Principal Components Analysis (PCA)

StatML
11.3.2014
Aasa Feragen
aasa@diku.dk

After today's lecture you should

- Know the definition of PCA
- Understand why PCA is useful for
  - Dimensionality reduction
  - Data preprocessing
  - Visualization of high dimensional data
- Be able to compute principal components for a given dataset
- Be able to use PCA for visualization of global dataset variation
- Be able to use PCA for interpretation of principal component variation for a certain class of data points including shapes
- Be able to show the equivalence between error minimization and variance maximization definitions of PCA

Optional additional reading:

- **Shlens tutorial:**

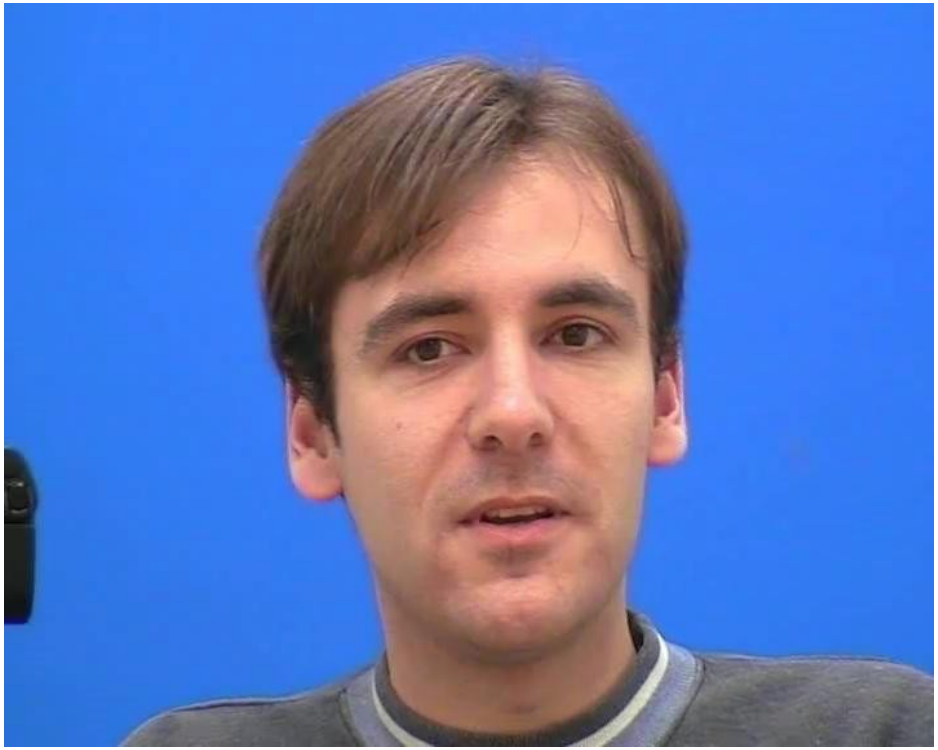
  Fantastic introductory PCA tutorial with matlab code

- **Jain, Duin, Mao, *Statistical Pattern Recognition: A Review*, TPAMI 22 (1), 2000**

  General overview paper; the dimensionality reduction/curse of dimensionality part is great.

Case: Face Shape

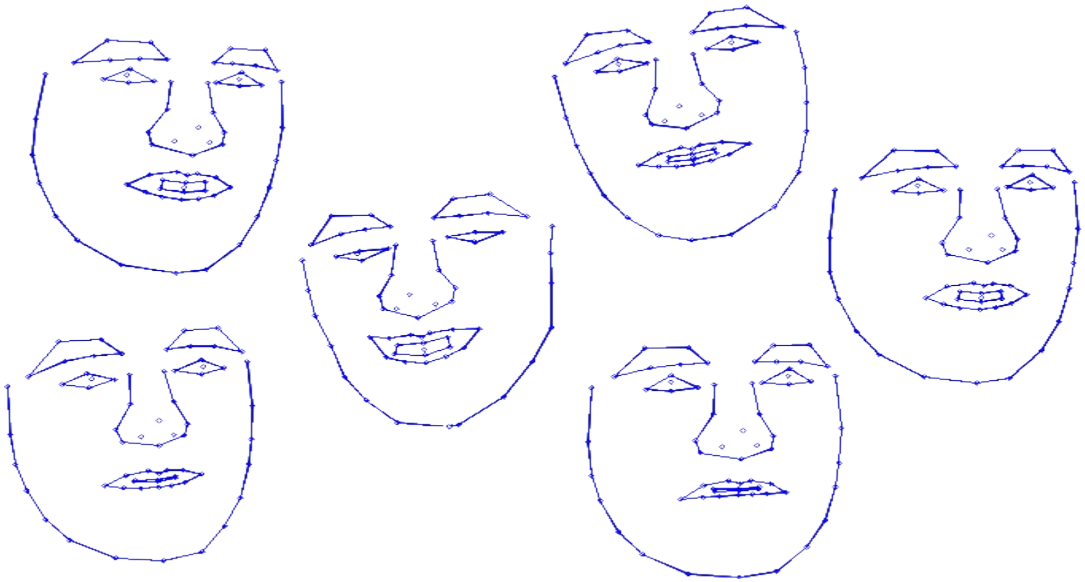
- Here is an image of a man's face
- How do you detect what the face looks like? Can a computer learn to do the same?



All face data from Tim Cootes' talking face dataset, http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html.

Case: Face Shape

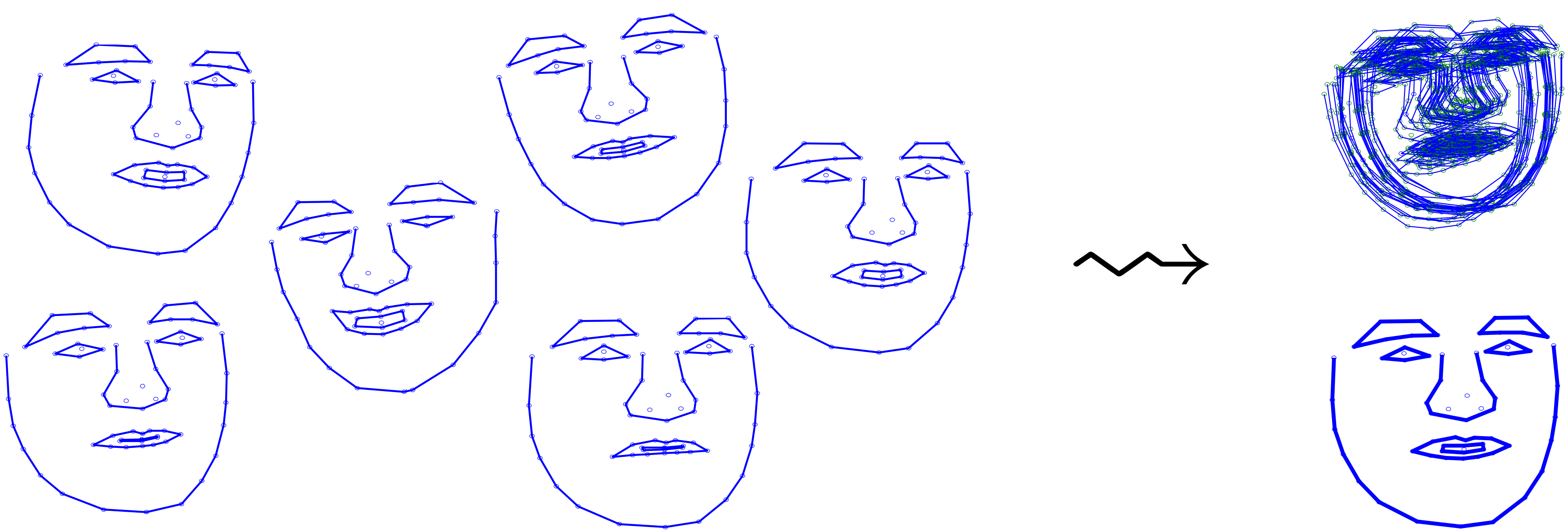
- Here are a set of connect-the-dots-figures describing the man's face while he is talking.
- How can I describe the variation in the face?



All face data from Tim Cootes' talking face dataset, http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html.

Case: Face Shape

- Here are a set of connect-the-dots-figures describing the man's face while he is talking.
- How can I describe the variation in the face?

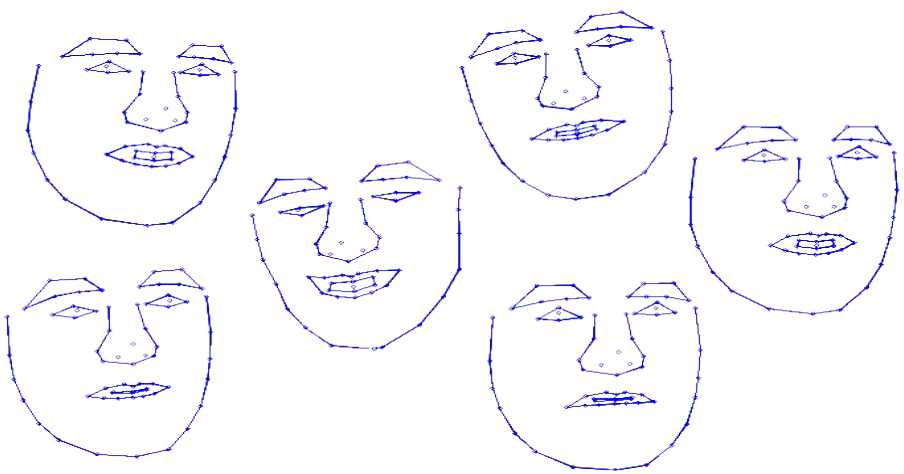


All face data from Tim Cootes' talking face dataset, http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html.

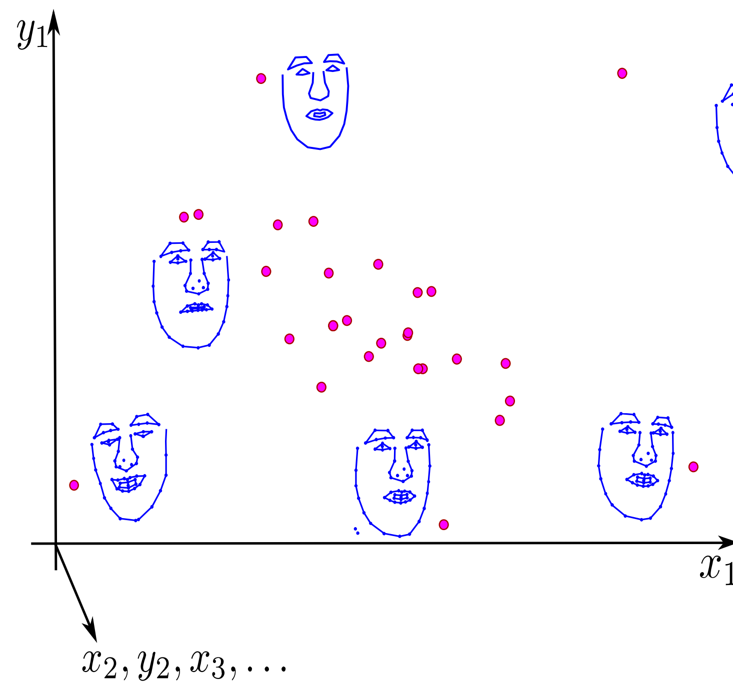
Continuous latent variable models

Often high dimensional data has a low intrinsic dimensionality or few degrees of freedom.

Example: Talking face



“Face space”



Degrees of freedom:

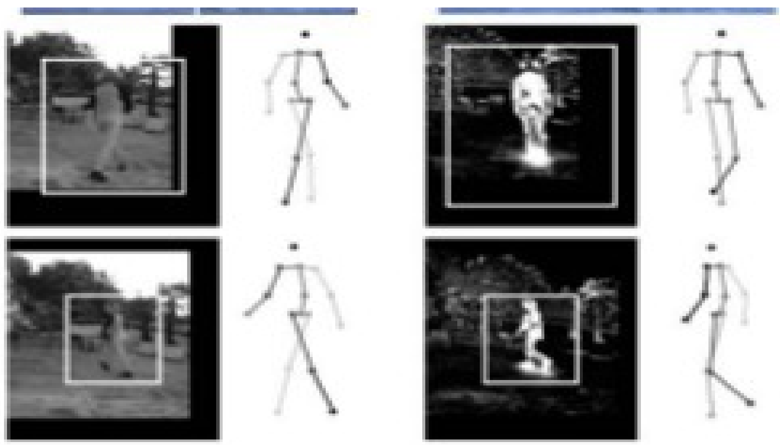
Easy: Translation (2), rotation (1)

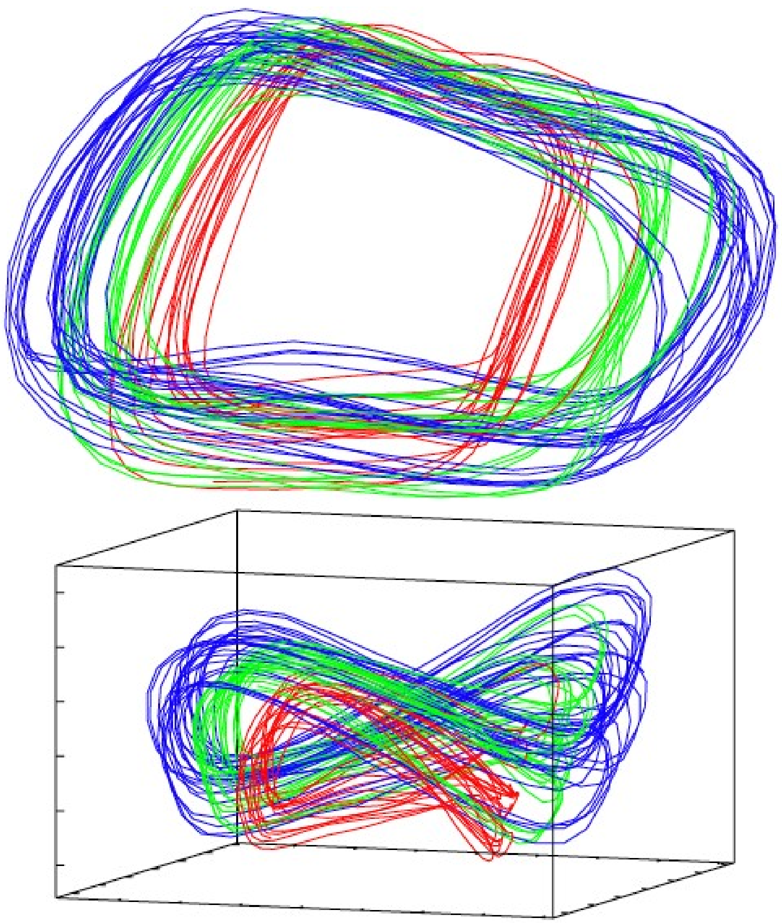
Complicated: coming from the variability in movement when talking.

