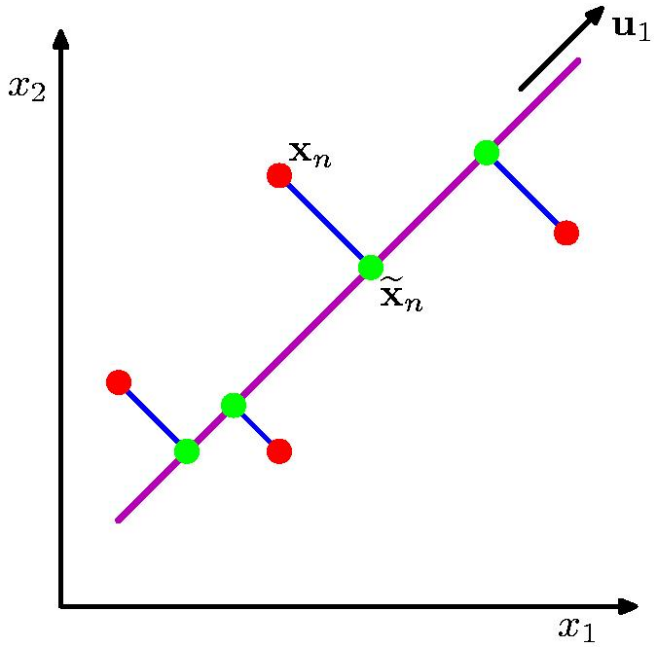
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Principal Component Analysis (PCA)
Dimensionality and noise reduction

Kim Steenstrup Pedersen

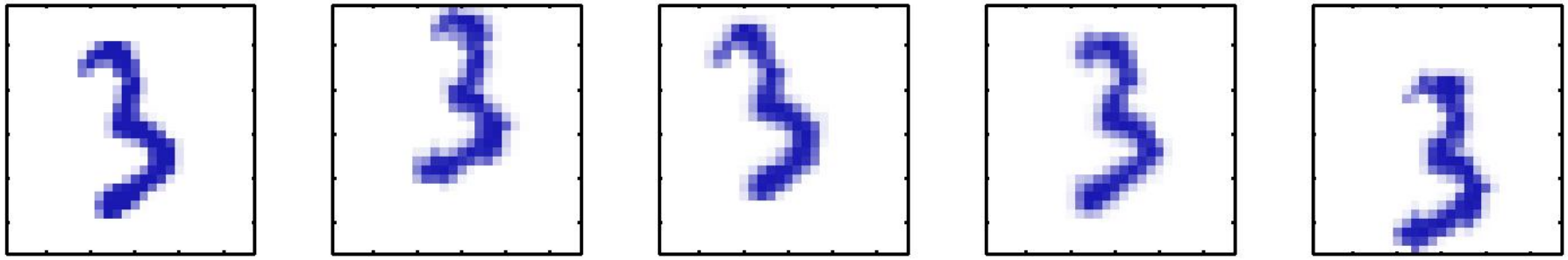
Plan for this lecture

- Continuous latent variable models:
  - Why linear PCA?
- Principal component analysis (PCA):
  - Maximum variance formulation
  - Minimum error formulation
- Applications of PCA:
  - Preprocessing of data for noise reduction
  - Dimensionality reduction
  - Visualization of high dimensional data
- And some computational tricks



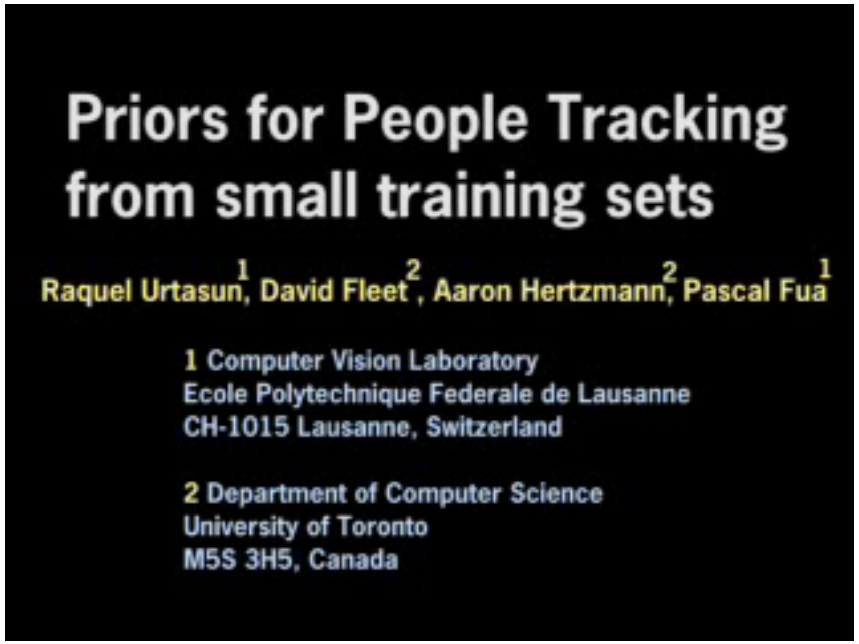
Continuous latent variable models

- Often high dimensional data have few degrees of freedom, i.e. a low intrinsic dimensionality.
- Example: Images of hand written digits



- Dimensionality: 64 x 64 pixels = high dimensionality
- Intrinsic degrees of freedom (< 64 x 64):
  - **Easy:** Translation (2), rotation (1)
  - **Complicated:** Degrees of freedom coming from the variability in how to write the digit 3.
  - Not all images represents valid digits – the set of digit images is sparsely distributed in the space of images.

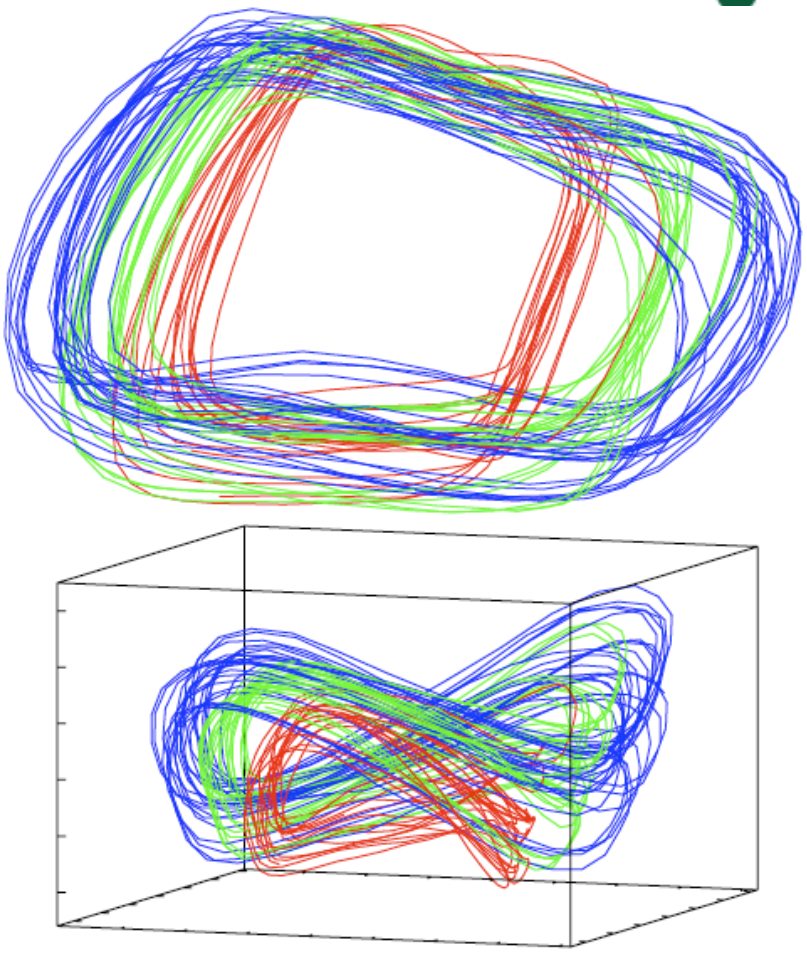
Continuous latent variable models : Walking occupies a low (3D?) dim. torus



[Urtasun et al, 2005b]

Dimensionality: 15 joints x 3 (3D)

The intrinsic dimensionality of walking is much lower



[Jaeggli et al, 2009]

Continuous latent variable models

- We can model the degrees of freedom as latent (hidden) variables **z**.
- The connection between data representation **x** and latent variables **z** is in general some non-linear mapping:

$$\mathbf{x} = \varphi(\mathbf{z}, \varepsilon)$$

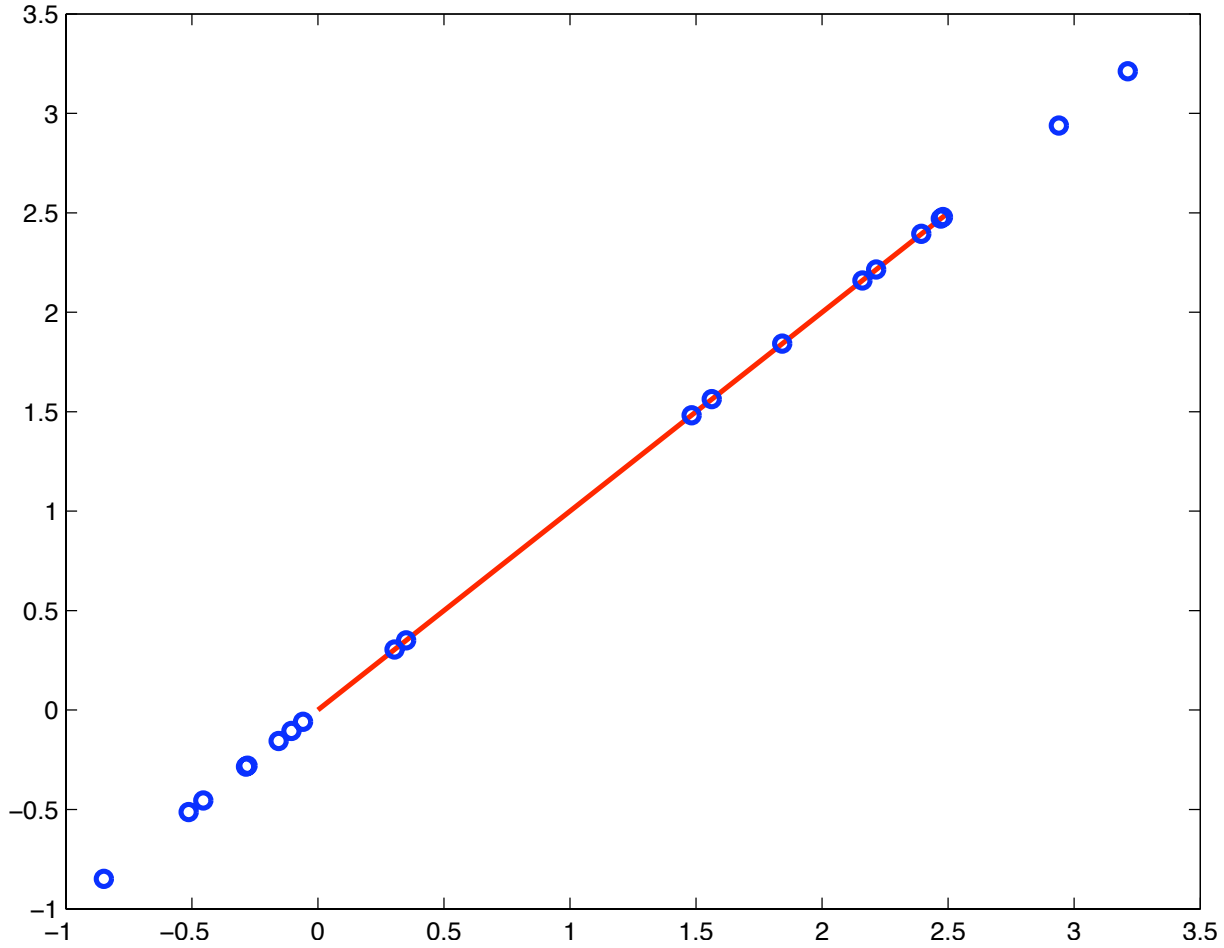
  Including some noise ε .
- The simplest choice is a linear model with additive noise:

