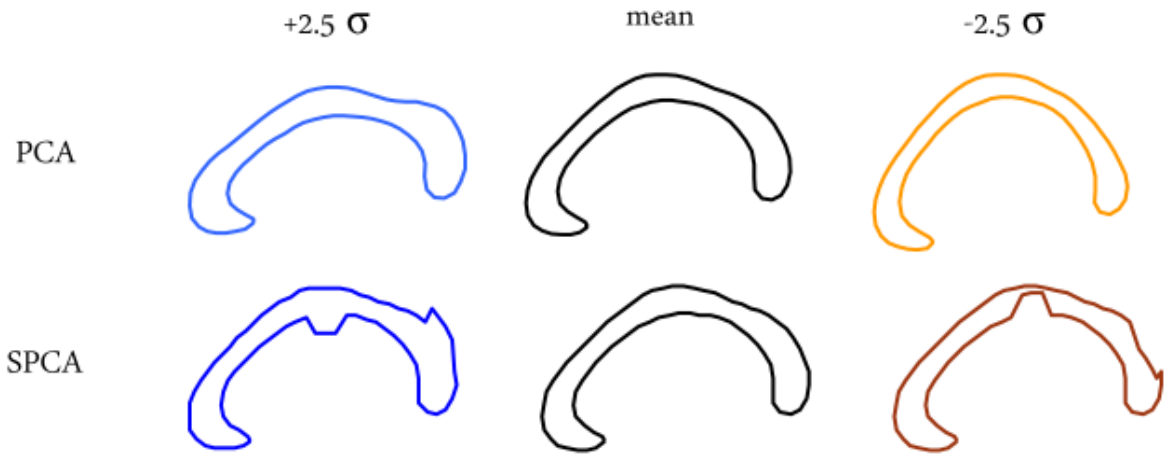
- Ok, we know why, but how?

Three techniques,

- Thresholding of loadings
- Varimax rotation of loadings
- Estimation using the elastic net

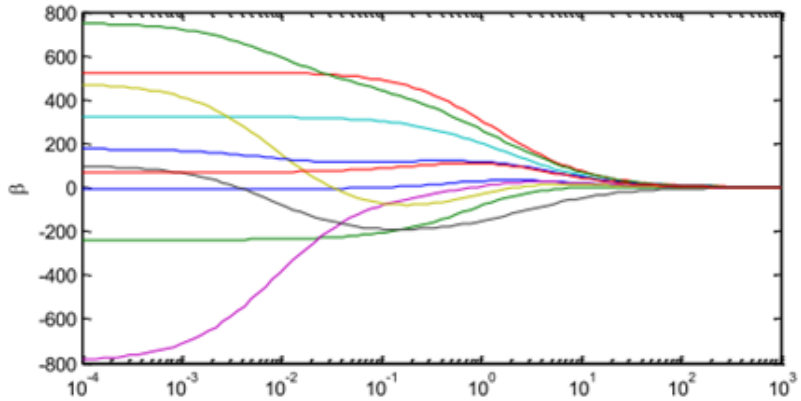
SPCA - Thresholding

- Simple!
- Set all loadings below some threshold to zero.
- Result



SPCA - Thresholding

- Why is this a bad idea?
- Think of ridge regression



- Coefficients are interdependent

Rotated Principal Components

The varimax criterion

- Produces approximately sparse loading vectors
- Quick to compute
  - ...but we skip the algorithm here
- Concept,

$$S = XL \Longleftrightarrow X = SL^T$$

$$X = SR^T RL^T = (SR^T)(LR^T)^T$$

$$\tilde{S} = SR^T \text{ rotated } S$$

$$\tilde{L} = LR^T \text{ rotated } L$$

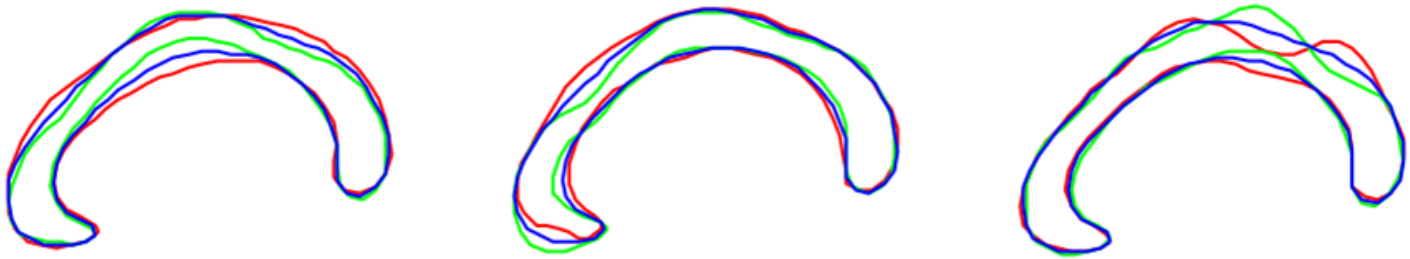
R rotates in $\mathbb{R}^p \Longrightarrow L$ stays orthogonal, S does not.