Continuous latent variable models

Often high dimensional data has a low intrinsic dimensionality or few degrees of freedom.

Example: Walking man





Degrees of freedom:

Easy: Translation (2), rotation (1)

Complicated: coming from the variability in movement when walking.

Not all limb positions correspond to natural walking (or are even physically possible). The “walk” points are sparse in the space of possible limb position configurations

Jaeggli et al 2009

Continuous latent variable models

Model degrees of freedom as latent variable $\mathbf{z}$

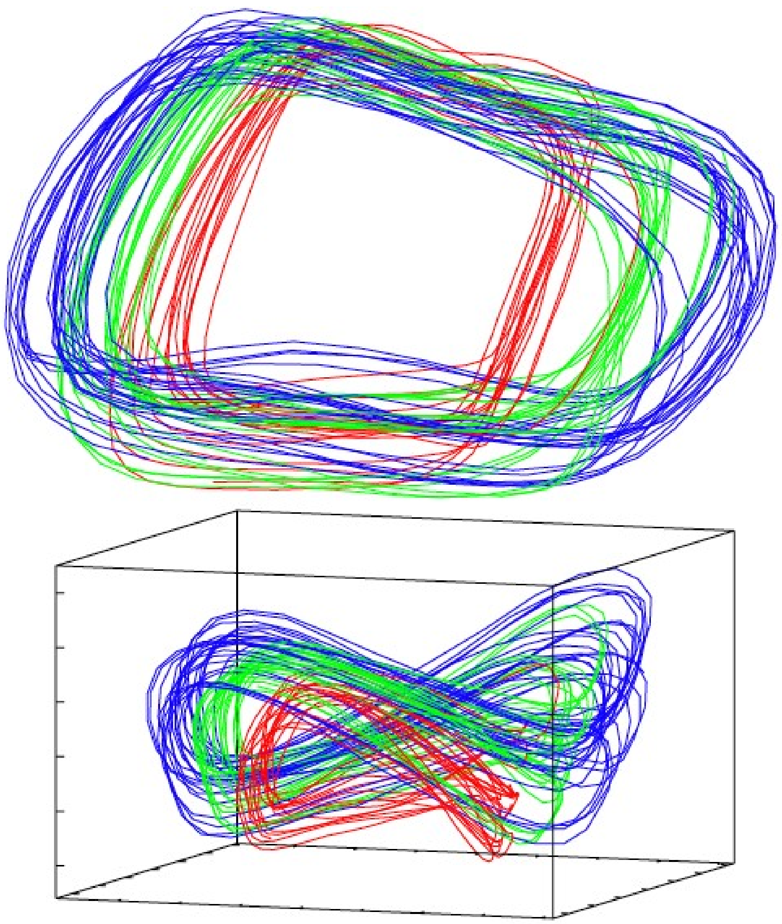
Connection between data representation $\mathbf{x}$ and latent variable $\mathbf{z}$ is generally some non-linear mapping:

$$\mathbf{x} = \phi(\mathbf{z}, \epsilon)$$

including some noise ϵ

Simplest version: Linear model with additive noise: $$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b} + \epsilon$$

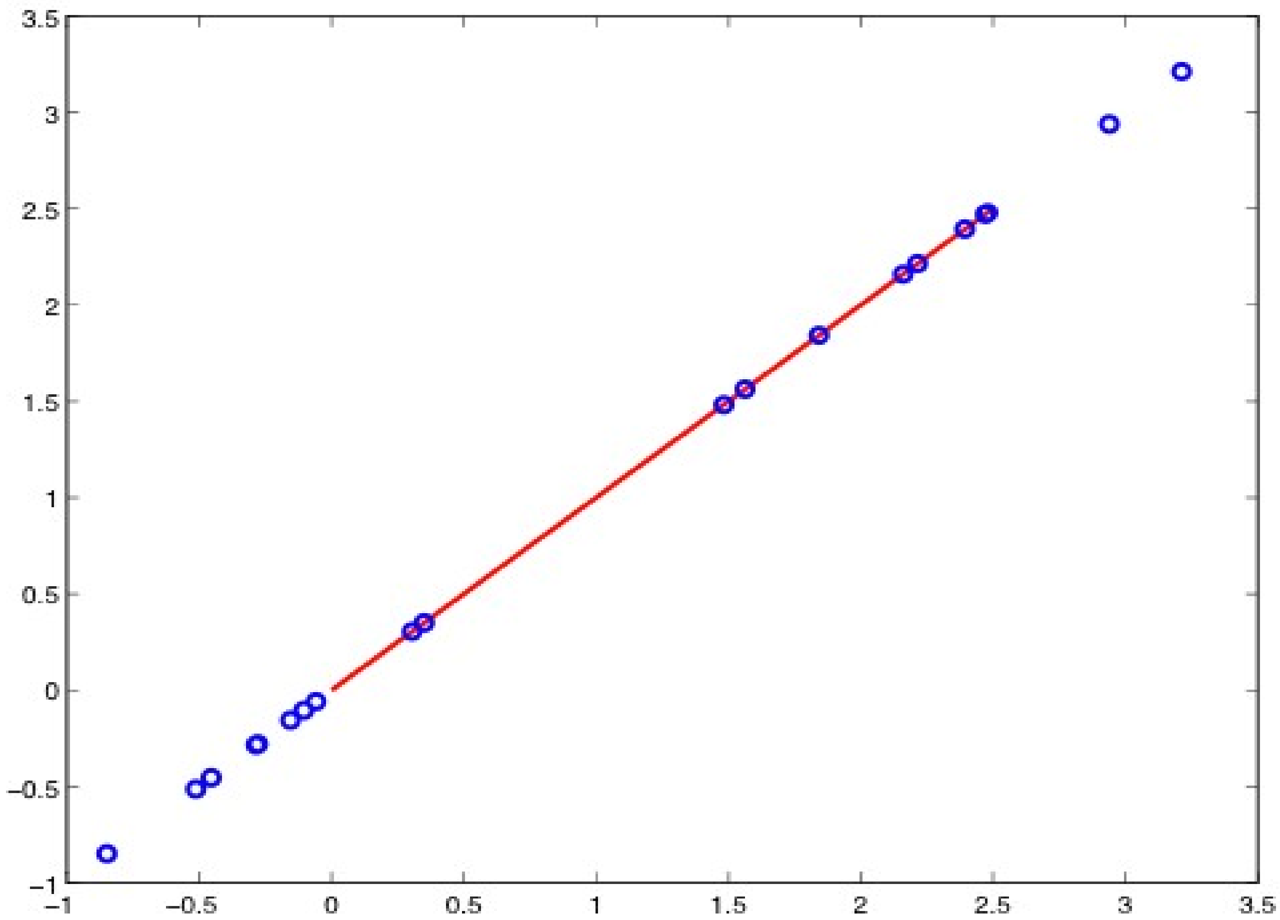
Principal component analysis (PCA) is based on a linear model.



Jaeggli et al 2009

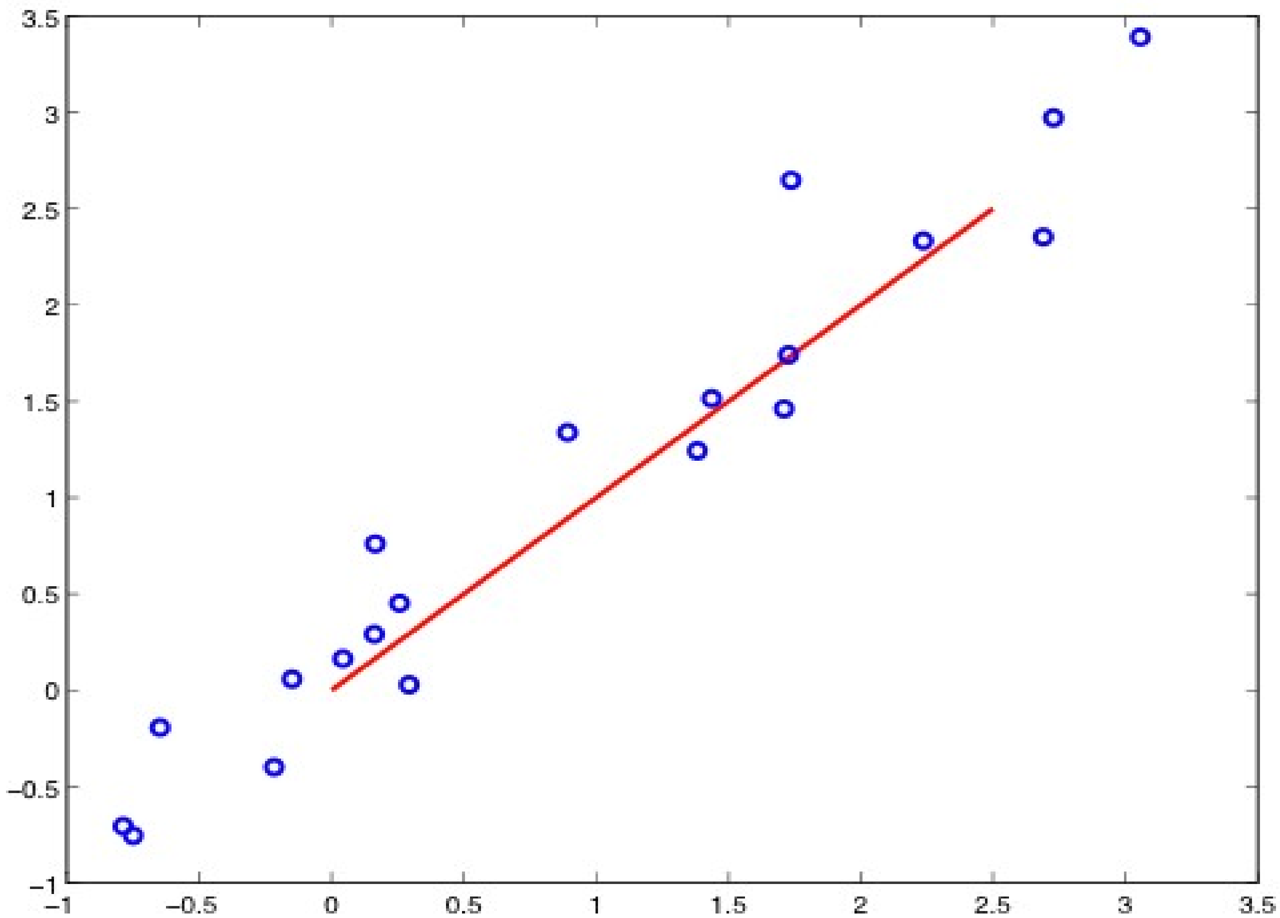
Continuous latent variable models : A synthetic linear 1D latent variable example