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b} + \varepsilon$$

- Principal component analysis (PCA) is based on a linear model.

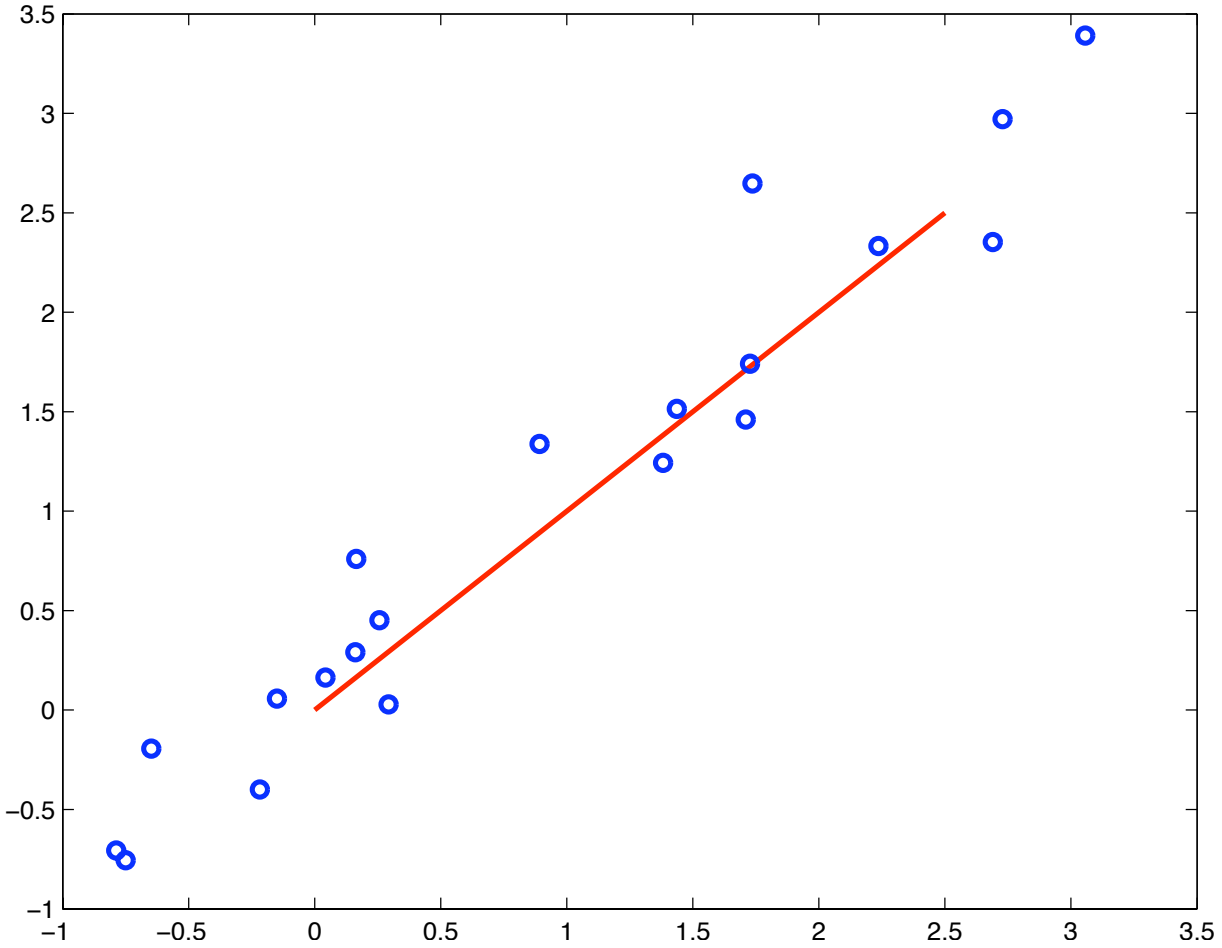
Continuous latent variable models : A synthetic linear 1D latent variable example

Data is 2D but it only has 1 degree of freedom and lies along a line (linear subspace)



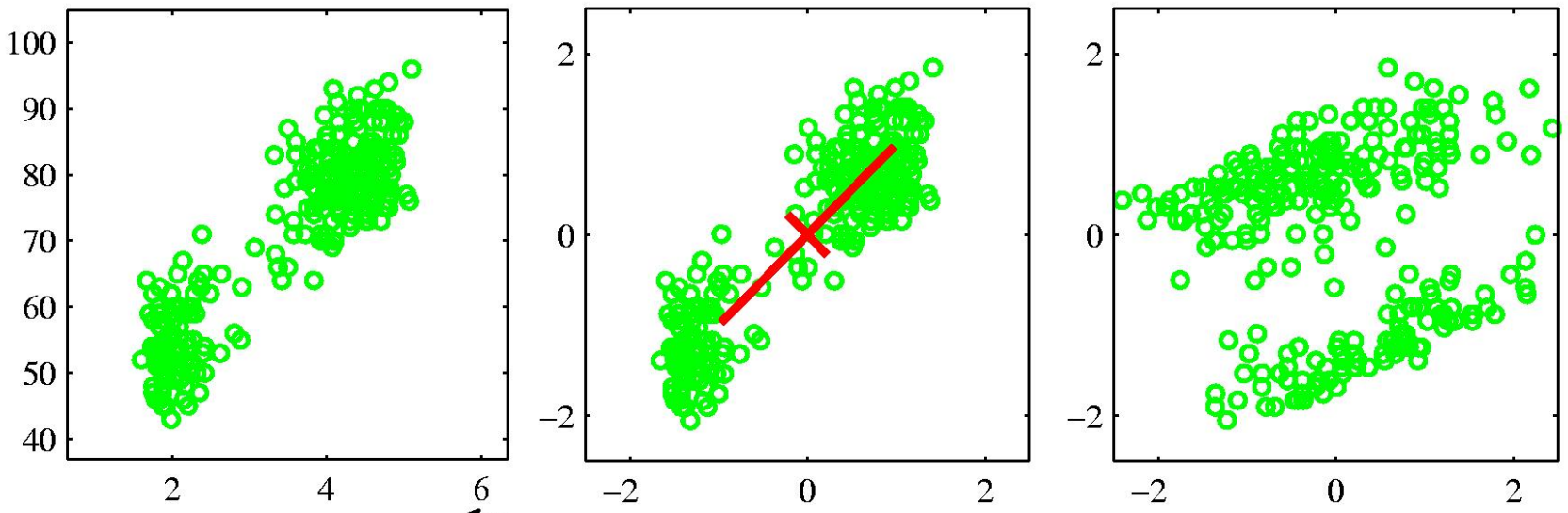
Continuous latent variable models :
In reality we often encounter data like this

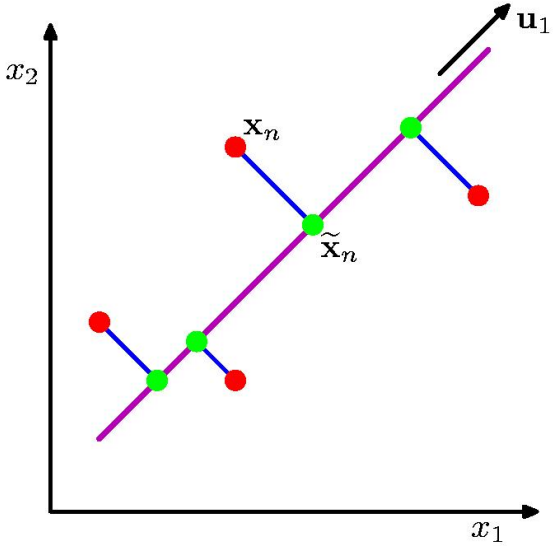
A bit more messy but data can still be approximated with a linear model:



$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b} + \boldsymbol{\varepsilon}$$

Principal Component Analysis (PCA) : The typical steps





1. Find principal directions
2. transform into this coordinate system

Principal directions?

- Either maximize variance along $\mathbf{u}_1$
- Or choose $\mathbf{u}_1$ to minimize projection error

It turns out that they are equivalent formulations!

PCA : Maximum variance formulation
(The first)

Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M \ll D$.

We find directions sequentially, $\mathbf{u}_1$ first.