The Varimax Criterion

Maximize the variance among the loadings in each principal axis gives approximate sparseness.

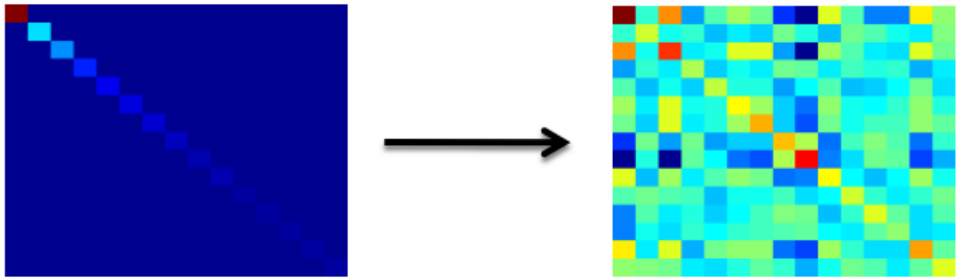
- Some loadings are high, some are close to zero.
- Amount of sparsity is determined by k, the number of retained PCs.

Example,

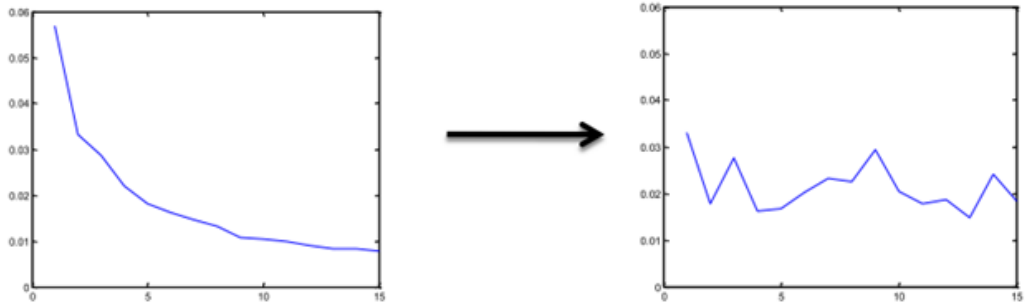


The Varimax Criterion

- Scores become correlated, $S^T S$,



- Eigenvalue spectrum becomes flattened



SPCA using the elastic net

Express each PC as a regression problem

$$\arg\min_{l} ||s_i - Xl||^2$$

Optimize wrt l using the scores s_i from PCA. This will give the loadings from PCA.

Problem: Cannot be solved when $p > n$.

Solution: Ridge regression

$$\arg\min_{l} ||s_i - Xl||^2 + \lambda||l||^2$$

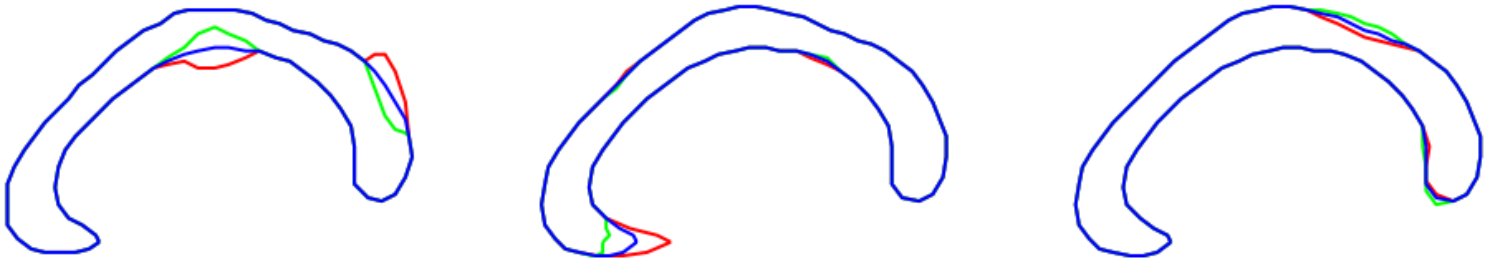
- Normalize solution to unit length
- Result: Standard PCA. Works well when $p > n$

SPCA using the elastic net

Add a L_1-penalty to get sparse solutions

$$\arg\min_{l} ||s_i - Xl||^2 + \lambda||l||^2 + \gamma||l||_1$$

Example,



Drawback, solution is guided by the original principal components.