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



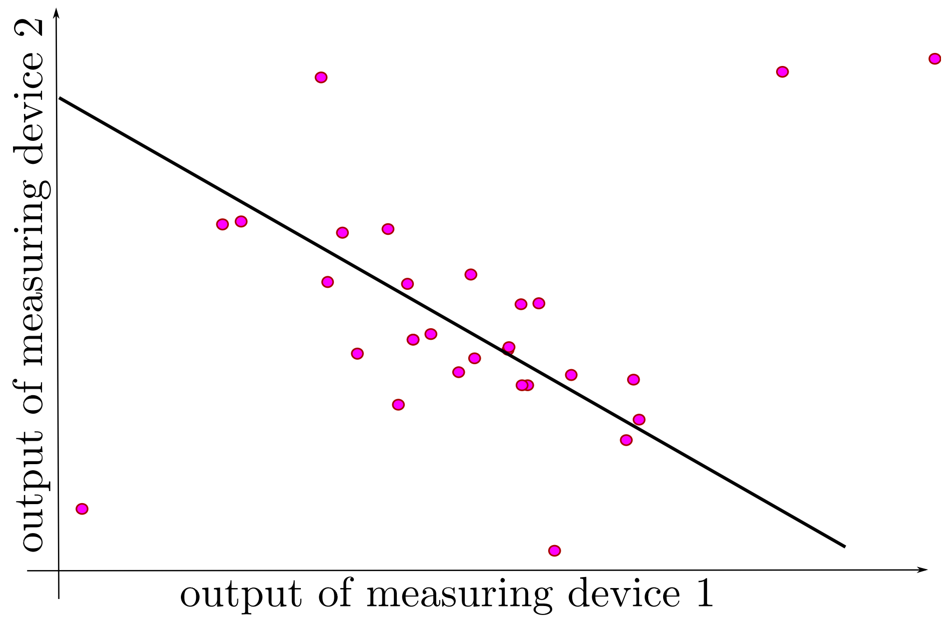
Continuous latent variable models : In reality we often encounter data like this

A bit more messy but data can still be approximated with a linear model: $\mathbf{X} = \mathbf{A}\mathbf{x} + \mathbf{b} + \epsilon$



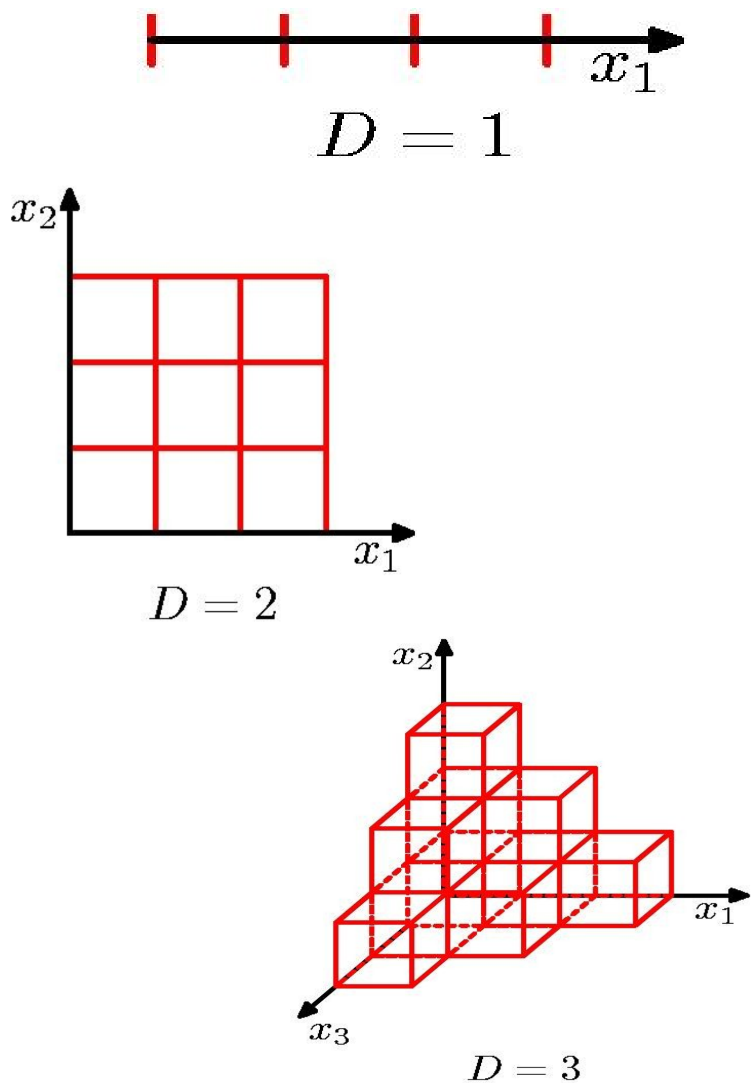
Motivation 1: Correlated features

- Latent variables underlying correlated measurements:
  - Difficulty of assignment
  - Number of points achieved
- Decorrelation might filter out noise and find hidden structure



Motivation 2: Curse of Dimensionality

- In order to sample the interval [0, 1] with density 0.1, I need 10 points
- To sample the cube [0,1] x [0,1] with the same density, I need 100 points
- Etc
- The more dimensions, the more data you need for drawing conclusions



Motivation 2: Curse of Dimensionality

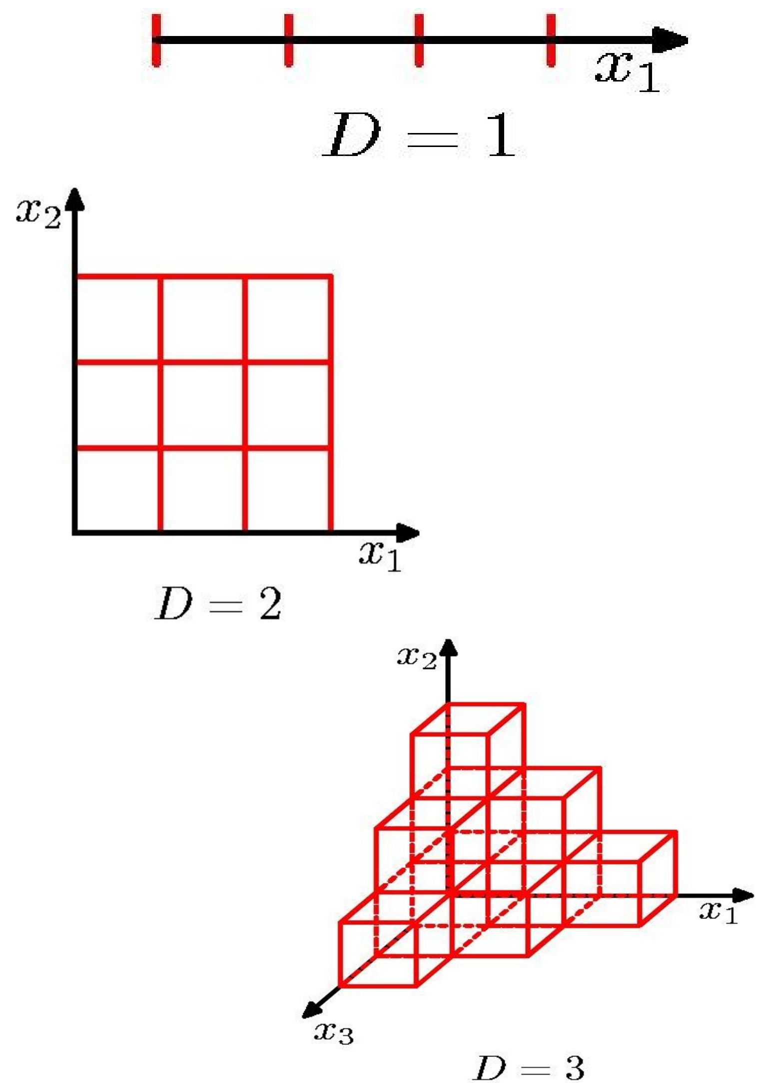
- Consider the d-cube

$$[\text{-}1,\ 1]^d$$

- The distance from the center to a corner is

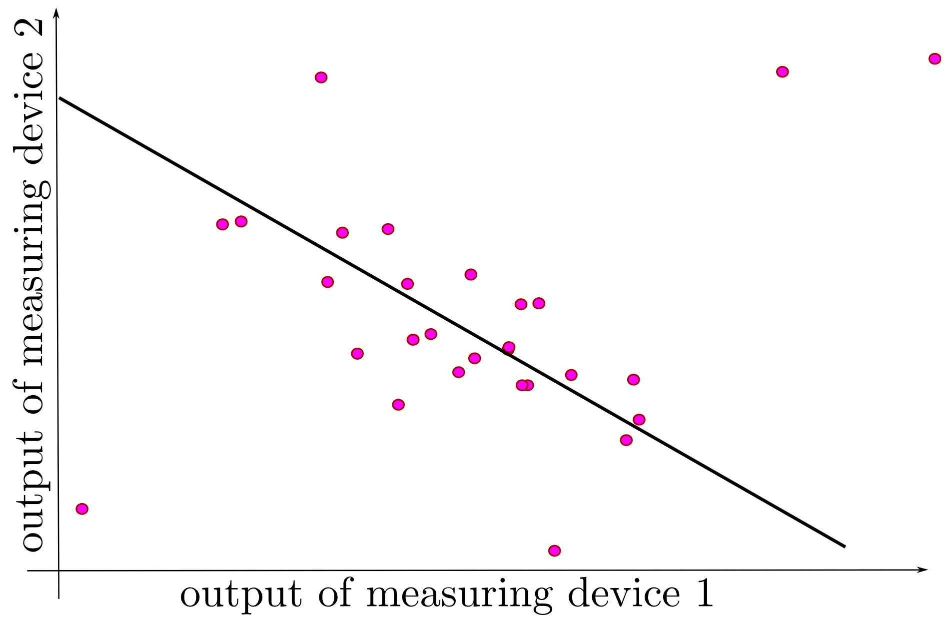
$$\sqrt{d} \to \infty \text{ as } d \to \infty$$

- When d gets large, everything gets large – including noise effects!
- By extracting the essential dimensions, we can avoid using unnecessary dimensions



Motivation 2: Curse of Dimensionality

- Consider the d-cube
$$[-1,\ 1]^d$$
- The distance from the center to a corner is
$$\sqrt{d} \to \infty \text{ as } d \to \infty$$
- When d gets large, everything gets large – including noise effects!
- By extracting the essential dimensions, we can avoid using unnecessary dimensions