Mean of projected data: $\mathbf{u}_1^T \bar{\mathbf{x}}$ with

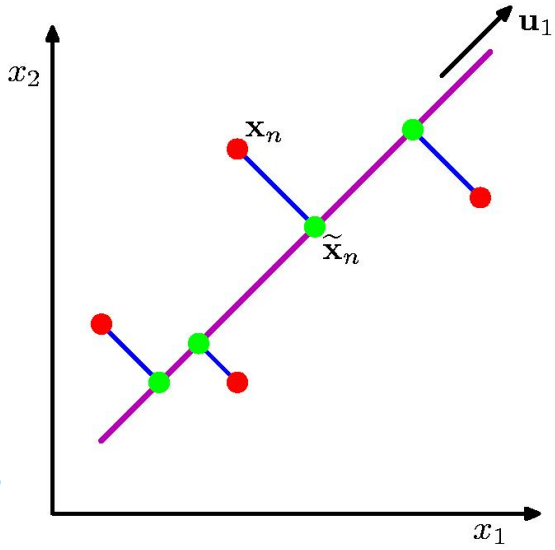
$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \ .$$

Variance of projected data:

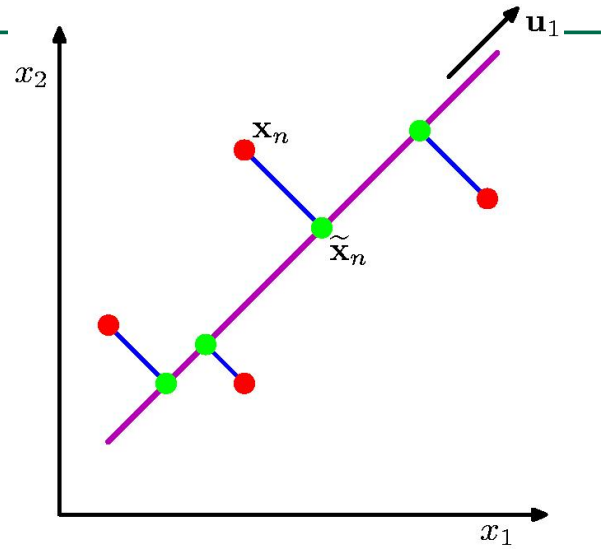
$$\frac{1}{N} \sum_{n=1}^{N} \left\{ \mathbf{u}_1^T \left(\mathbf{x}_n - \bar{\mathbf{x}} \right) \right\}^2 = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 \ .$$

with the empirical co-variance

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}_n - \bar{\mathbf{x}} \right) \left(\mathbf{x}_n - \bar{\mathbf{x}} \right)^T$$



PCA : Maximum variance formulation
(The first)



Maximize variance

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$$

with respect to $\mathbf{u}_1$.

But we need a constraint to avoid $\mathbf{u}_1 \to \infty$:

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)$$

Student exercise – take derivative wrt $\mathbf{u}_1$ and show

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \ .$$

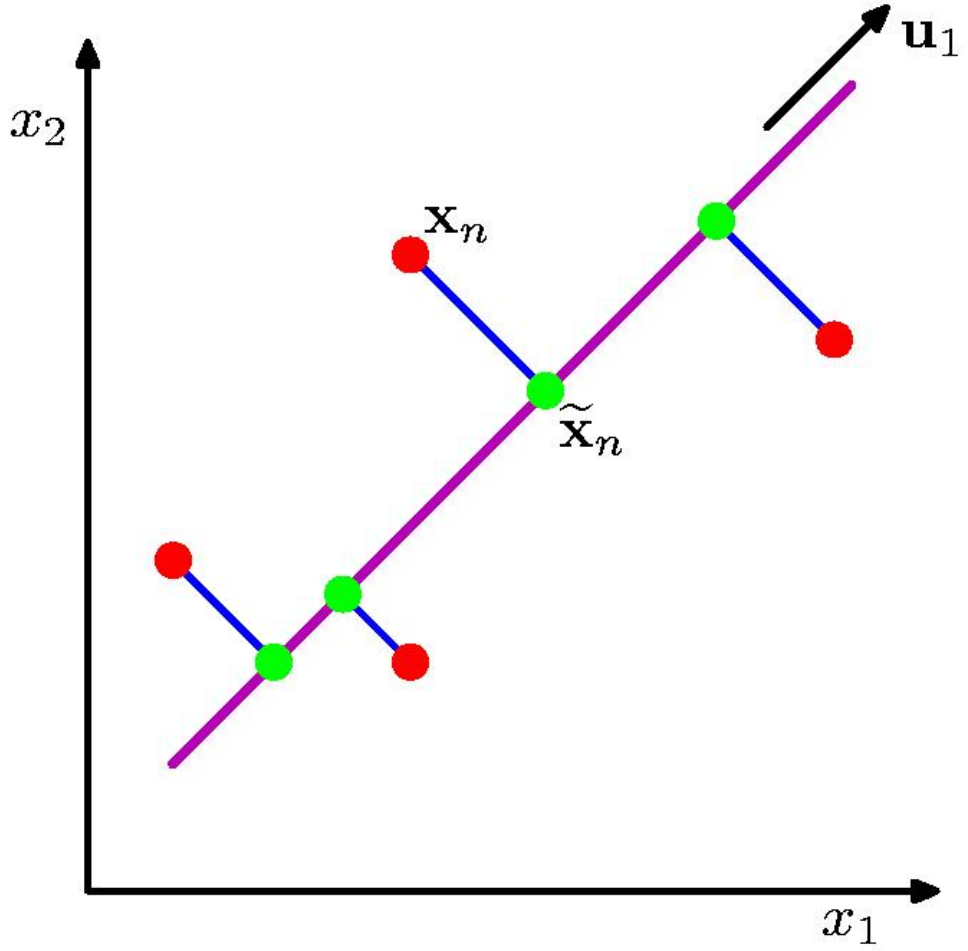
Multiply by $\mathbf{u}_1^T$ and show

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \lambda_1 \ .$$

How to interpret these two results?

Recall: The eigenvectors are orthonormal and form a coordinate system

PCA : Minimum error formulation
(The second)



Goal: Find the optimal reconstructing orthonormal directions $\{\mathbf{u}_i\}$

PCA : Minimum error formulation
(The second)

- Assume complete orthonormal basis given: $\{\mathbf{u}_i\}$, $i = 1,\ldots,D$

$$\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij} \qquad \delta_{ij} = 1 \text{ if } i = j, \text{ otherwise } \delta_{ij} = 0$$

- In this basis, data $\mathbf{x}_n$ may be represented as $(z_{n1}, z_{n2}, \ldots, z_{nD})$

$$z_{ni} = \mathbf{x}_n^T \mathbf{u}_i \text{ (projection on PC)} \qquad \text{And back: } \mathbf{x}_n = \sum_{i=1}^{D} z_{ni}\mathbf{u}_i$$

- Goal: Approximate data with fewer dimensions $M < D$ by an M dimensional linear subspace:

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{j=M+1}^{D} b_j \mathbf{u}_j$$

$$z_{ni} = \mathbf{x}_n^T \mathbf{u}_i \quad , \quad i = 1,\ldots,\mathrm{M}$$

$$b_j = \bar{\mathbf{x}}^T \mathbf{u}_j \quad , \quad j = M+1,\ldots,D$$

- By minimizing the sum of squares error:

$$J = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \tilde{\mathbf{x}}_n\right)^2 = \sum_{j=M+1}^{D} \mathbf{u}_j^T \mathbf{S} \mathbf{u}_j \quad \text{subject to constraint } \mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$$

PCA : Minimum error formulation
(The second)

- Minimizing the sum of squares error:

$$J = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \tilde{\mathbf{x}}_n\right)^2 = \sum_{i=M+1}^{D}\mathbf{u}_i^T\mathbf{S}\mathbf{u}_i \quad \text{subject to constraint } \mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$$

- It can be proven that the general solution is given by the eigenvector equation: $\mathbf{S}\mathbf{u}_i = \lambda_i\mathbf{u}_i$
- Also the corresponding error for this solution is

$$J = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \tilde{\mathbf{x}}_n\right)^2 = \sum_{i=M+1}^{D}\lambda_i$$

- Hence for fixed *M* choose the eigenvectors corresponding to the *M* largest eigenvalues to minimize the sum of squares reconstruction error.

PCA : Summary

- The two formulations are equivalent.
- Solve the eigenvector equation for the data covariance:

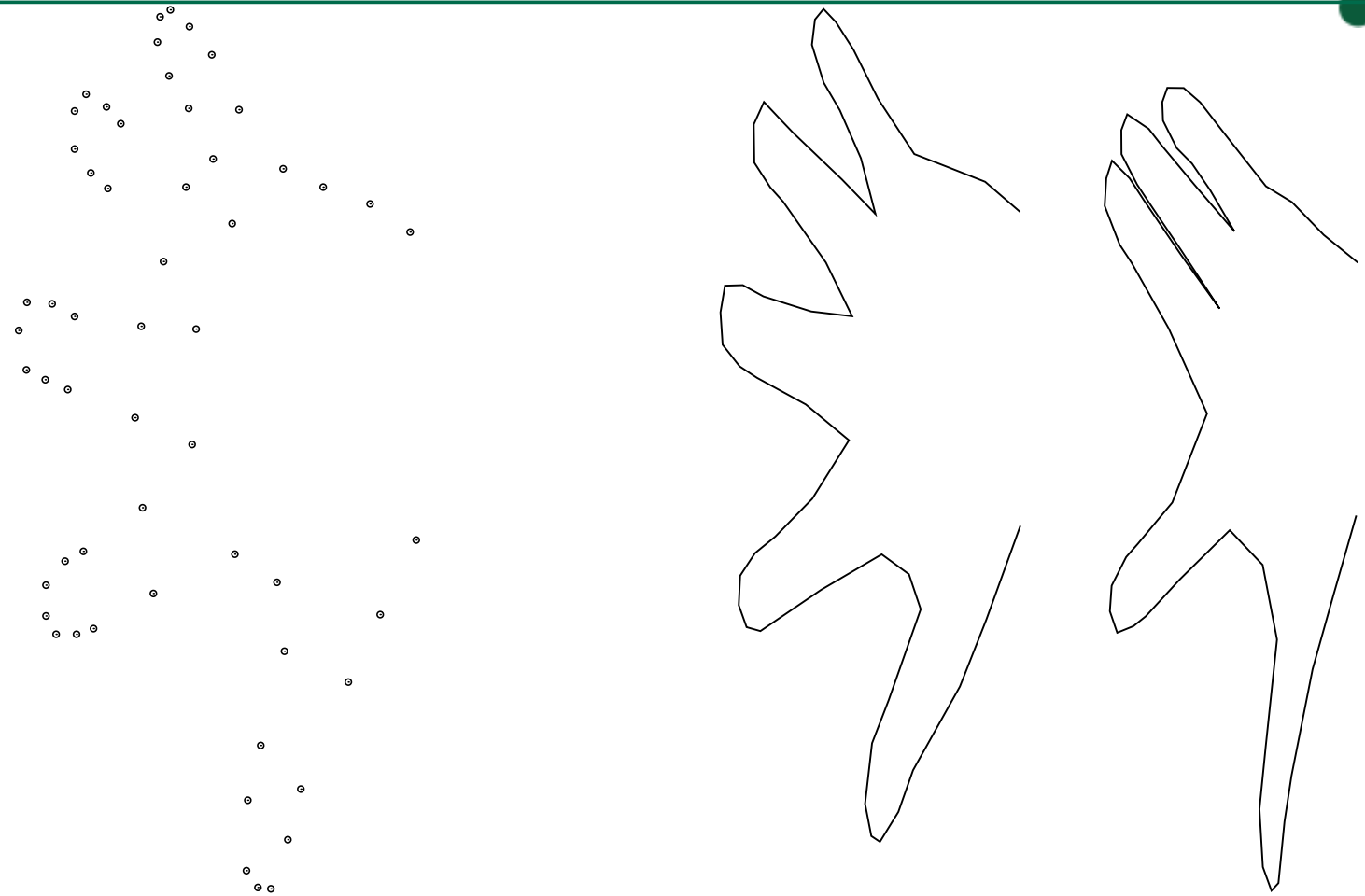
$$\mathbf{S}\mathbf{u}_i = \frac{1}{N}\mathbf{X}^T\mathbf{X}\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

  Where the data matrix is $\mathbf{X} = \left(\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_N - \bar{\mathbf{x}}\right)^T \in R^{N\times D}$
- Eigenvectors form the Principal Components (PC) and an orthogonal coordinate basis.