Principal Component Regression

Linear regression on the PCA scores

Principal Component Regression, PCR

From a PCA analysis of data X we have the scores S. Use $[s_1,\ s_2\ ..., s_M]$ for some $M \leq p$ and we have a standard regression problem in the new variables,

$$y = \beta_0 + [s_1, ..., s_M]\beta + e$$

- PCR handles $n < p$ by operating on a subset of PCs.
- PCR performs similar to ridge regression
- Equivalent to OLS when $M = p$

Partial Least Squares

- Supervised method with latent variable structure
- Seeks directions which have high variance and have high correlation with the response
- Tune number of PLS components

Partial Least Squares

The mth PLS direction φ_m solves

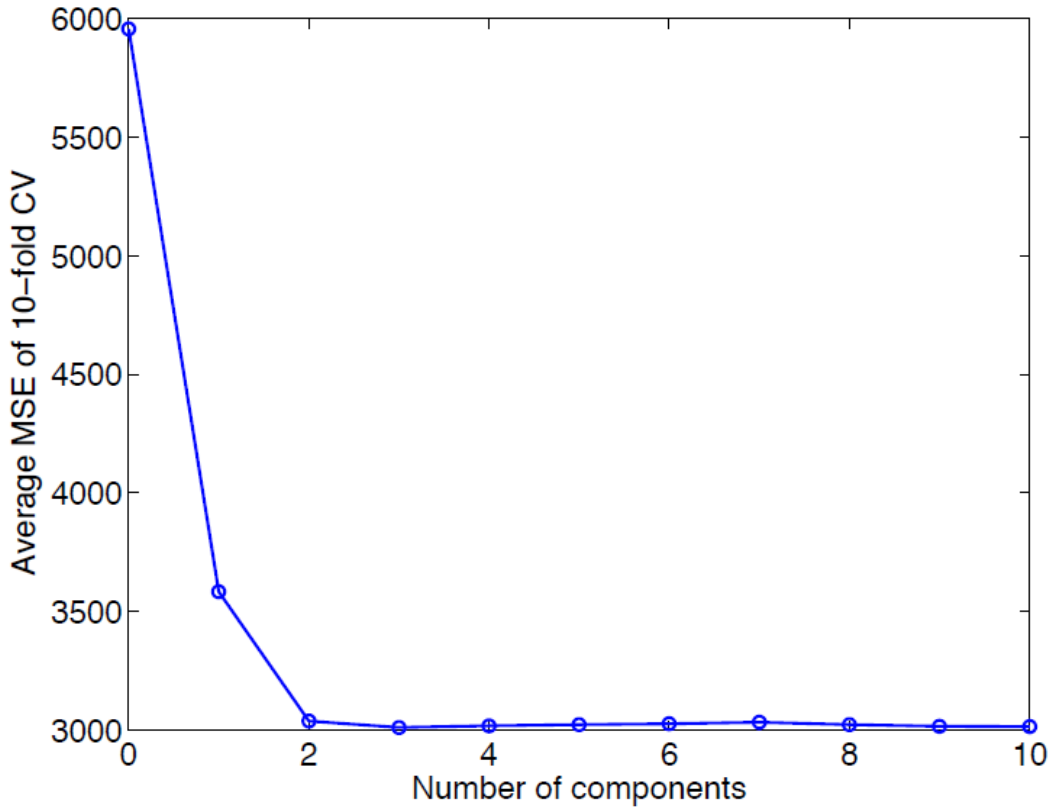
$$\max_{\alpha} Corr^2(y, X\alpha) Var(X\alpha)$$

$$\text{subject to } ||\alpha|| = 1,\ \alpha^T \Sigma \varphi_l = 0,\ l = 1, ..., m-1$$

- Behaves similar to ridge regression and principal component regression in shrinking coefficient estimates.
- Shrink low variance directions (like ridge)
- Can inflate high variance directions