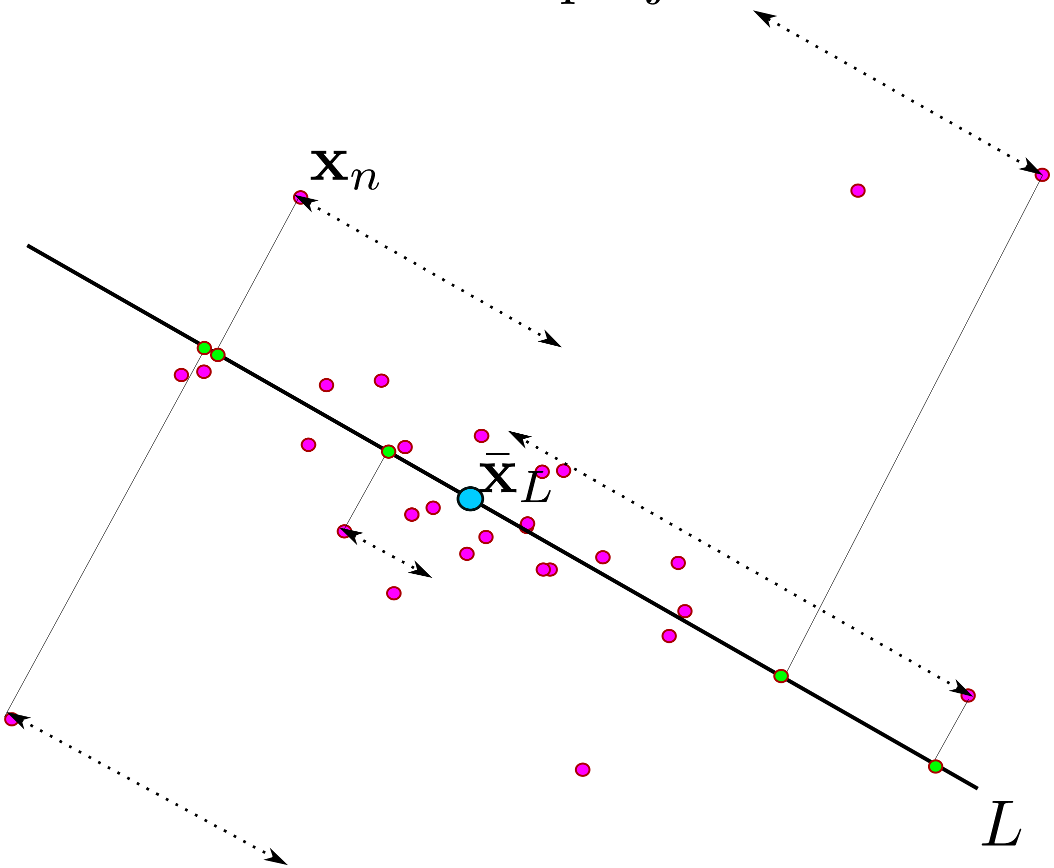


PCA definition 1: Variance maximization

Find M-dimensional hyperplane L which maximizes projected variance

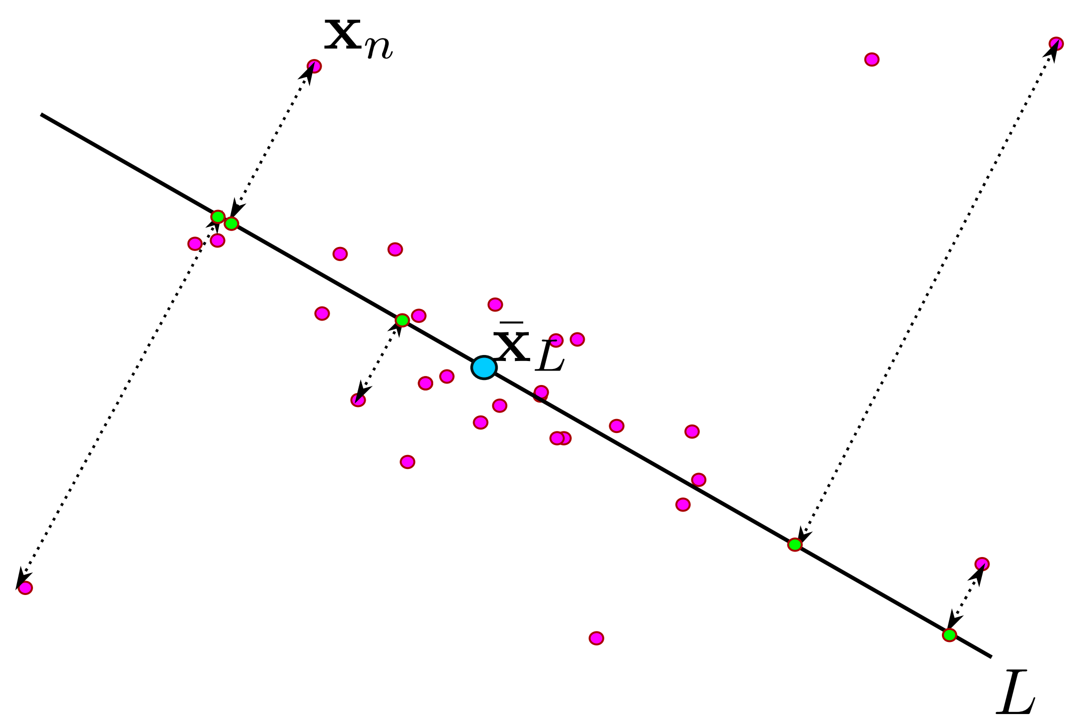
$\sum_{n=1}^{N} \|\mathrm{pr}_L \mathbf{x}_n - \bar{\mathbf{x}}_L\|^2$

$\bar{\mathbf{x}}_L$ mean of $\{\mathrm{pr}_L(\mathbf{x}_n)\}$



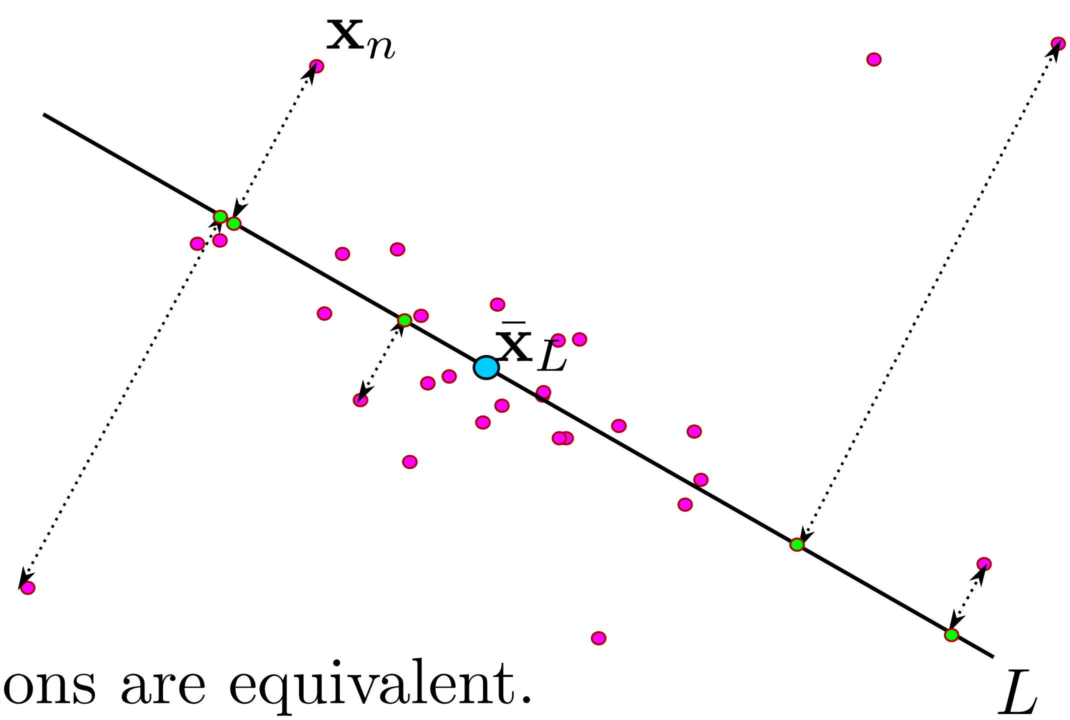
PCA definition 2: Error minimization

Find M-dimensional hyperplane L which minimizes quadratic projection error $\sum_{n=1}^{N} \|\mathbf{x}_n - \mathrm{pr}_L \mathbf{x}_n\|^2$



PCA definition 2: Error minimization

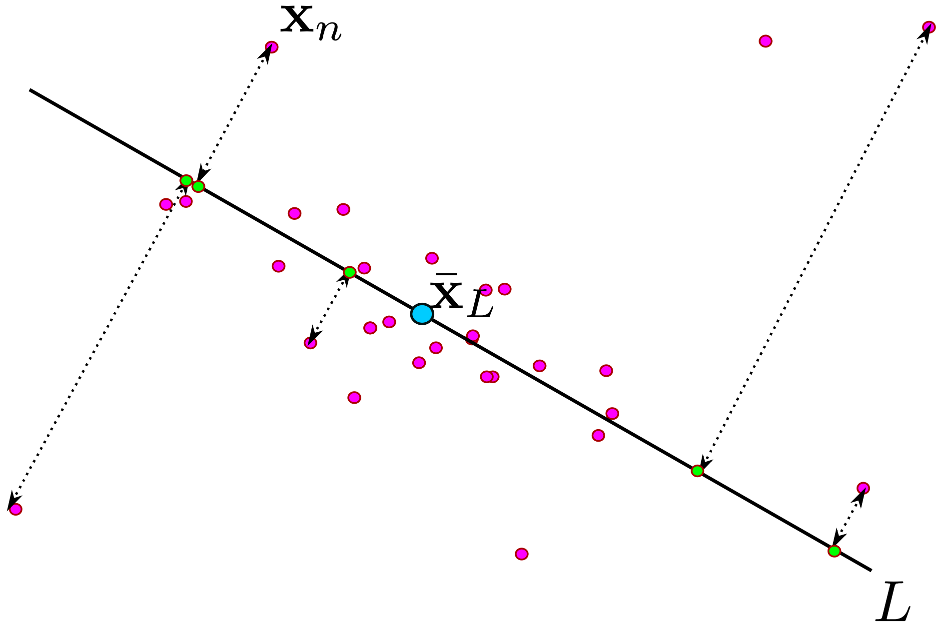
Find M-dimensional hyperplane L which minimizes quadratic projection error $\sum_{n=1}^{N} \|\mathbf{x}_n - \mathrm{pr}_L \mathbf{x}_n\|^2$



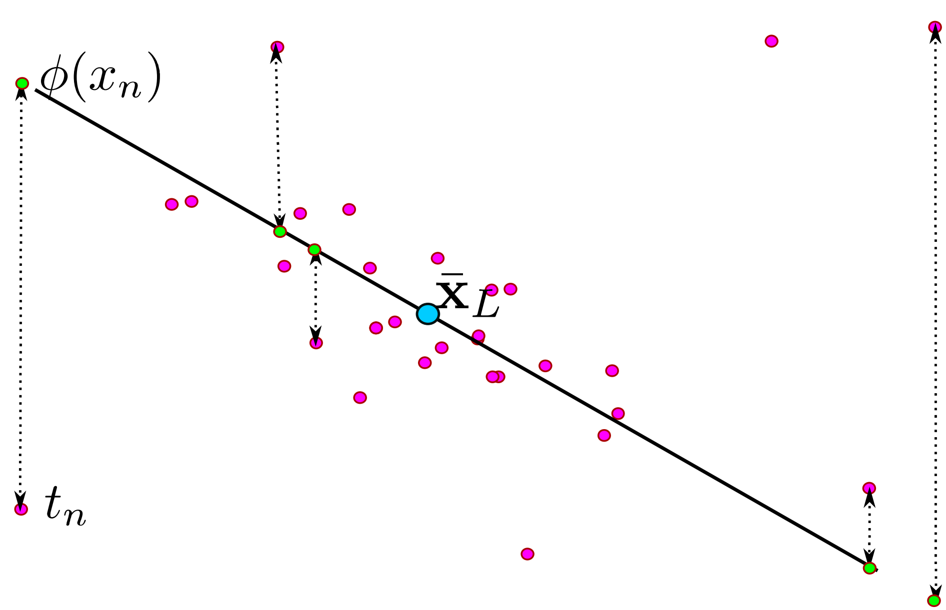
Later today: These definitions are equivalent.

PCA versus geometric formulation of linear regression

$\sum_{n=1}^{N} \|\mathbf{x}_n - \mathrm{pr}_L \mathbf{x}_n\|^2$

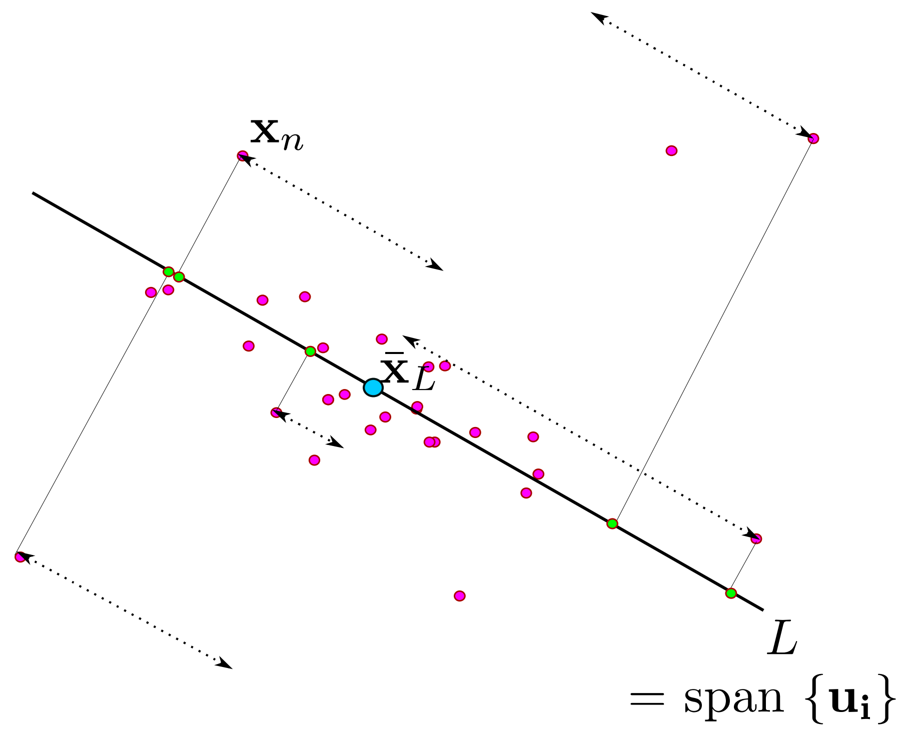


$\sum_{n=1}^{N} \|t_n - \phi(x_n)\|^2$



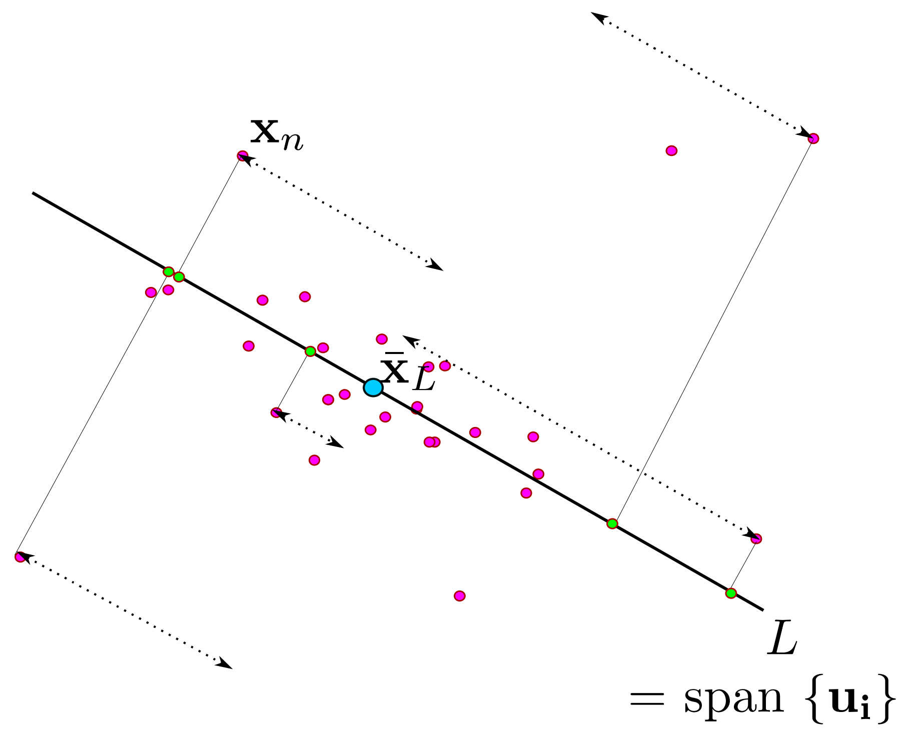
Computing principal components

Task: Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M << D$.
in a way which maximizes projected variance