  $z_{ni} = \mathbf{x}_n^T\mathbf{u}_i$ (projection on PC) , $\mathbf{x}_n$ in PC space $(z_{n1}, \ldots, z_{nM})^T$
- Eigenvalues represents the projected data variance along the corresponding PC.

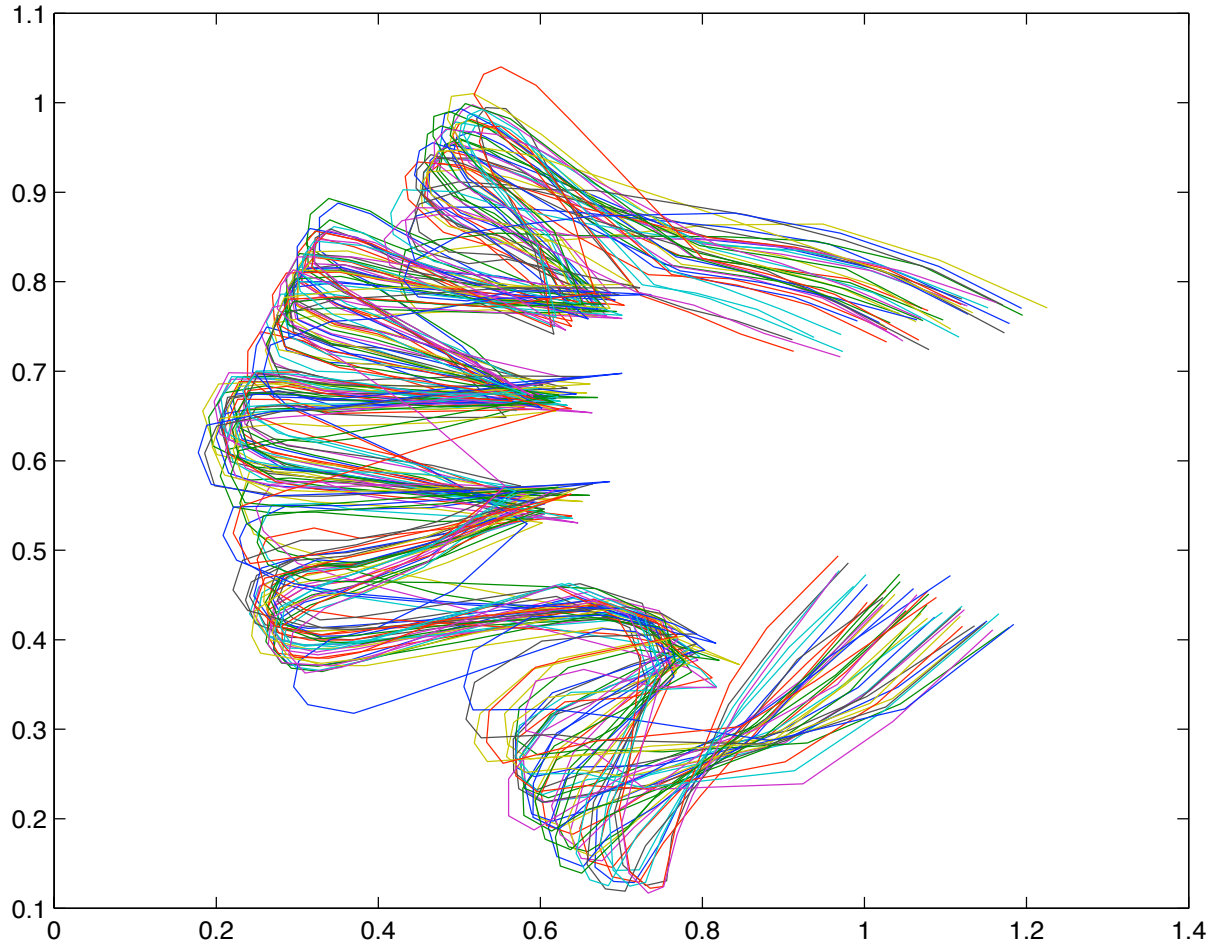
- Now what can this be used for?

The IMM-DTU hand data set



$$\mathbf{X} = \left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)^T \in R^{N\times D} \ , N = 40 \ , D = 2\times 56 = 112$$

The hand data set:
Visualization of all *N* data points as hands in 2D

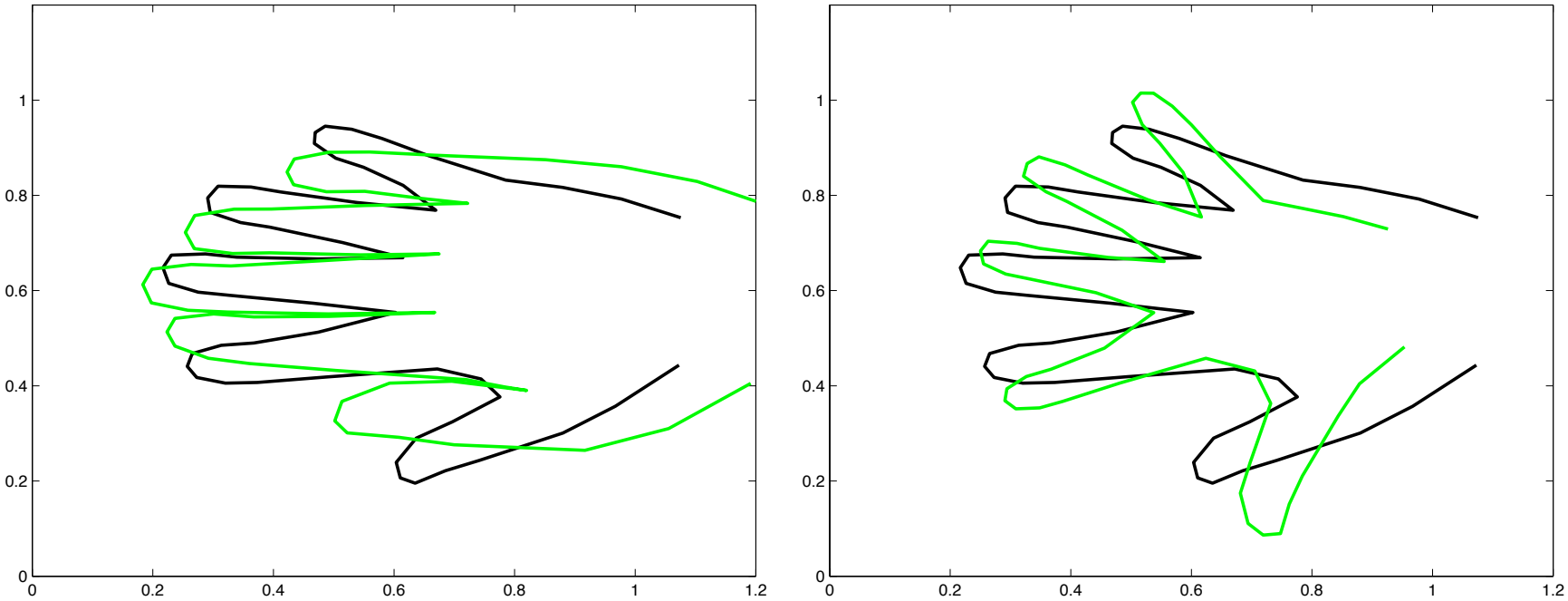


Visualizing the first principle component

Black: Mean data point (mean hand) $\bar{\mathbf{x}}$

Green: Visualizing +/- 2 standard deviations along first PC $\bar{\mathbf{x}} \pm 2\sqrt{\lambda_1}\mathbf{u}_1$

Conclusion: PC 1 captures finger spread variation and size

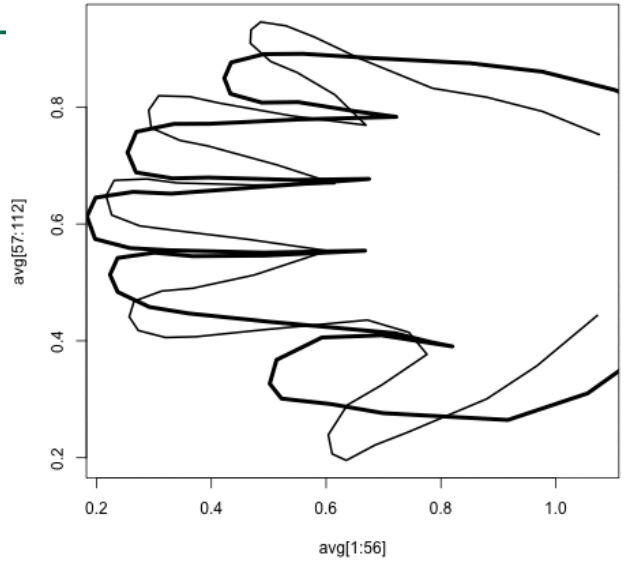


PCA can be used to visualize the largest / important variation in the data set.