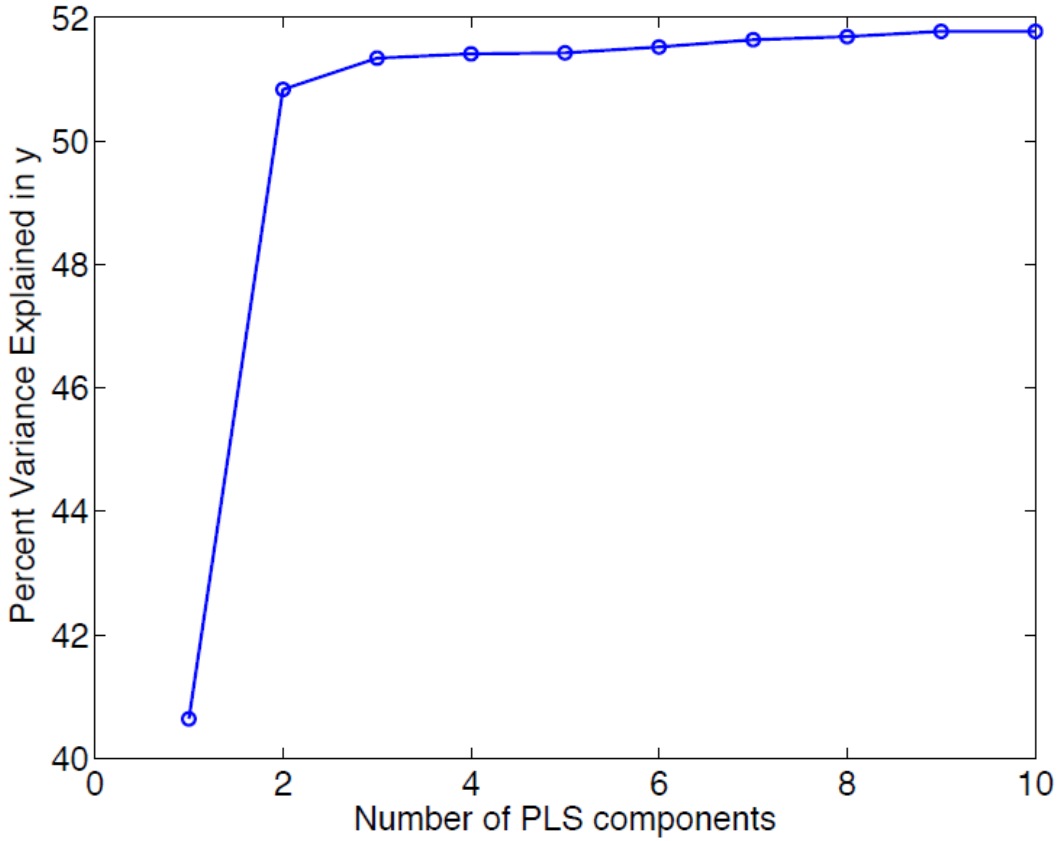
PLS on diabetes example

```
[XL,yl,XS,YS,beta,PCTVAR,MSE,stats]...
    = plsregress(X,y,10,'cv',10);
plot(0:10,MSE(2,:),'-bo')
```



PLS on diabetes example

```
plot(1:10,cumsum(100*PCTVAR(2,:)),'-bo');
```



PLS example on satisfaction survey

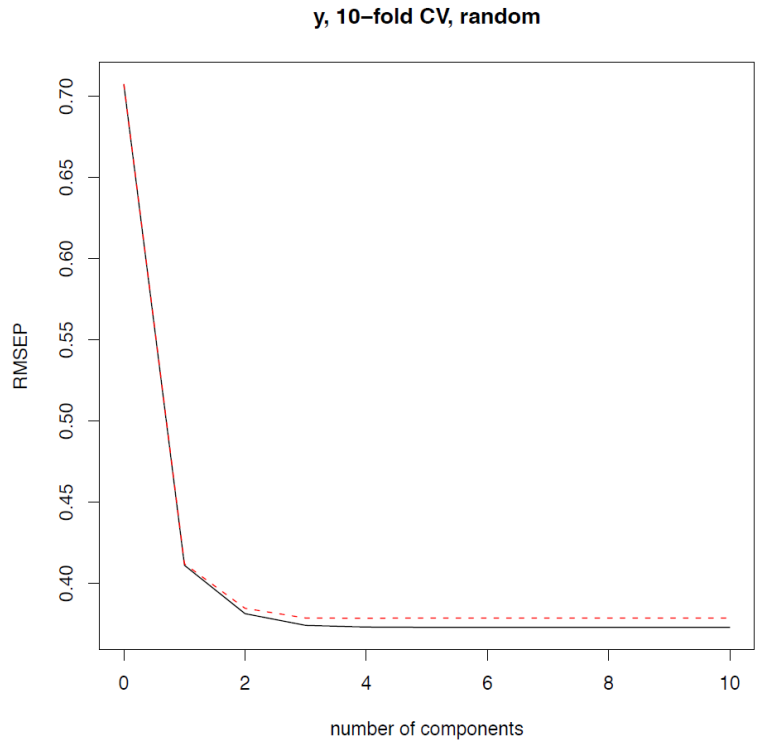
- The student environment survey conducted at DTU had approximately 2900 students fill out a questionnaire with more than 35 questions related to various aspects of student life at DTU rated on a scale from 1 to 6.

- The students general satisfaction was modeled using a PLS analysis as a function of questions related to the physical, the environmental and the psychological aspects of their student life.

PLS example on satisfaction survey

The number of PLS components was assessed using a 10-fold cross validation and the one-standard-error rule.