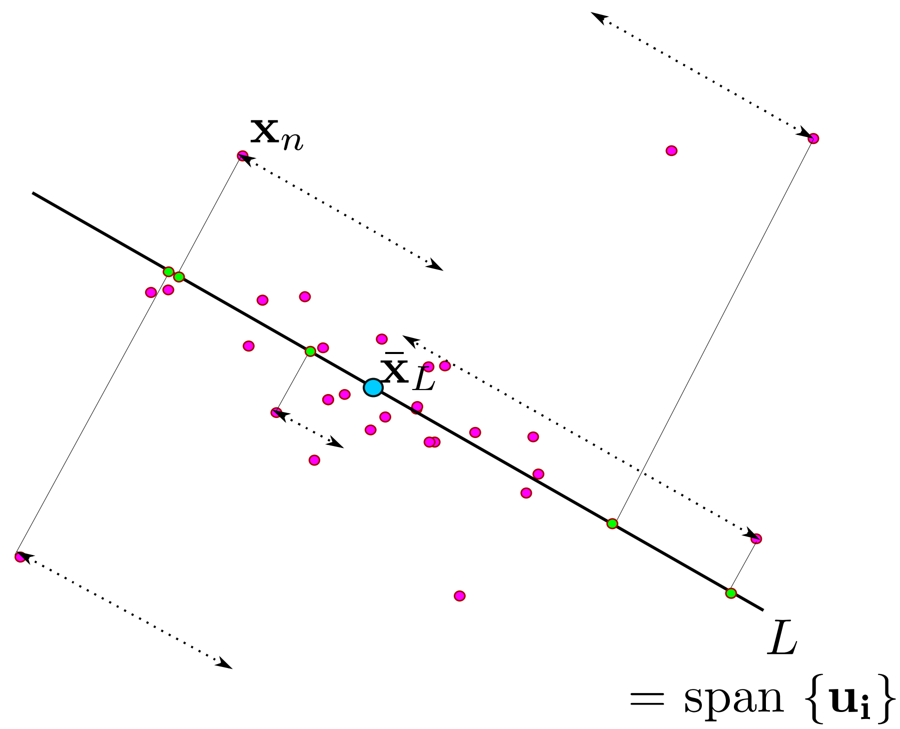
Computing principal components

Task: Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M << D$.
in a way which maximizes projected variance
The $\mathbf{u}_i$ are defined up to translation;



Computing principal components

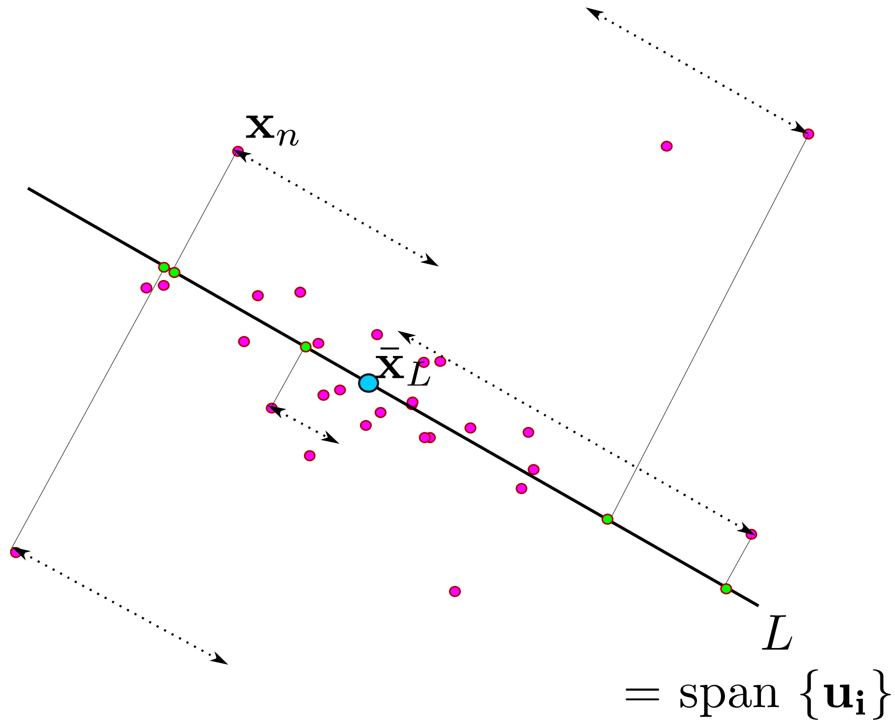
Task: Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M << D$.
in a way which maximizes projected variance
The $\mathbf{u}_i$ are defined up to translation;
use unit vectors $\mathbf{u_i}$ s.t. $\mathbf{u_i}^T\mathbf{u_j} = \delta_{\mathbf{ij}}$



Computing principal components

Task: Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M << D$.
in a way which maximizes projected variance
The $\mathbf{u}_i$ are defined up to translation;
use unit vectors $\mathbf{u_i}$ s.t. $\mathbf{u_i}^T \mathbf{u_j} = \delta_{\mathbf{ij}}$

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



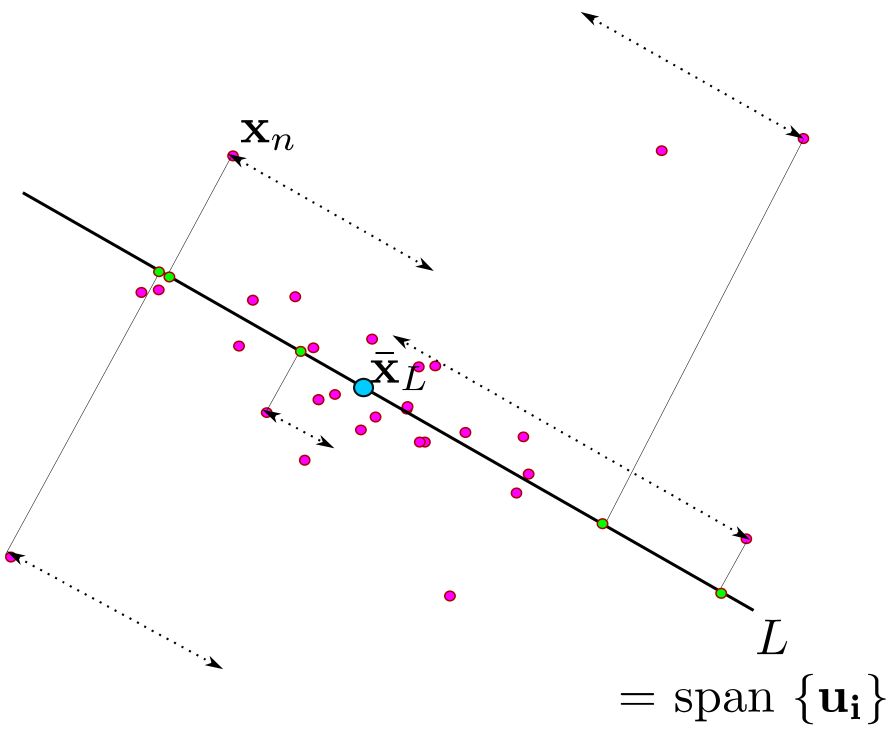
Computing principal components

Task: Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M << D$.
in a way which maximizes projected variance
The $\mathbf{u}_i$ are defined up to translation;
use unit vectors $\mathbf{u_i}$ s.t. $\mathbf{u_i}^T\,\mathbf{u_j} = \delta_{\mathbf{ij}}$

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

Mean of projected data: $\mathbf{u_1^T}\bar{\mathbf{x}}$ with

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



Computing principal components

Task: Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M << D$.
in a way which maximizes projected variance
The $\mathbf{u}_i$ are defined up to translation;
use unit vectors $\mathbf{u_i}$ s.t. $\mathbf{u_i}^T\,\mathbf{u_j} = \delta_{\mathbf{ij}}$

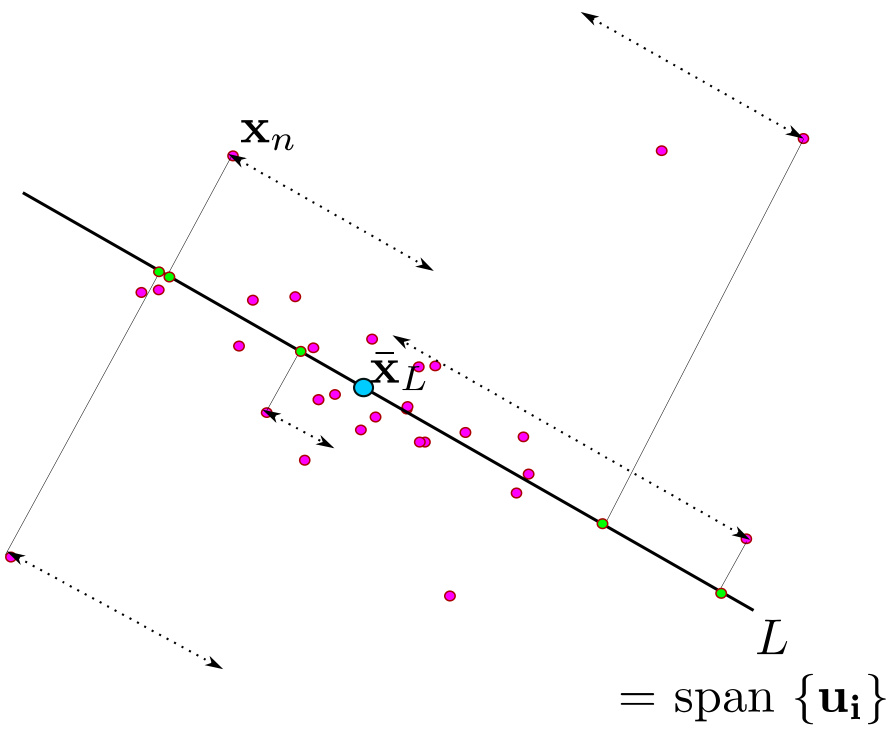
We find directions sequentially, $\mathbf{u}_1$ first.

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:

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



Computing principal components

Task: Project data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ onto directions $\{\mathbf{u}_i\}_{i=1,\ldots,M}$ with $M << D$.
in a way which maximizes projected variance
The $\mathbf{u}_i$ are defined up to translation;
use unit vectors $\mathbf{u_i}$ s.t. $\mathbf{u_i}^T \mathbf{u_j} = \delta_{\mathbf{ij}}$

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

Mean of projected data: $\mathbf{u_1^T}\bar{\mathbf{x}}$ with

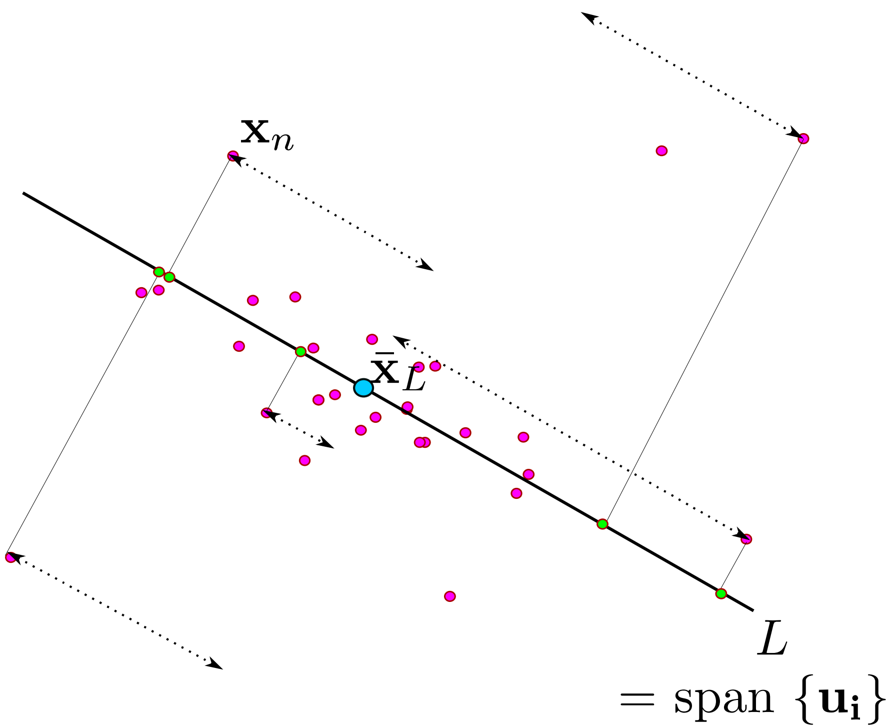
$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

Variance of projected data:

$$\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^T(\mathbf{x}_n - \bar{\mathbf{x}})\}^2 = \mathbf{u_1^T S u_1},$$