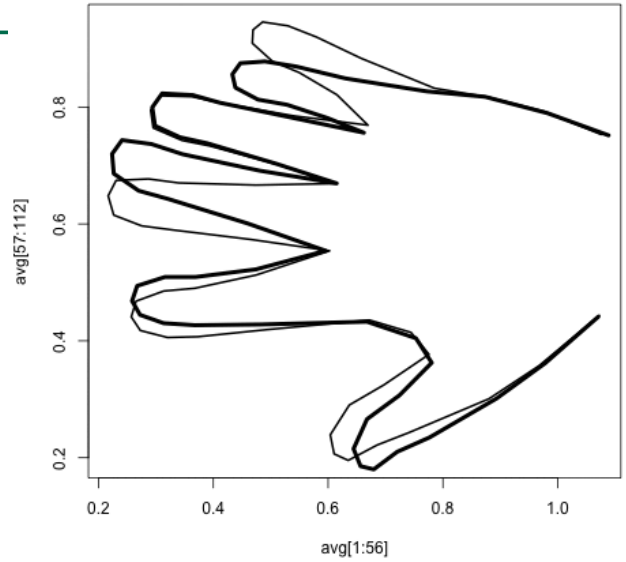
Visualizing the first principle components

Videos by R. Fonseca http://www.diku.dk/hjemmesider/ansatte/rfonseca/HandPC/

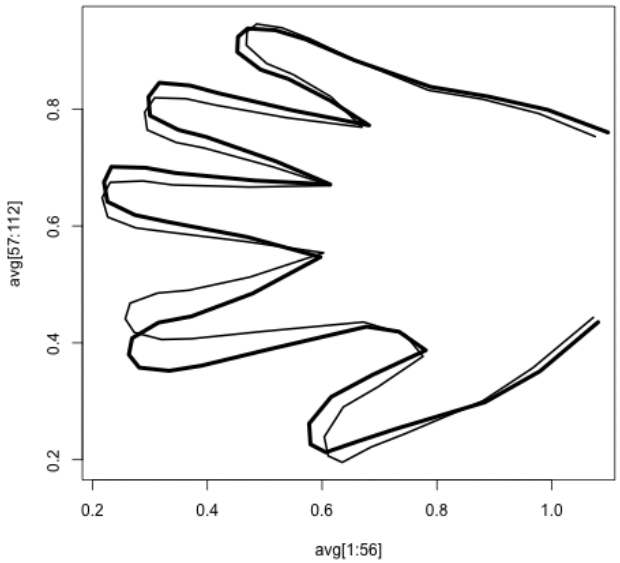
PC1



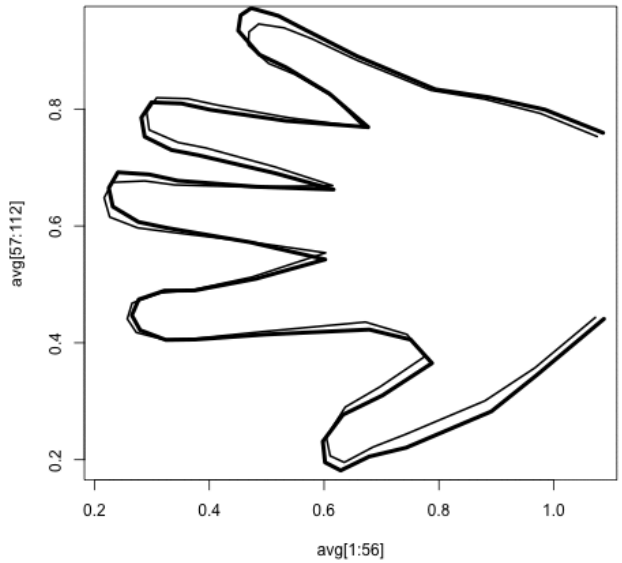
PC2



PC3



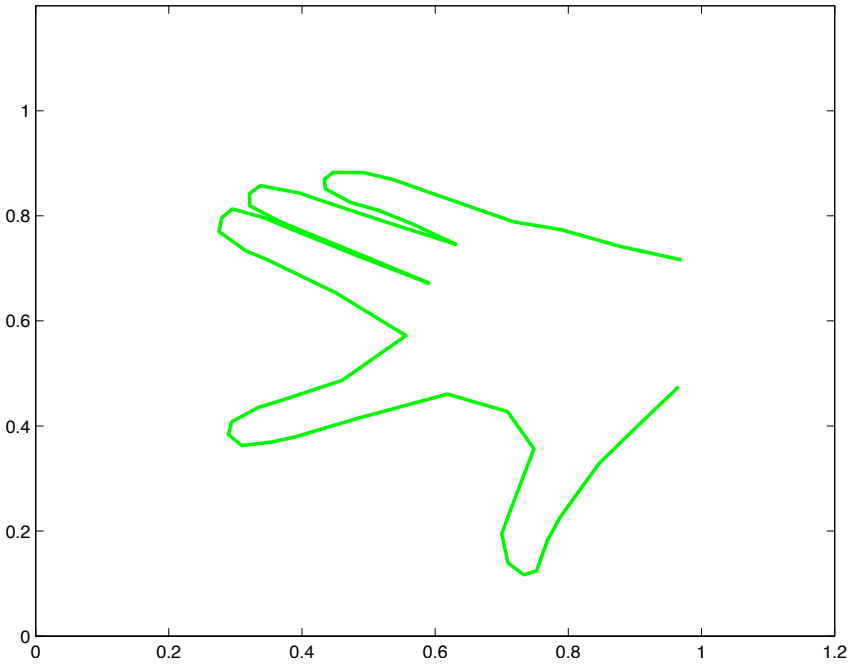
PC4

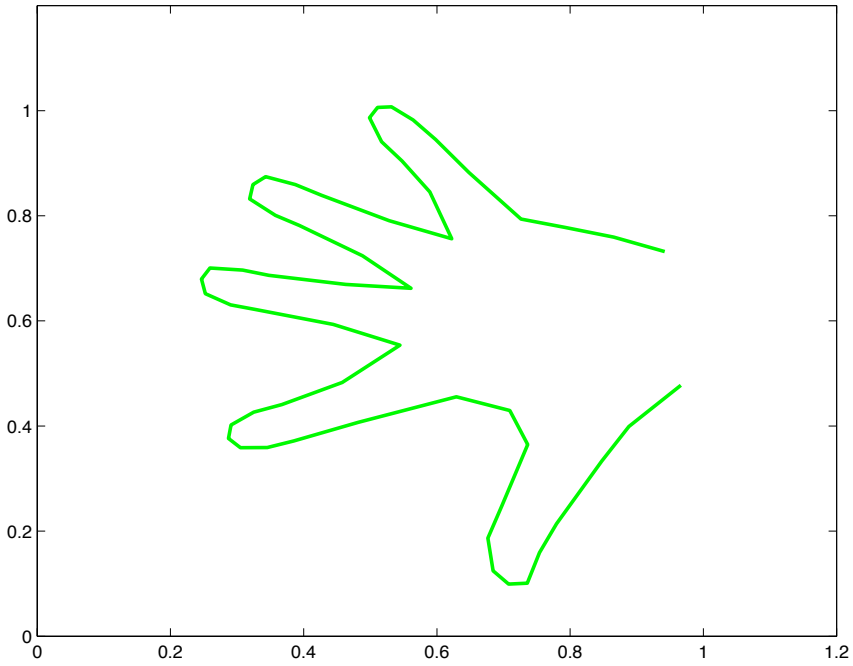


Using PCA for filtering (noise reduction): Find the closest hand with all fingers spread

Outlier hand $\mathbf{x}_i$

Projected onto PC1 $\left(\mathbf{x}_i^T \mathbf{u}_1\right)\mathbf{u}_1 + \bar{\mathbf{x}}$





Filtering: We remove irrelevant variation by projection onto principal components. In this example we concentrate on the spread of all fingers.

Remember the body fat data set from earlier ?
A regression problem

- **Data set (measurements from *N*=252 men):**
  - Density determined from underwater weighing
  - Percentage body fat from Siri's (1956) equation
  - Age (years)
  - Weight (lbs)
  - Height (inches)
  - Circumferences (cm): Neck, Chest, Abdomen 2, Hip, Thigh, Knee, Ankle, Biceps (extended), Forearm, Wrist
- *Observations* **x**: circumferences, weight, age and height (*D*=15 dimensional vector)
  Target values t: Percentage body fat (scalar)
  Data set: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T \quad , \quad \mathbf{T} = (t_1, \ldots, t_N)^T$

Percentage body fat data set revisited : PCA as dimensionality reduction

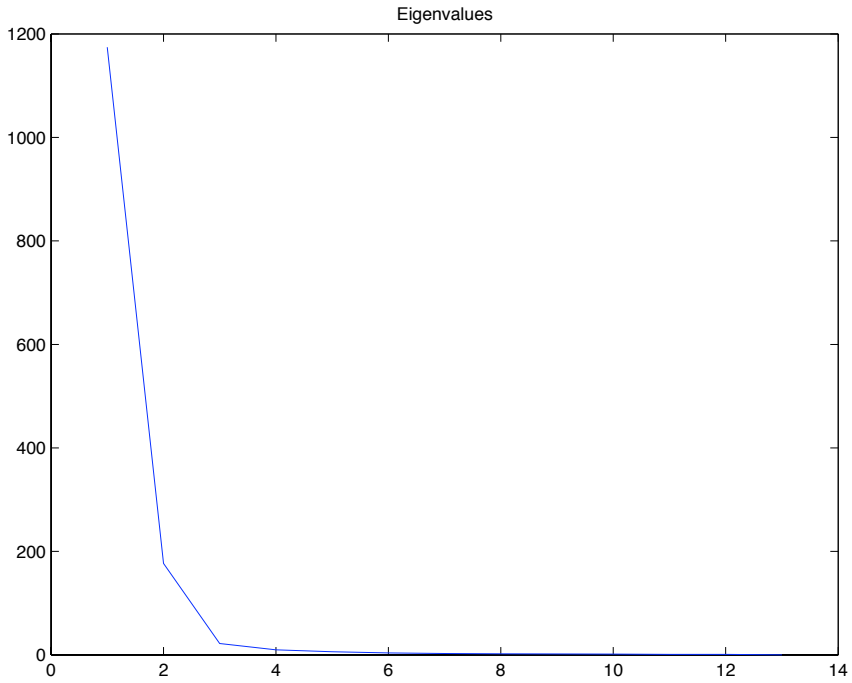
- Picking features by hand, Ex.:
  - **Selection 1 :** Columns 4 (weight), 7 (chest), 8 (abdomen), and 9 (hip)
  - **Selection 2 :** Column 8 (abdomen)

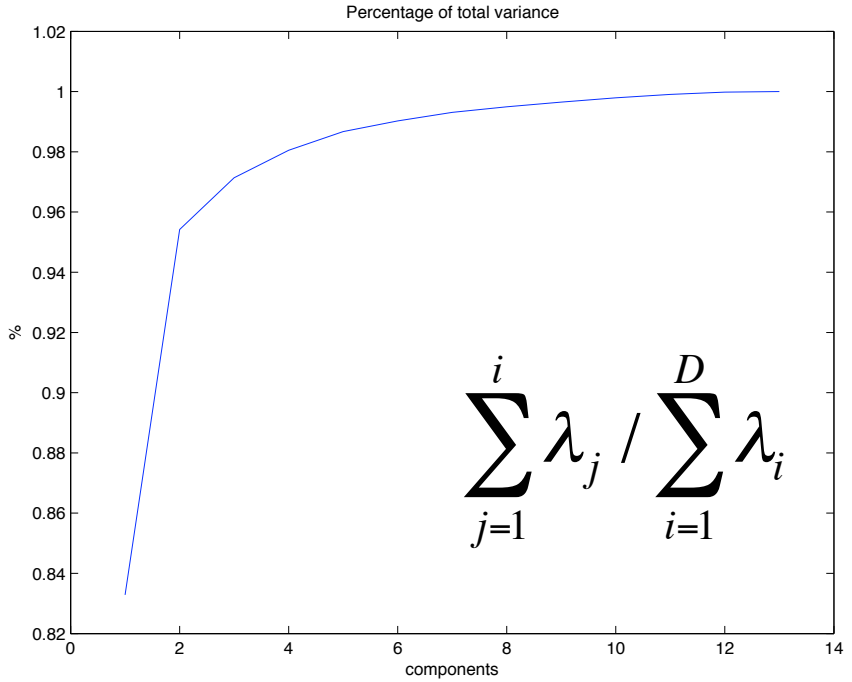
- This is based on intuition and picking variables with the largest covariance with the target variable in column 2 (percentage body fat).