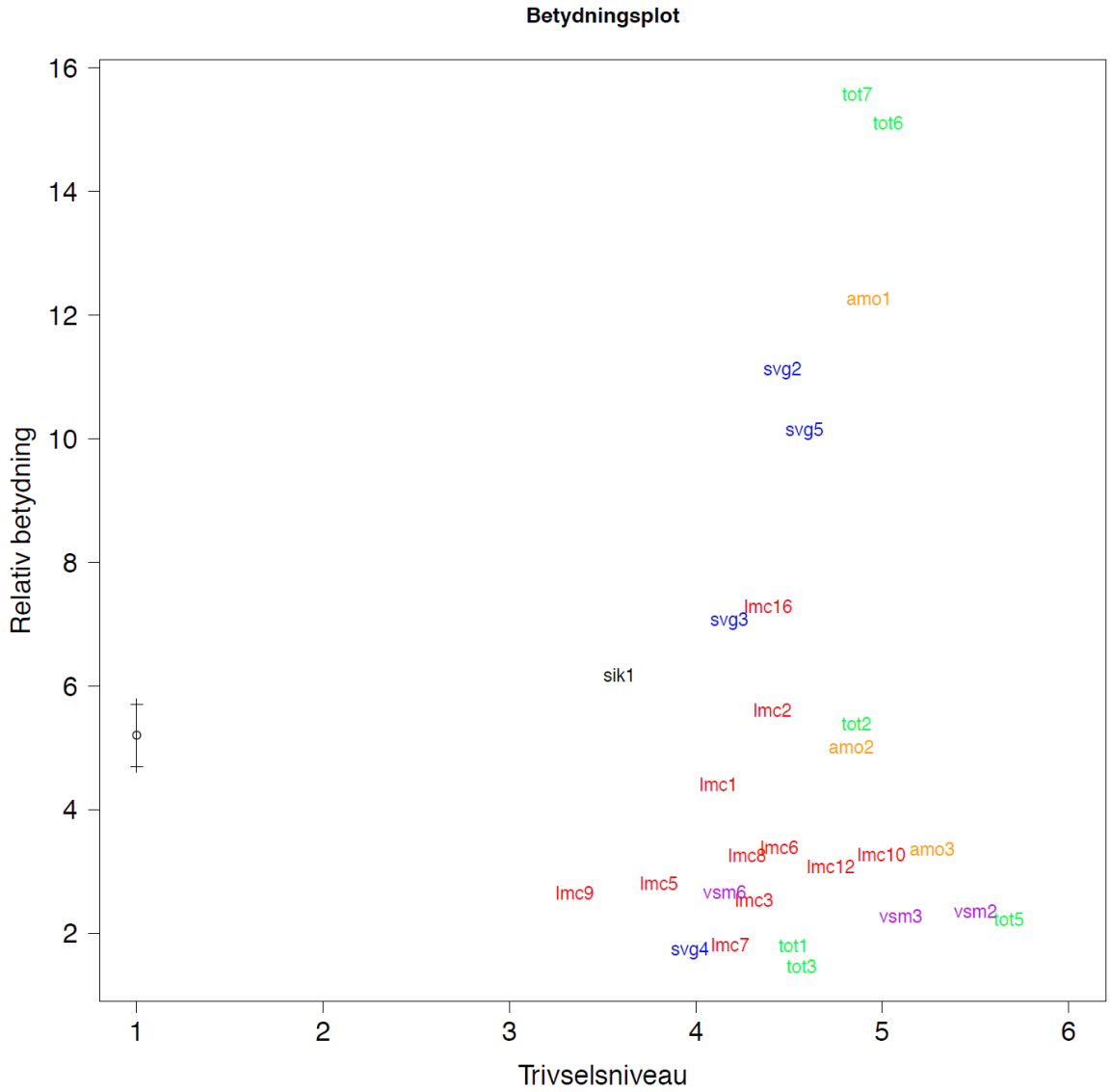
Resulting in 3 components,



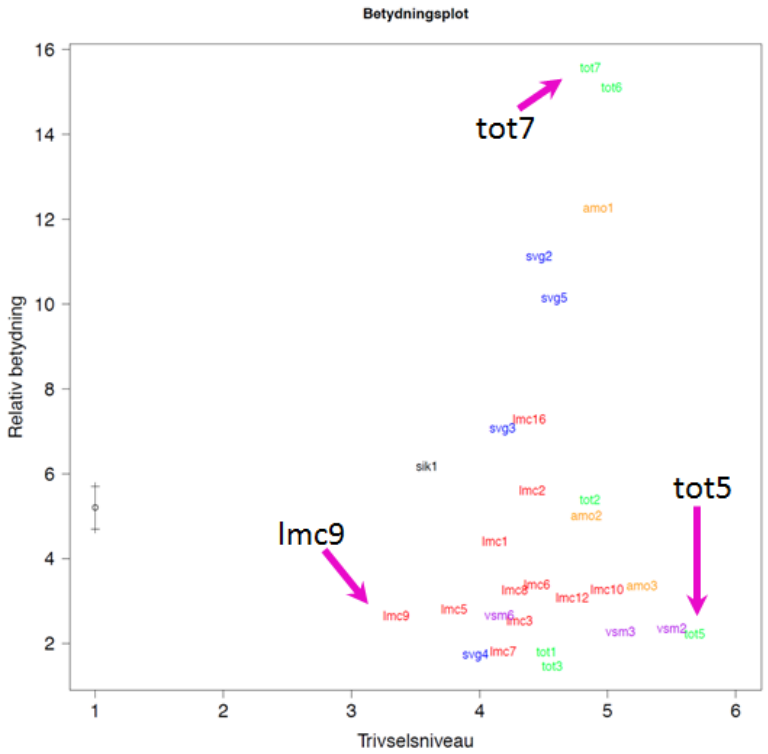
PLS example on satisfaction survey

- Using bootstrapping the variance of the coefficient estimates was estimated as s_j.
- The coefficients b_j were considered "significant" if the estimate divided by the variance was greater than 1.5, ie $b_j/s_j > 1.5$

PLS example on satisfaction survey



PLS example on satisfaction survey



tot7
Jeg føler mig sjældent ensom på DTU

tot5
Jeg er ikke udsat for mobning eller chikane fra underviser eller anden ansat

lmc9
Der er tilfredsstillende strømforsyning på campus

Sparse PLS

- A sparse version of PLS exists where the partial least squares directions are elastic net regularized.
- Implemented in the R package `spls`.

Chun, H., Keles, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society Series B, 72 (1), 3-25.

Canonical Correlation Analysis

- Finds associations between two data sets.