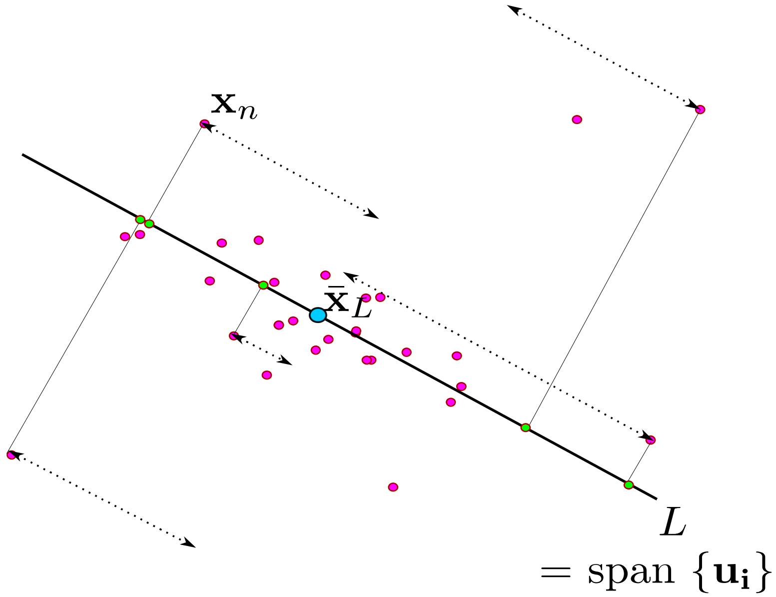
with the empirical co-variance

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



Computing principal components

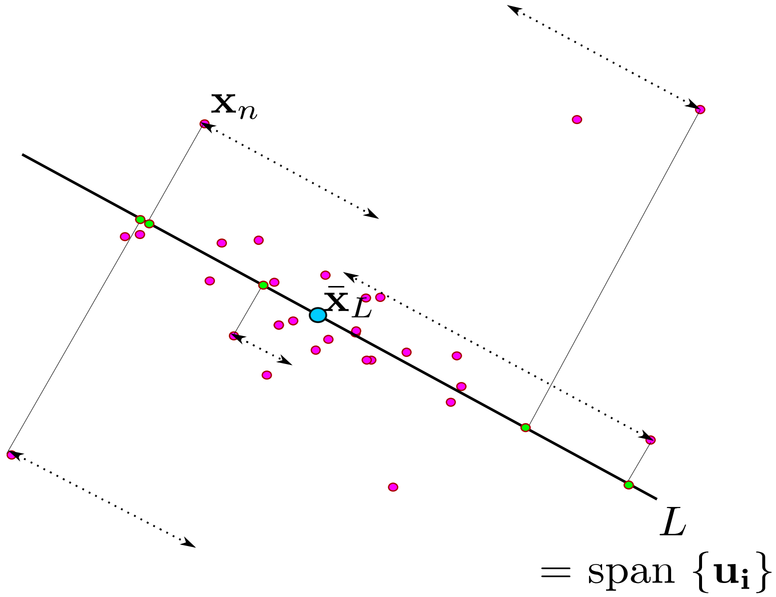
Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$



Computing principal components

Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$

The constraint $\mathbf{u_1^T u_1} = \|\mathbf{u_1}\|^2 = \mathbf{1}$ lets us avoid $\mathbf{u_1} \to \infty$.



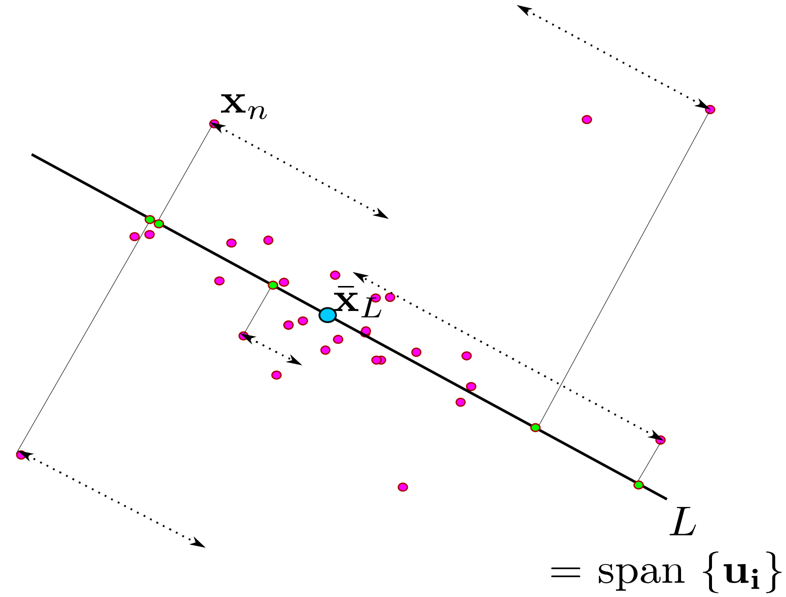
Computing principal components

Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$

The constraint $\mathbf{u_1^T u_1} = \|\mathbf{u_1}\|^{\mathbf{2}} = \mathbf{1}$ lets us avoid $\mathbf{u_1} \to \infty$.

Trick: Lagrangian multipliers $\Rightarrow$

$$\text{argmax}_{\{\mathbf{u_1} | \mathbf{u_1^T u_1 = 1}\}} \mathbf{u_1^T S u_1} = \text{argmax}_{\mathbf{u_1}} \mathbf{u_1^T S u_1} + \lambda_1 (1 + \mathbf{u_1^T u_1})$$



Computing principal components

Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$

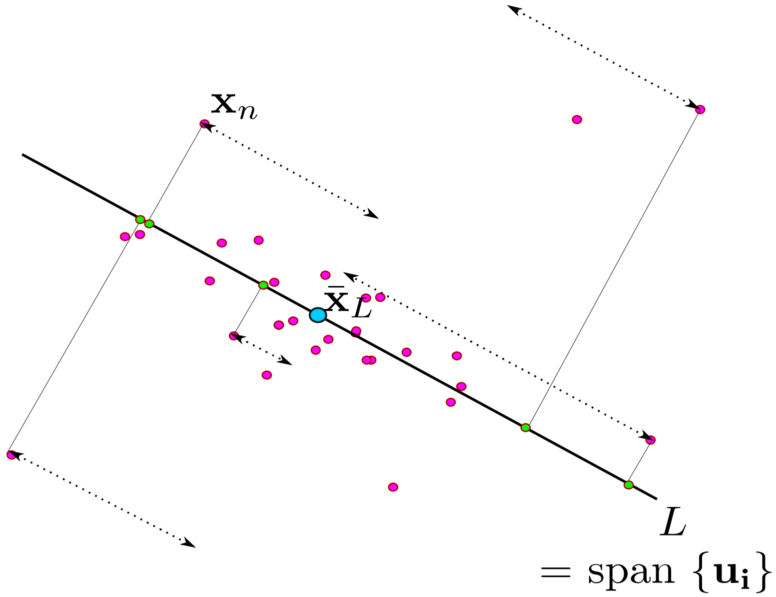
The constraint $\mathbf{u_1^T u_1} = \|\mathbf{u_1}\|^\mathbf{2} = \mathbf{1}$ lets us avoid $\mathbf{u_1} \to \infty$.

Trick: Lagrangian multipliers $\Rightarrow$

$$\text{argmax}_{\{\mathbf{u_1} | \mathbf{u_1^T u_1 = 1}\}} \mathbf{u_1^T S u_1} = \text{argmax}_{\mathbf{u_1}} \mathbf{u_1^T S u_1} + \lambda_1 (1 + \mathbf{u_1^T u_1})$$

Exercise: Take derivative w.r.t. $\mathbf{u_1}$ and show that for the optimal $\mathbf{u_1}$,

$$\mathbf{S u_1} = \lambda_1 \mathbf{u_1}$$



Computing principal components

Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$

The constraint $\mathbf{u_1^T u_1} = \|\mathbf{u_1}\|^\mathbf{2} = \mathbf{1}$ lets us avoid $\mathbf{u_1} \to \infty$.

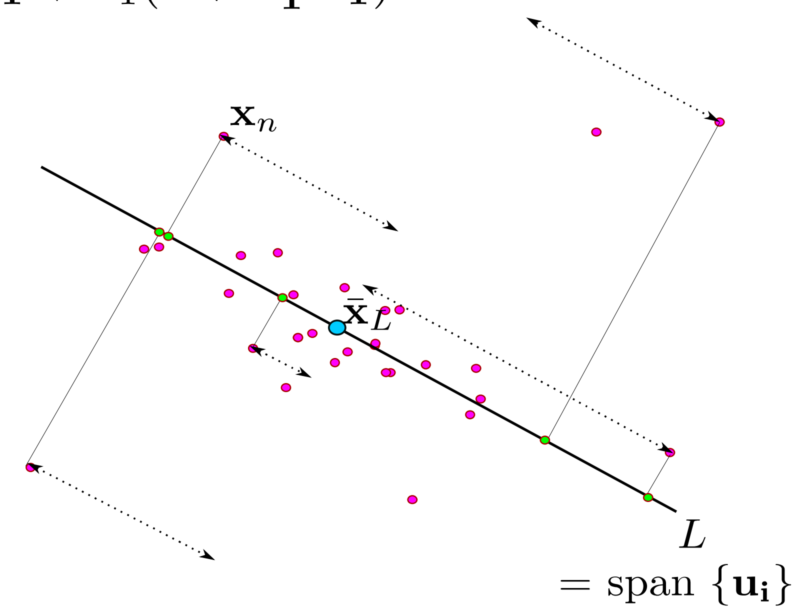
Trick: Lagrangian multipliers $\Rightarrow$

$$\text{argmax}_{\{\mathbf{u_1}|\mathbf{u_1^T u_1 = 1}\}} \mathbf{u_1^T S u_1} = \text{argmax}_{\mathbf{u_1}} \mathbf{u_1^T S u_1} + \lambda_1(1 + \mathbf{u_1^T u_1})$$

Exercise: Take derivative w.r.t. $\mathbf{u_1}$ and show that for the optimal $\mathbf{u_1}$,

$$\mathbf{S u_1} = \lambda_1 \mathbf{u_1}$$

How to interpret this?



Computing principal components

Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$

The constraint $\mathbf{u_1^T u_1} = \|\mathbf{u_1}\|^2 = \mathbf{1}$ lets us avoid $\mathbf{u_1} \to \infty$.

Trick: Lagrangian multipliers $\Rightarrow$

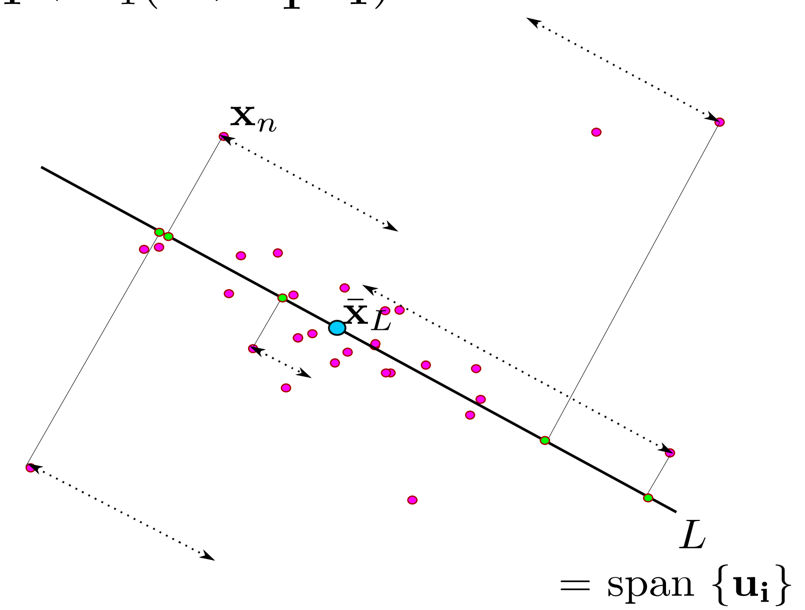
$$\text{argmax}_{\{\mathbf{u_1} | \mathbf{u_1^T u_1} = \mathbf{1}\}} \mathbf{u_1^T S u_1} = \text{argmax}_{\mathbf{u_1}} \mathbf{u_1^T S u_1} + \lambda_1 (1 + \mathbf{u_1^T u_1})$$

Exercise: Take derivative w.r.t. $\mathbf{u_1}$ and show that for the optimal $\mathbf{u_1}$,

$$\mathbf{S u_1} = \lambda_1 \mathbf{u_1}$$

How to interpret this?

$\mathbf{u_1}$ eigenvector of covariance $\mathbf{S}$ with eigenvalue λ_1



Computing principal components

Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$

The constraint $\mathbf{u_1^T u_1} = \|\mathbf{u_1}\|^2 = \mathbf{1}$ lets us avoid $\mathbf{u_1} \to \infty$.