- Lets see what we get out of performing PCA on this data set (dimensionality reduction).
  - Do PCA on columns 3 to 15, that is perform PCA on

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T \in R^{N \times D} \ , N = 252 \ , D = 13$$

Eigenvalues and total variance

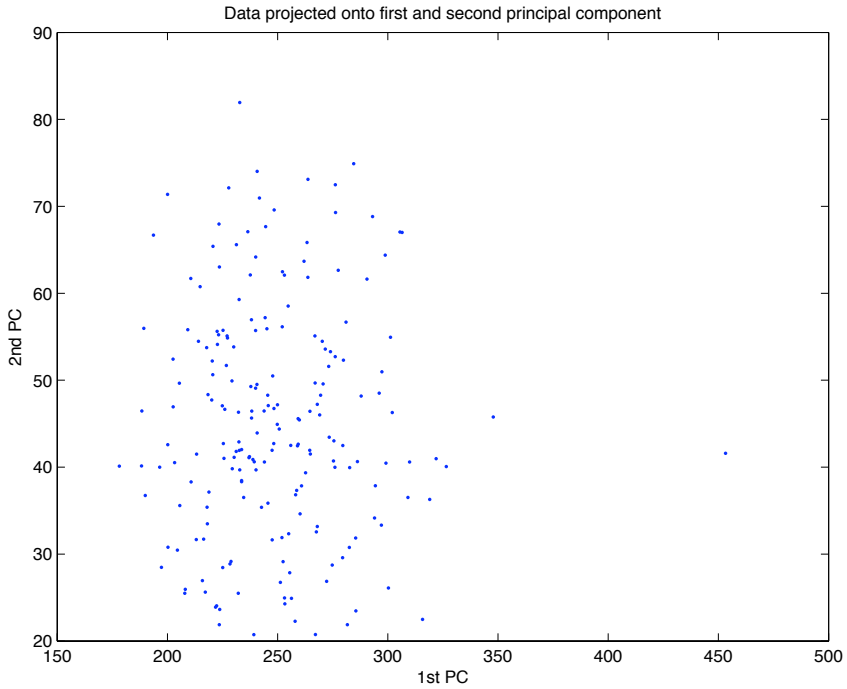




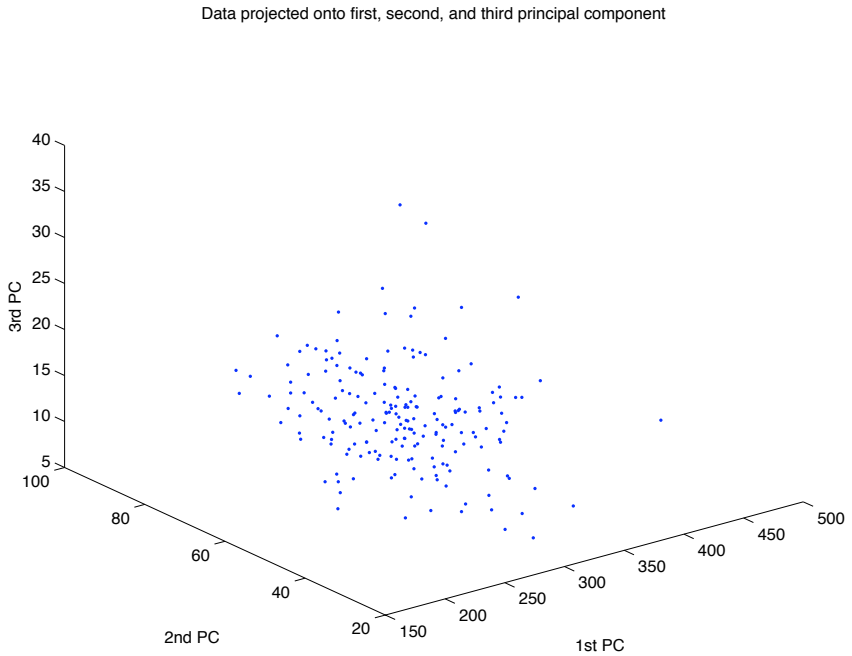
$$\text{Total variance} = \sum_{i=1}^{D} \lambda_i$$

PCA for columns 3 to 15: Projections of data

Data on PC 1 and 2

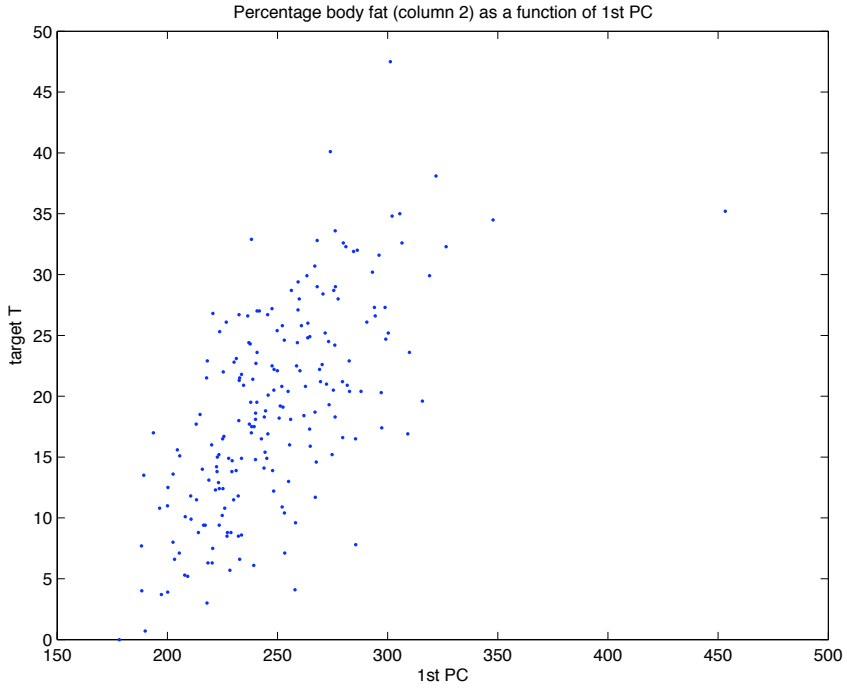


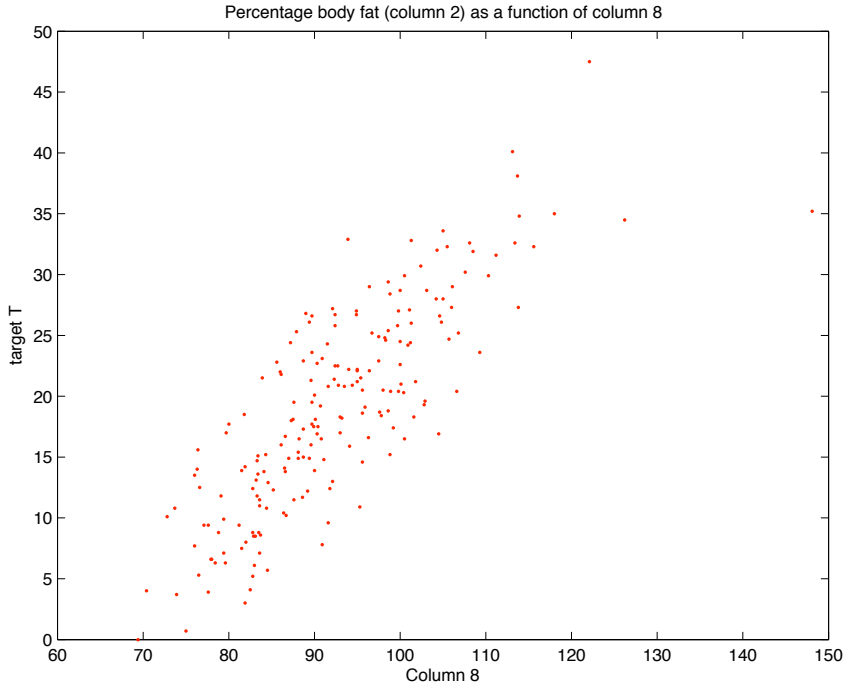
Data on PC 1, 2, and 3



What dimensionality for the latent variable should we choose? 1, 2, or 3?

Target variable plotted against respectively 1st PC and column 8 (abdomen)





Regression results revisited: Do we improve our results?