Canonical correlation analysis

The canonical correlation of the data matrices X and Y are given as,

$$\max_{u_m, v_m} Corr^2(Yu_m, Xv_m)$$

$$\text{subject to } u_m u_j = 0 \text{ and } v_m v_j = 0 \; m \neq j, m = 1, ..., M$$

The linear combinations are uncorrelated and there is at most the minimum dimension of the two data matrices.

CCA example - course evaluations

We examine the relations between DTU form A (the student evaluations of the course) and DTU form B (the student evaluations of the instructor)

Form A, 8 questions

Form B, 3 questions

We look at one DTU course for one specific semester.

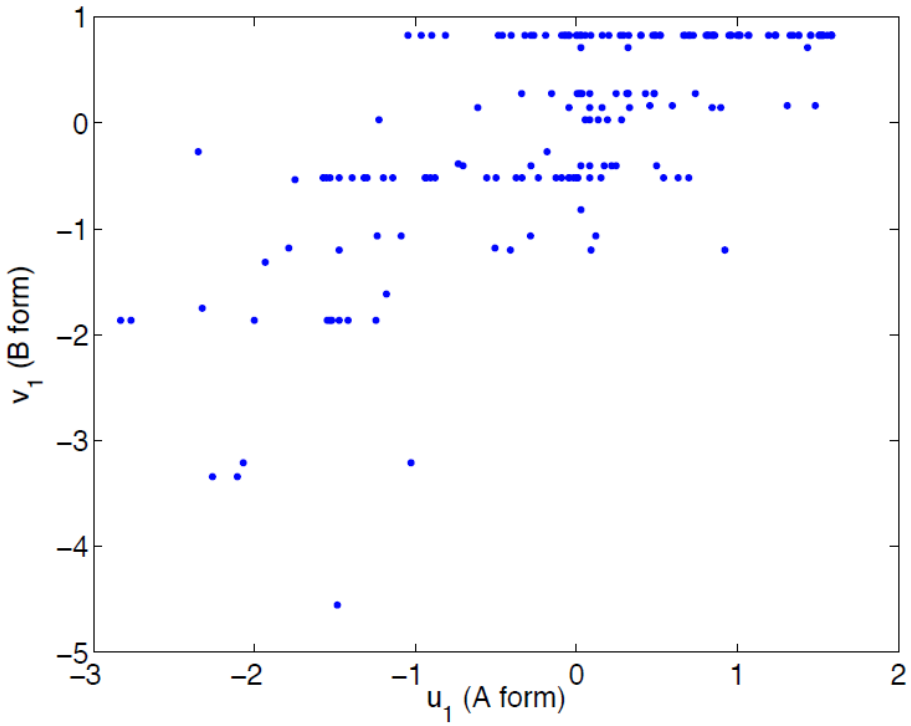
CCA example - course evaluations

```
% load data
load CourseEval
X = A; % A-form replies
X(:,7) = 5-abs(2*X(:,7)-6); % transform prerequisites
X(:,6) = 5-abs(2*X(:,6)-6); %transform English
Y = B; % B-form replies

% perform canoncial correlation analysis (CCA)
[A,B,R,U,V,STATS] = canoncorr(X,Y);
```