Trick: Lagrangian multipliers $\Rightarrow$

$$\operatorname{argmax}_{\{\mathbf{u_1} | \mathbf{u_1^T u_1 = 1}\}} \mathbf{u_1^T S u_1} = \operatorname{argmax}_{\mathbf{u_1}} \mathbf{u_1^T S u_1} + \lambda_1 (1 + \mathbf{u_1^T u_1})$$

Exercise: Take derivative w.r.t. $\mathbf{u_1}$ and show that for the optimal $\mathbf{u_1}$,

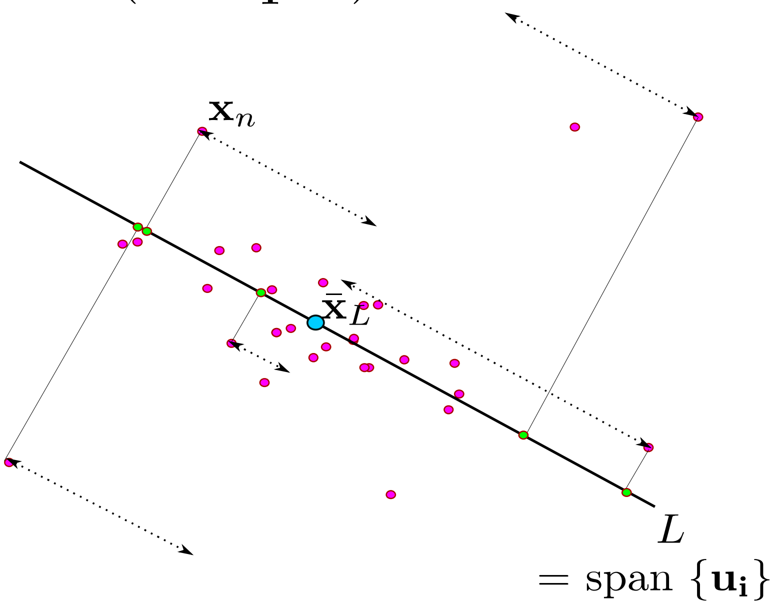
$$\mathbf{S u_1} = \lambda_1 \mathbf{u_1}$$

How to interpret this?

$\mathbf{u_1}$ eigenvector of covariance $\mathbf{S}$ with eigenvalue λ_1

Multiply with $\mathbf{u_1^T}$ and see

$$\mathbf{u_1^T S u_1} = \lambda_1 \mathbf{u_1^T u_1} = \lambda_1$$



Computing principal components

Task: Maximize variance $\mathbf{u_1^T S u_1}$ with respect to $\mathbf{u_1}$

The constraint $\mathbf{u_1^T u_1} = \|\mathbf{u_1}\|^{\mathbf{2}} = \mathbf{1}$ lets us avoid $\mathbf{u_1} \to \infty$.

Trick: Lagrangian multipliers $\Rightarrow$

$$\operatorname{argmax}_{\{\mathbf{u_1} | \mathbf{u_1^T u_1 = 1}\}} \mathbf{u_1^T S u_1} = \operatorname{argmax}_{\mathbf{u_1}} \mathbf{u_1^T S u_1} + \lambda_1 (1 + \mathbf{u_1^T u_1})$$

Exercise: Take derivative w.r.t. $\mathbf{u_1}$ and show that for the optimal $\mathbf{u_1}$,

$$\mathbf{S u_1} = \lambda_1 \mathbf{u_1}$$

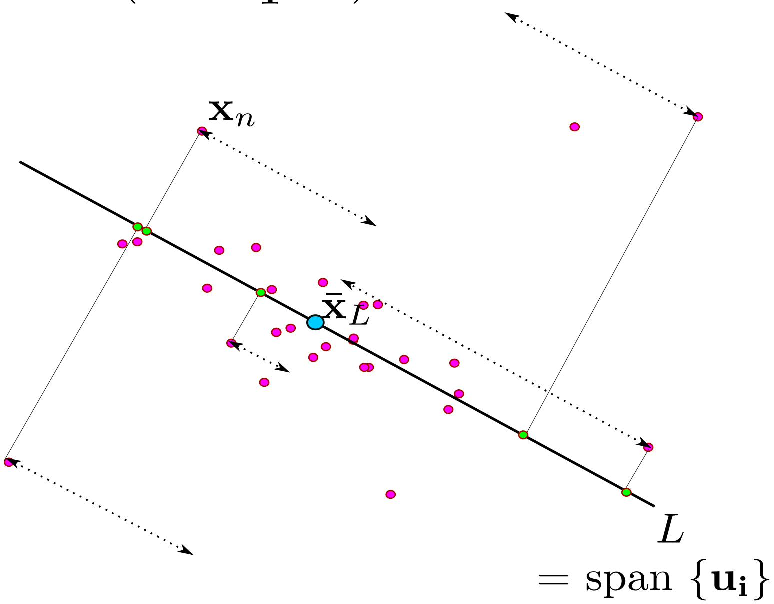
How to interpret this?

$\mathbf{u_1}$ eigenvector of covariance $\mathbf{S}$ with eigenvalue λ_1

Multiply with $\mathbf{u_1^T}$ and see

$$\mathbf{u_1^T S u_1} = \lambda_1 \mathbf{u_1^T u_1} = \lambda_1$$

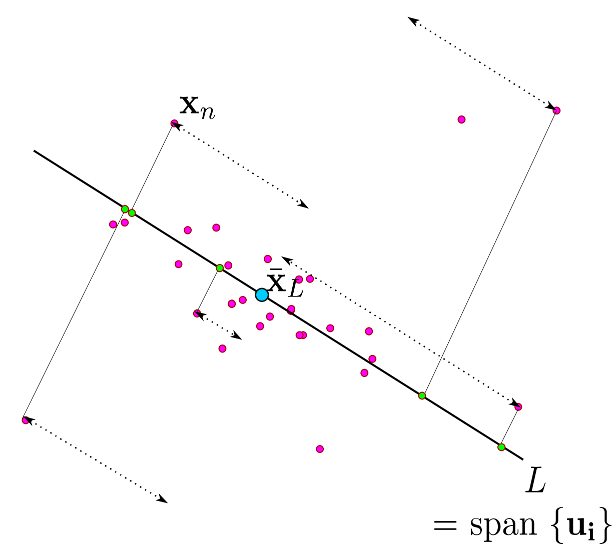
$\mathbf{u_1}$ is eigenvector of maximal eigenvalue λ_1!



Computing principal components

To compute the M-dimensional hyperplane minimizing projected variance:

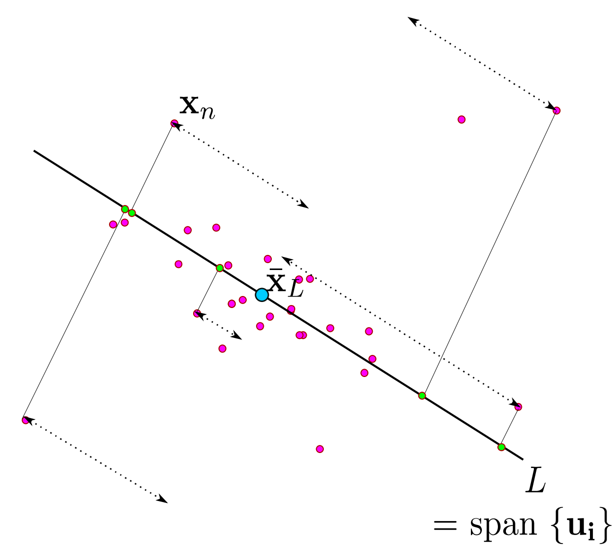
1. Compute covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \bar{\mathbf{x}})(\mathbf{x_n} - \bar{\mathbf{x}})^{\mathbf{T}}$



Computing principal components

To compute the M-dimensional hyperplane minimizing projected variance:

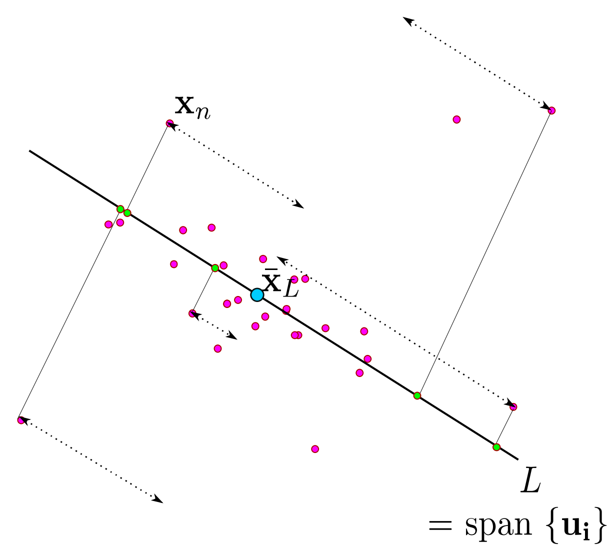
1. Compute covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \bar{\mathbf{x}})(\mathbf{x_n} - \bar{\mathbf{x}})^{\mathbf{T}}$
2. Compute its eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq \ldots \geq \lambda_D$
   and their eigenvectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_M}, \ldots \mathbf{u_D}$



Computing principal components

To compute the M-dimensional hyperplane minimizing projected variance:

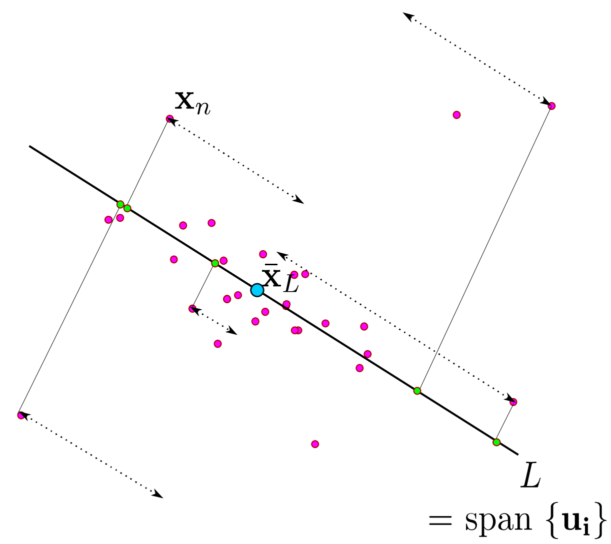
1. Compute covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \bar{\mathbf{x}})(\mathbf{x_n} - \bar{\mathbf{x}})^{\mathbf{T}}$
2. Compute its eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq \ldots \geq \lambda_D$
   and their eigenvectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_M}, \ldots \mathbf{u_D}$
3. Optimal hyperplane is span $\{\mathbf{u_i}\}_{\mathbf{i=1}}^{\mathbf{M}}$



Computing principal components

To compute the M-dimensional hyperplane minimizing projected variance:

1. Compute covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \bar{\mathbf{x}})(\mathbf{x_n} - \bar{\mathbf{x}})^{\mathbf{T}}$
2. Compute its eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq \ldots \geq \lambda_D$
   and their eigenvectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_M}, \ldots \mathbf{u_D}$
3. Optimal hyperplane is span $\{\mathbf{u_i}\}_{\mathbf{i=1}}^{\mathbf{M}}$
4. Projected variance is $\lambda_1 + \lambda_2 + \ldots + \lambda_M$

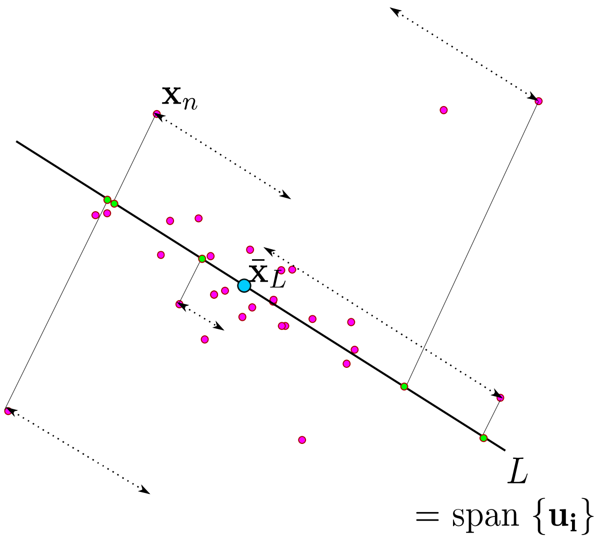


Computing principal components

To compute the M-dimensional hyperplane minimizing projected variance:

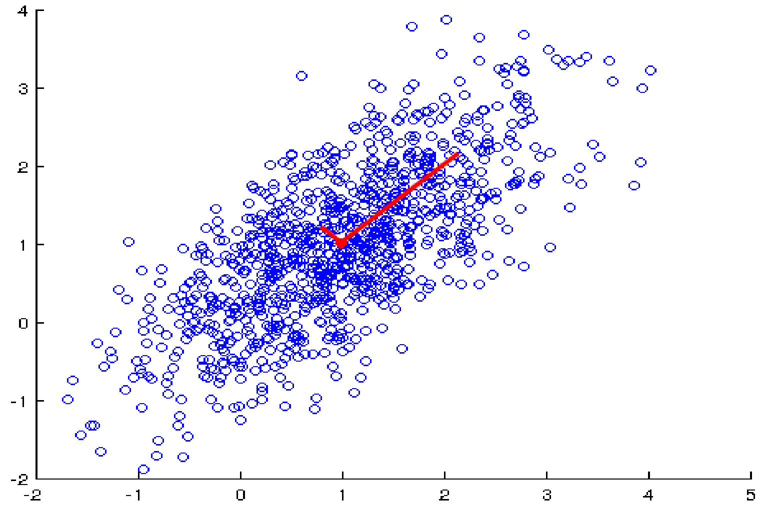
1. Compute covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \bar{\mathbf{x}})(\mathbf{x_n} - \bar{\mathbf{x}})^{\mathbf{T}}$
2. Compute its eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq \ldots \geq \lambda_D$
   and their eigenvectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_M}, \ldots \mathbf{u_D}$
3. Optimal hyperplane is span $\{\mathbf{u_i}\}_{\mathbf{i=1}}^{\mathbf{M}}$
4. Projected variance is $\lambda_1 + \lambda_2 + \ldots + \lambda_M$

Look familiar?