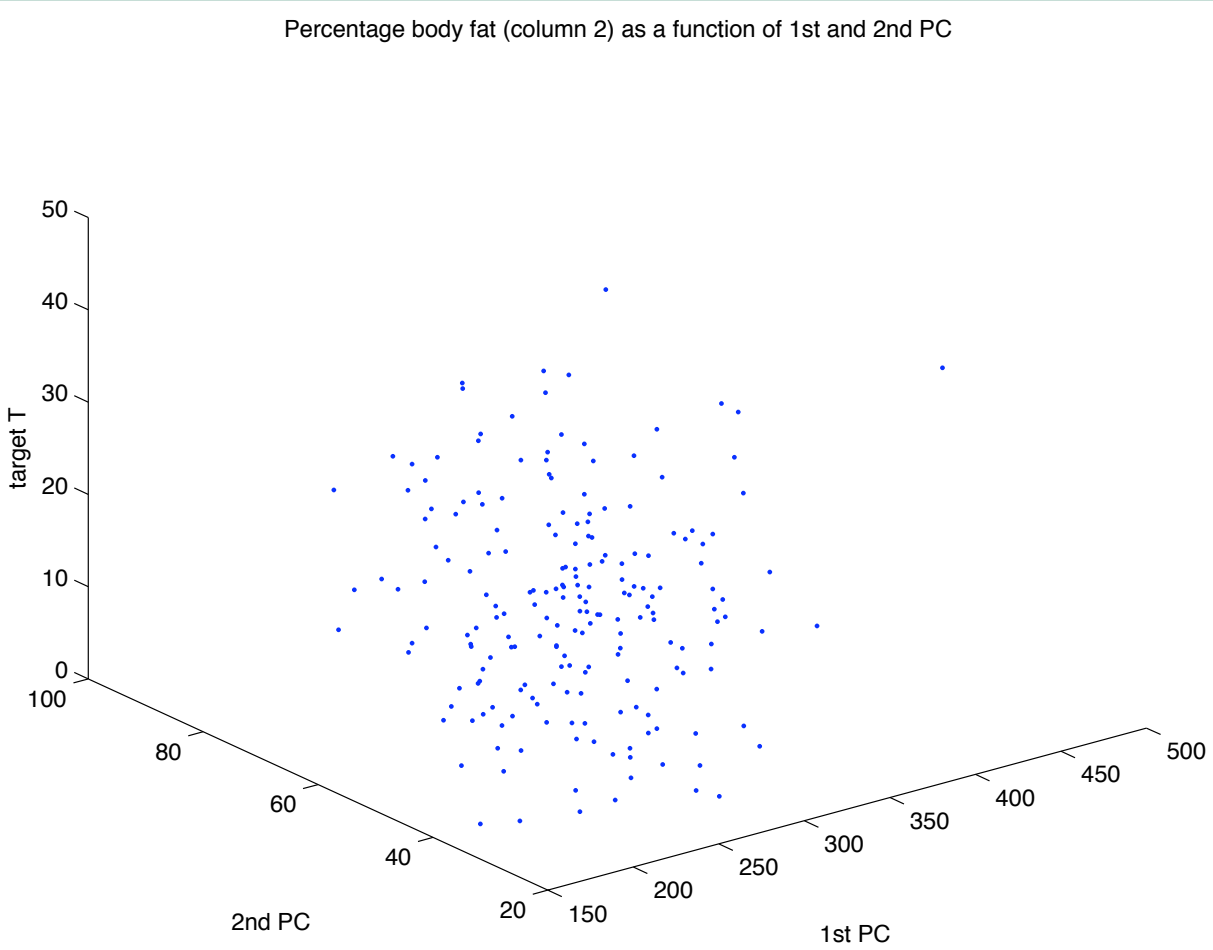
- Maximum likelihood RMS error with column 8:
  RMS = 10.9 (% body fat)

- Maximum likelihood RMS error with PC 1:
  RMS = 9.9 (% body fat)

- (Average results over 50 trials (bootstrapping) to compensate for sensitivity to random partitioning of data into training and test set).

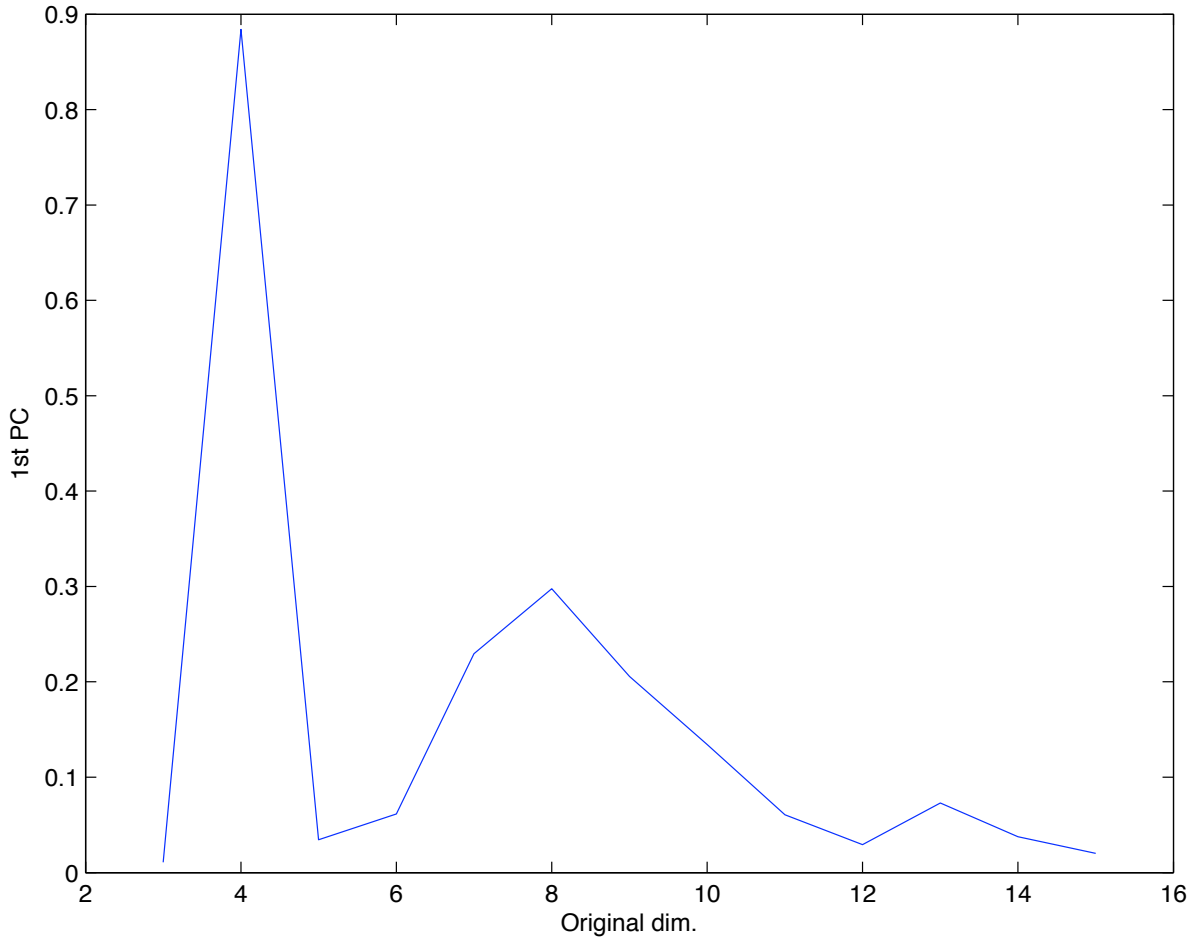
- And maybe we can do better by using both PC 1 and 2?

Target variable plotted against 1st PC and 2nd PC



Since the first 2 PCs capture more than 95 % of the total variance a good choice would be these.

Visualizing the 1st PC vector
(Sometimes useful visualization trick)



PCA picks a linear combination of variables 4, 7, 8, 9, and 10 (the rest are fairly small).

PCA as a data exploration tool - summary

- Projection of data onto PC's can visualize the major modes of variation in the dataset.
- PC's versus target values / class labels can help with our the selection of models.
- By inspecting the individual PC vector we can see which features contributes the most to this PC (feature selection).
- PCA with M < D gives a lossy reconstruction of data and can be used for noise removal.

PCA for small sample size (*N*<*D*): For instance relevant for the hand data set

- *N*<*D*: The problem is singular with only *N* eigenvalues!
- The eigenvector equation for the empirical covariance

$$\mathbf{S}\mathbf{u}_i = \frac{1}{N}\mathbf{X}^T\mathbf{X}\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

- Pre-multiply by **X**: $N^{-1}\mathbf{X}\mathbf{X}^T\left(\mathbf{X}\mathbf{u}_i\right) = \lambda_i\left(\mathbf{X}\mathbf{u}_i\right)$
- Introduce $\mathbf{v}_i = \mathbf{X}\mathbf{u}_i$ and substitute: $N^{-1}\mathbf{X}\mathbf{X}^T\mathbf{v}_i = \lambda_i\mathbf{v}_i$
- Find eigenvectors for this smaller *N* x *N* problem.
- Eigenvector for covariance: Pre-multiply by $\mathbf{X}^T$

$$\frac{1}{N}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{v}_i\right) = \mathbf{S}\left(\mathbf{X}^T\mathbf{v}_i\right) = \lambda_i\left(\mathbf{X}^T\mathbf{v}_i\right)$$

- and back to eigenvectors of **S** : $\mathbf{u}_i = \frac{1}{\left(N\lambda_i\right)^{1/2}}\mathbf{X}^T\mathbf{v}_i$

Singular Value Decomposition (SVD) : A computational trick

- SVD – a simple way of doing PCA: $[\mathbf{U},\mathbf{K},\mathbf{V}] = \mathrm{SVD}(\mathbf{X})$

$$\mathbf{X} = \mathbf{U}\mathbf{K}\mathbf{V}^T \in R^{N\times D}$$

- **U** and **V** are *N* x *N* and *D* x *D* orthogonal matrices:

$$\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_N \qquad \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_D$$

- **K** is a *N* x *D* diagonal matrix of sorted singular values (≥0).
- The covariance may be expressed as

$$\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X} = \frac{1}{N}\mathbf{V}\mathbf{K}^T\mathbf{U}^T\mathbf{U}\mathbf{K}\mathbf{V}^T = \frac{1}{N}\mathbf{V}\mathbf{K}^T\mathbf{K}\mathbf{V}^T$$

- Columns of **V** are eigenvectors of **S** (the PCs):

$$\mathbf{S}\mathbf{v}_i = \frac{\mathbf{K}_{ii}^2}{N}\mathbf{v}_i \qquad \lambda_i = \frac{\mathbf{K}_{ii}^2}{N}$$

- Projection of data onto *i*'th PC: $\mathbf{X}\mathbf{v}_i = \mathbf{U}\mathbf{K}\mathbf{V}^T\mathbf{v}_i = \mathbf{u}_i\mathbf{K}_{ii}$

PCA in preprocessing: Standardization

- It is common to preprocess data by normalizing the individual variables to have zero mean and unit variance:

$$\tilde{x}_{ni} = \frac{\left(x_{ni} - \bar{x}_i\right)}{\sigma_i}$$

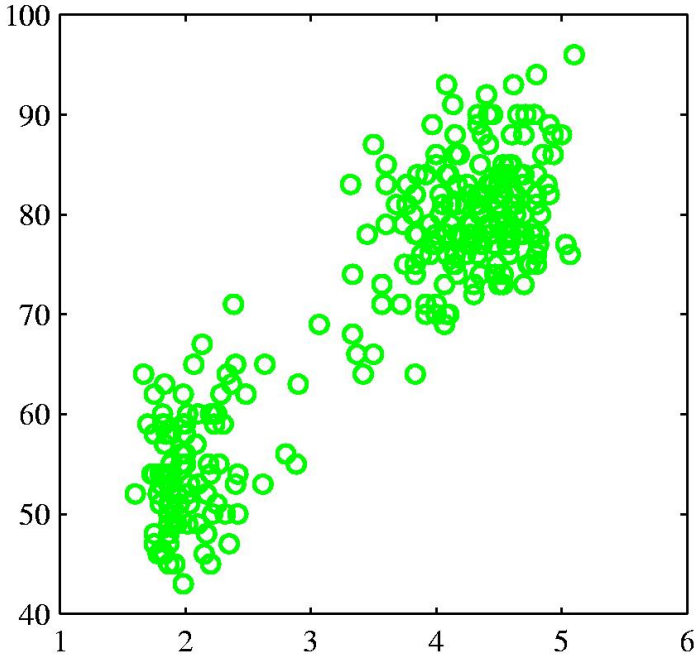
- Covariance matrix becomes the correlation matrix

$$\rho_{ij} = \tilde{\mathbf{S}}_{ij} = \frac{1}{N}\sum_{n=1}^{N} \frac{\left(x_{ni} - \bar{x}_i\right)}{\sigma_i} \frac{\left(x_{nj} - \bar{x}_j\right)}{\sigma_j}$$

- Necessary for some ML algorithms, e.g. distance based methods.