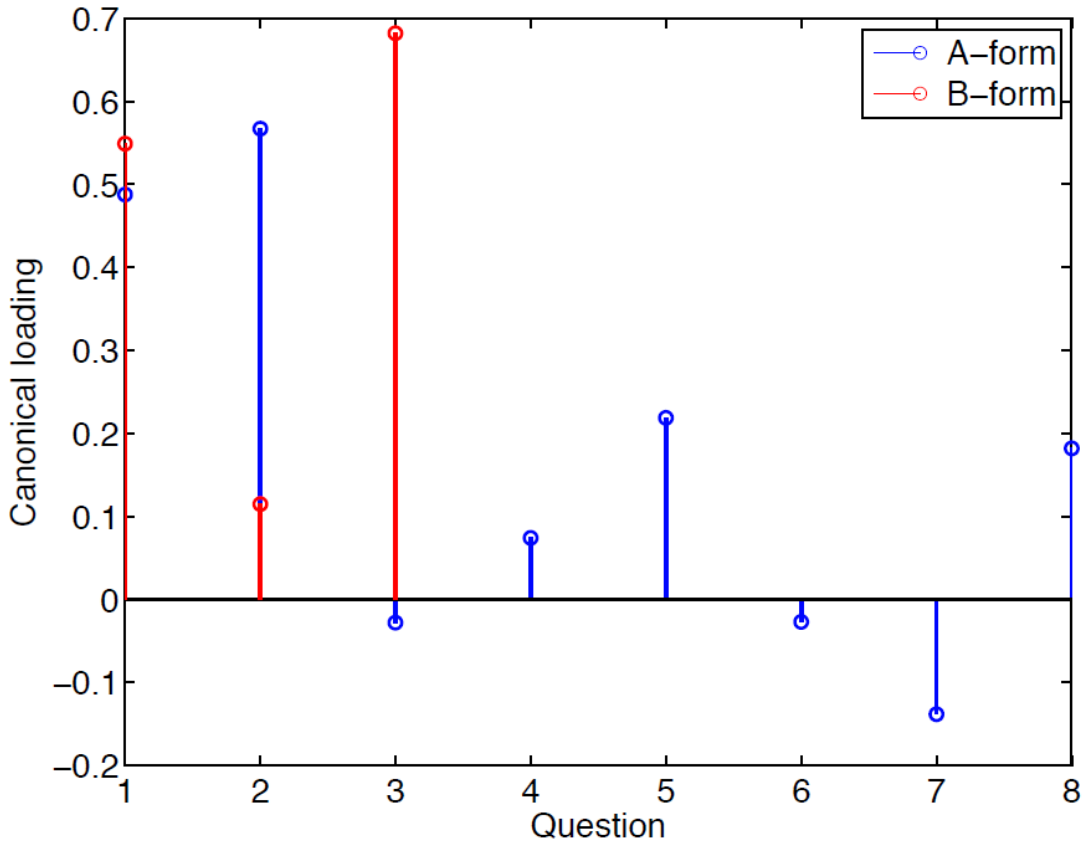
CCA example - course evaluations

```
% plot the first pair of canonical variates
plot(U(:,1),V(:,1),'.','MarkerSize',12)
```



CCA example - course evaluations

```
% plot the loadings of the first canonical variables
h1 = stem(A(:,1),'b'), hold on
h2 = stem(B(:,1),'r')
legend([h1 h2],{'A-form','B-form'})
```



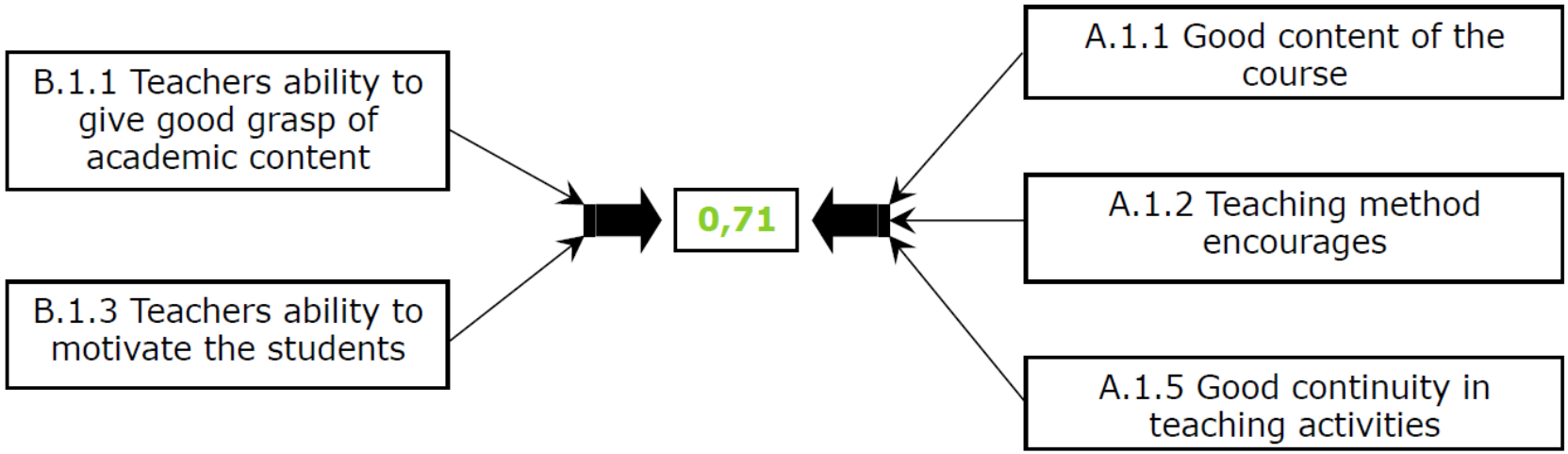
CCA example - course evaluations

There exist a test for the null-hypothesis that the correlation of the canonical variables is zero (when n>p)

```
R % The canonical correlations
>> 7.1195e-01 2.6177e-01 1.5497e-01
STATS.p % The p-values
>> 1.9962e-18 2.8024e-01 5.4442e-01
```

The first set of canonical variables are significantly correlated

CCA example - course evaluations



There is a strong association between how students evaluate the course and how students evaluate the teacher

Sparse CCA

- Several regularized and sparse versions of CCA exist
- Ridge regularization: Regularized CCA (Vinod 1976, Leurgans et al 1993)
- Sparse version: Penalized Matrix Decomposition (Witten et al, 2009)

Exercises

- Shape analysis of faces
- PCA inside-and-out
- Sparse PCA
  - Thresholding
  - Varimax rotation
  - Simple Elastic Net version
- PLS on sand data

Exercise - Shape analysis of faces

1. Apply Principal Component Analysis (PCA) to the face data set in *faces.mat*
   (a) Load the face shape data in the file *faces.mat*
   (b) Compute the mean shape and center the data. Plot the mean shape (see details in exercise file)
   (c) Compute a principal component analysis of the data. Try using both an eigen value decomposition (EVD) and a singular value decomposition (SVD). Remember that the EVD is computed on the correlation or covariance matrix and the SVD on the data matrix itself. See calculation hints in exercise file.

Exercise 1 - Shape analysis of faces

PCA continued - illustrations

(d) Plot the first mode of variation. This will provide a view of the most important variation in the data set. Let the mean face be the origin, we will use the mean face as a reference. Plot the mean face in black.

- Compute the face obtained by moving from the mean face along the first principal axis (first column of the loading matrix) out to a distance of +2.5 standard deviations ($\mu + 2.5\sigma_{l_1}$). The standard deviations can be obtained from the singular values (see the slides on how this is done). Plot the resulting face in red.
- Repeat this procedure to obtain a face at -2.5 standard deviations from the mean face. Plot this in blue.

(d) Explore the first few modes of variation using the provided shape inspector.

Exercise 2 - Sparse PCA

Extract Sparse Principal Components for the face data using **three different methods**:

(a) Compute a sparse PCA by **thresholding** all loadings from a regular PCA with absolute value less than 0.15.
- Now that the loading matrix has been changed, the scores matrix must be recomputed. How can you use this new scores matrix to compute the variance of the data along each principal axis?
- Use this result to plot the first mode of variation of the threshold SPCA for -2.5, 0 and +2.5 standard deviations.
- Investigate what has happened with the uncorrelatedness of the loadings and scores matrices of regular PCA. Are these properties fulfilled here?

(b) Use the **Varimax criterion**. Experiment with the number of columns from the loading matrix to rotate, and see how this affects sparsity.
- Again, what has happened with the uncorrelatedness of the scores and loading matrices? See the code listings for functions which perform Varimax rotation.

(c) Compute the first sparse principal loading vector using the **Elastic net** (remember to normalize to unit length).
- Start by using 10 non-zero loadings. Plot this solution and try different number of nonzero components.
- Can you put a meaningful anatomical label on this deformation?
- Would you be able to label the first mode of variation from regular PCA?

Exercise 3 - PLS on sand data

Apply Partial Least Squares regression to the sand data. *(Only a Matlab solution provided)*

a. Load data *sand.mat* and run a cross validation of partial least squares regression to decide the number of components that is adequate to model the sand data. Plot both the cross validation error and the percentage of explained variance in y to determine the number of components. See the code listings for useful PLS implementations.

b. How would you plot the coefficients of the final PLS regression model (β)? Which variables are important for the prediction of y? (In terms of loadings, remember the pitfall of PCA concerning scaled vs non-scaled loadings - this also holds for PLS).