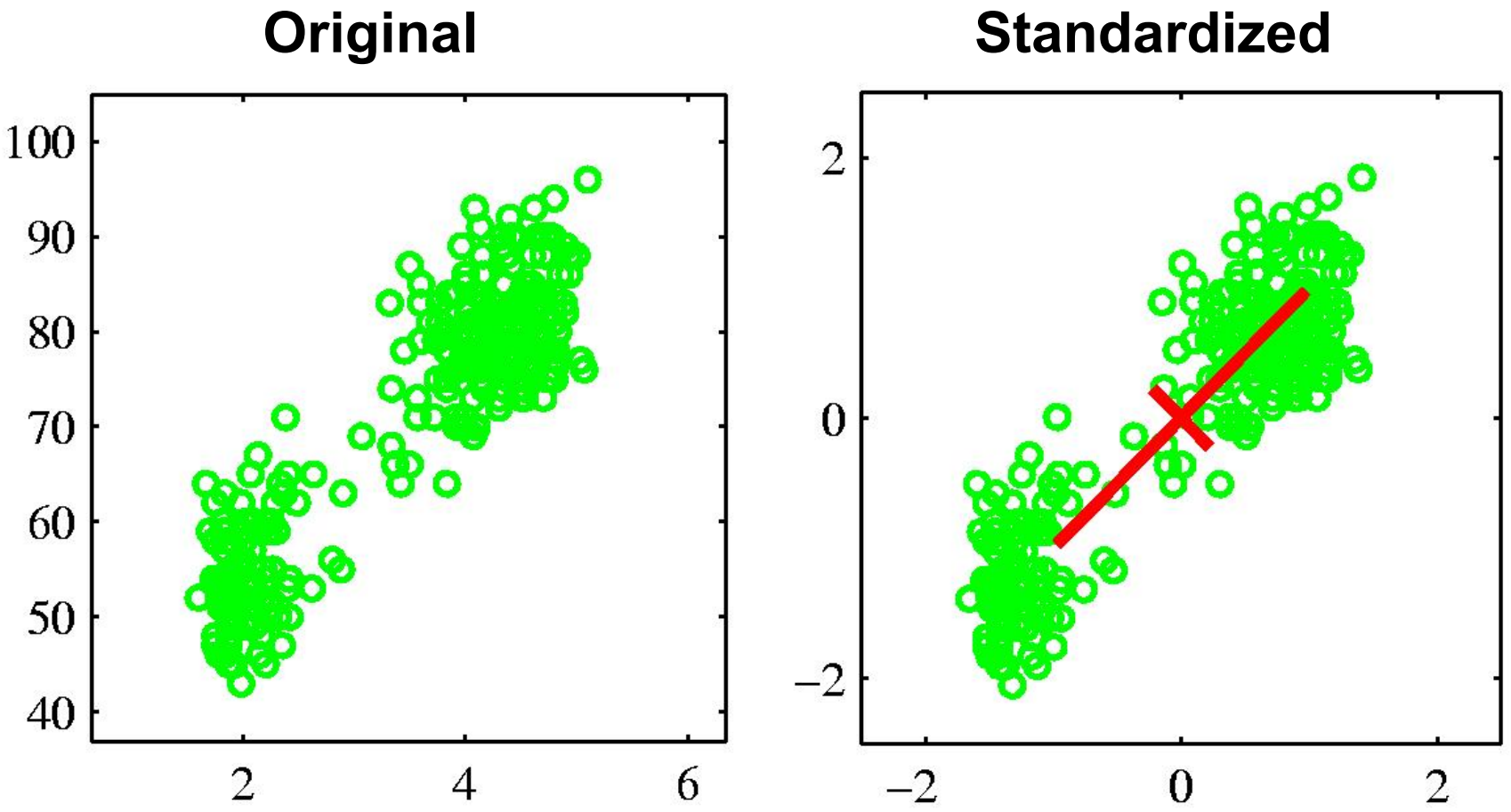
Old Faithful data set

- Hydrothermal geyser in Yellowstone National Park, Wyoming, USA.



- x-axis duration of eruption in minutes
- y-axis time to next eruption in minutes
- Notice the big difference in magnitude of the two axes.

Example of standardization



PCA in preprocessing: Whitening
(Scaling and decorrelating the variables)

- Write the eigenvector equation as $\mathbf{SU} = \mathbf{UL}$, where

$$\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_D) \quad \text{and} \quad \mathbf{L} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_D \end{pmatrix}$$

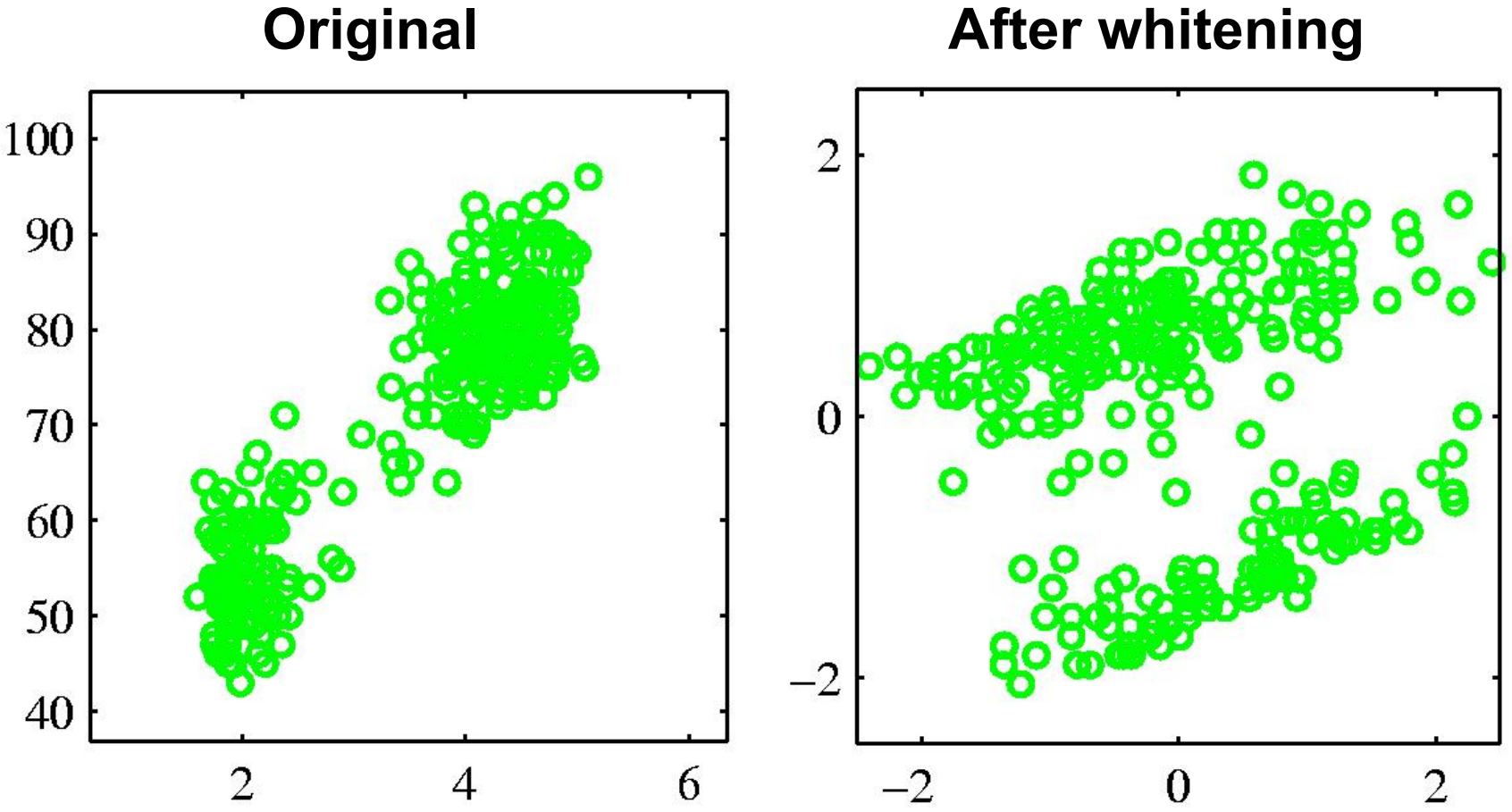
- Translate, rotate, and scale the data into the coordinate system of the PCs: $\mathbf{y}_n = \mathbf{L}^{-1/2}\mathbf{U}^T(\mathbf{x}_n - \bar{\mathbf{x}})$
- In this coordinate system the data is zero mean and have identity covariance

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T = \frac{1}{N}\sum_{n=1}^{N}\mathbf{L}^{-1/2}\mathbf{U}^T(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T\mathbf{U}\mathbf{L}^{-1/2}$$

$$= \mathbf{L}^{-1/2}\mathbf{U}^T\mathbf{S}\mathbf{U}\mathbf{L}^{-1/2} = \mathbf{L}^{-1/2}\mathbf{U}^T\mathbf{U}\mathbf{L}\mathbf{L}^{-1/2} = \mathbf{L}^{-1/2}\mathbf{L}\mathbf{L}^{-1/2} = \mathbf{I}_D$$

- This is referred to as whitening the data.

Example of whitening



Summary

- Continuous latent variable models
- Linear PCA : Find eigenvectors of data covariance
  - Maximum variance formulation
  - Minimum error formulation
- Applications
  - Preprocessing of data (whitening and filtering)
  - Dimensionality reduction
  - Visualization of high dimensional data

Literature

- Continuous latent variable models: CB page 559 – 561
- and Principal Component Analysis (PCA):
  CB page 561 – 570