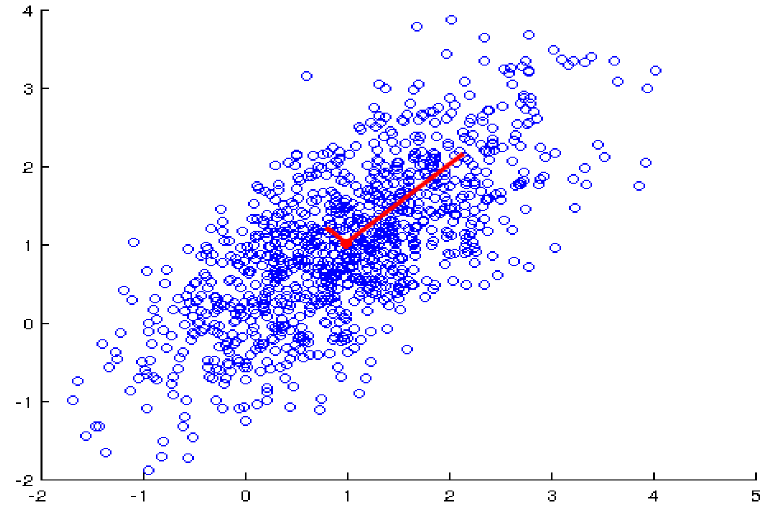
Computing principal components

- Eigenvalue decomposition shows that PCA is equivalent to
  - Fitting a Gaussian distribution to your data using mean and covariance
  - Using the eigenvectors of the covariance as principal directions



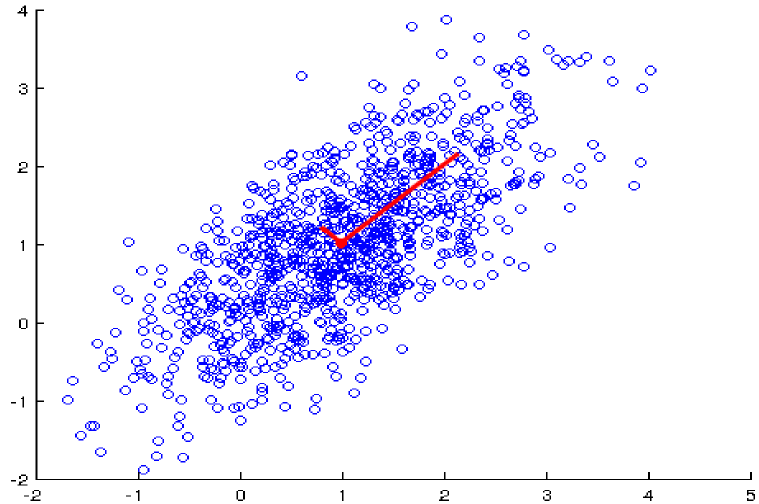
Computing principal components

- Eigenvalue decomposition shows that PCA is equivalent to
  - Fitting a Gaussian distribution to your data using mean and covariance
  - Using the eigenvectors of the covariance as principal directions
- What does that tell you about PCA?



Computing principal components

- Eigenvalue decomposition shows that PCA is equivalent to
  - Fitting a Gaussian distribution to your data using mean and covariance
  - Using the eigenvectors of the covariance as principal directions
- What does that tell you about PCA?



- Remember:

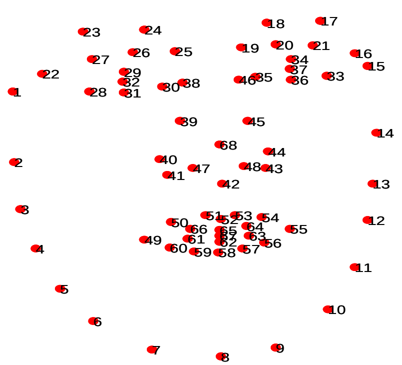
$$\mathbf{S} = \mathrm{R}_\theta * \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \dots & 0 \\ 0 & 0 & \dots & \lambda_D \end{pmatrix} R_\theta^{-1}$$

Case: Face shape – visualizing data variance

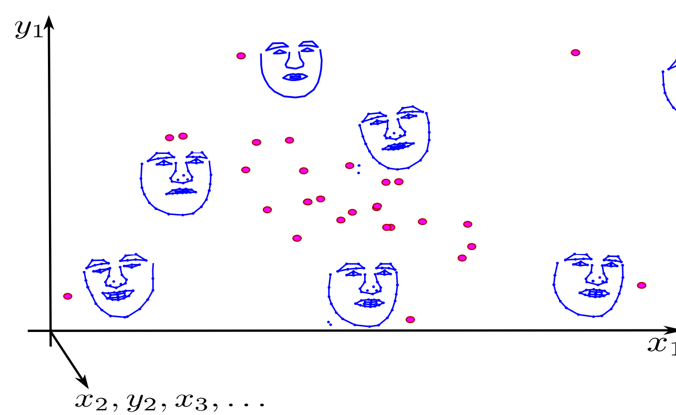
- Before doing statistics, let's describe the image with vectors
    1. Manual annotation of specific *landmark points*: A specific set of dots
    2. Connect the right dots
-

Case: Face shape – visualizing data variance

- Before doing statistics, let's describe the image with vectors
  1. Manual annotation of specific *landmark points*: A specific set of dots
  2. Connect the right dots
- Decide on an order of dots and record their coordinates in order



$$\rightsquigarrow (\underbrace{x_1, y_1}, \underbrace{x_2, y_2}, \dots, \underbrace{x_{68}, y_{68}}) \rightsquigarrow$$

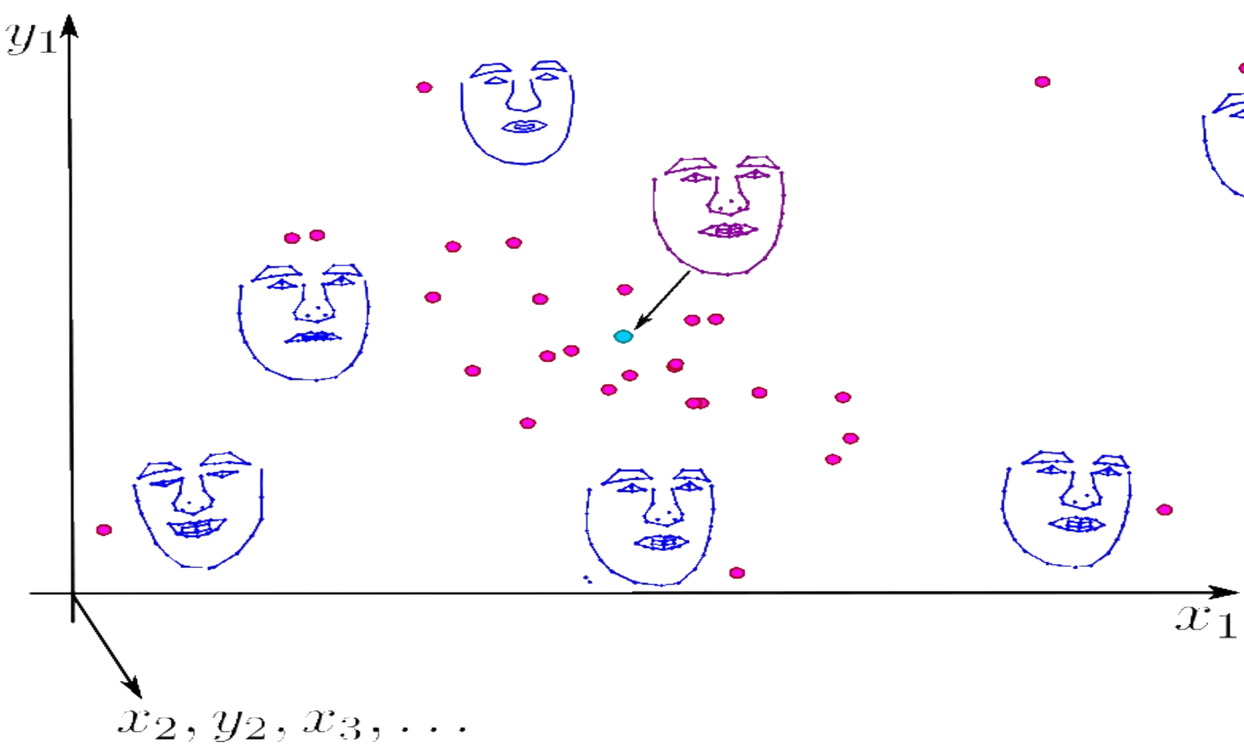


Case: Face shape – visualizing data variance

Denote ith face: $f^i = (x_1^i, y_1^i, x_2^i, y_2^i, \ldots, x_{68}^i, y_{68}^i)$

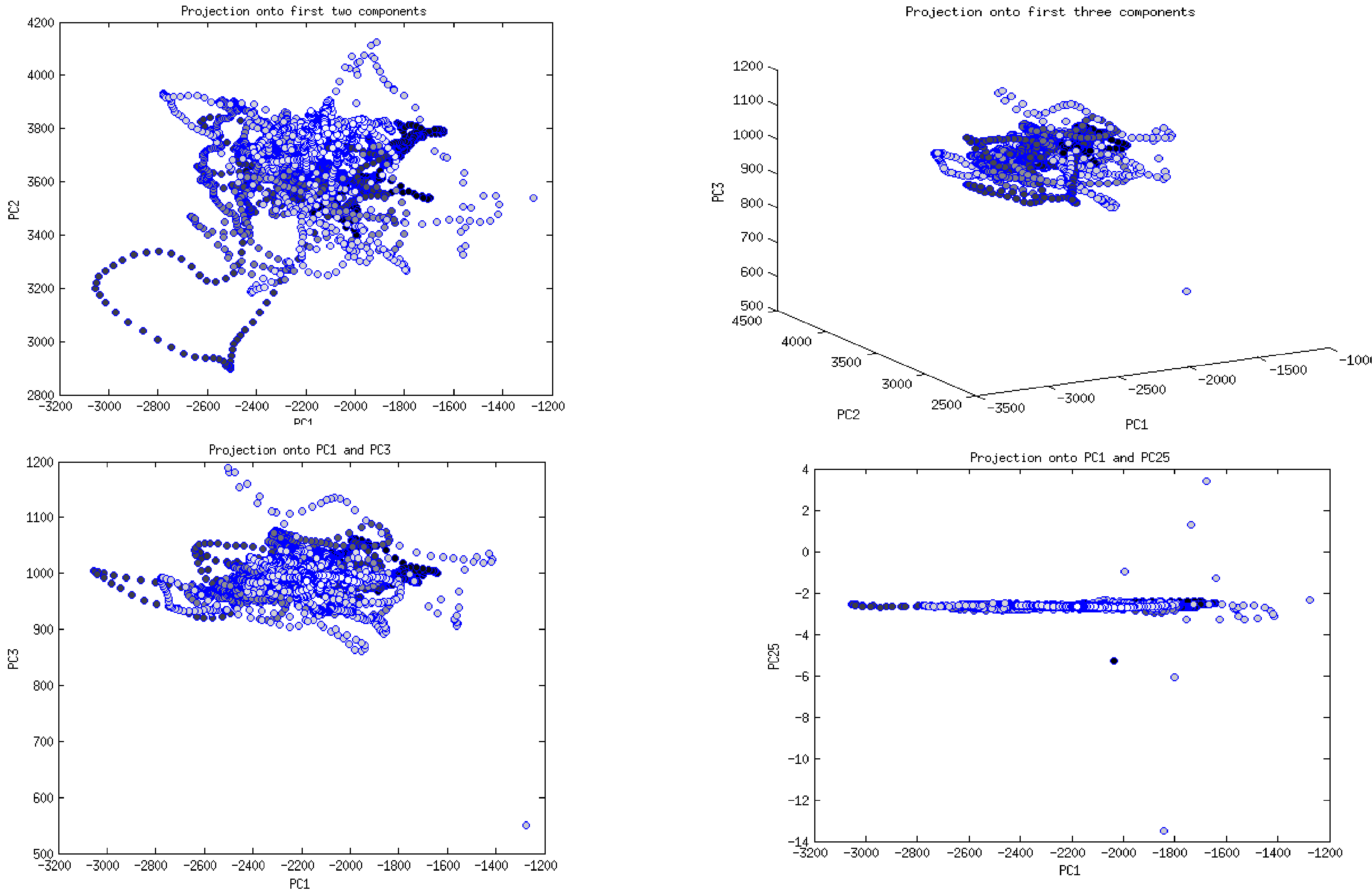
Mean face: $\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f^i = \frac{1}{N}\sum_{i=1}^{N}(x_1^i, y_1^i, \ldots, x_{68}^i, y_{68}^i)$

$= (\bar{x}_1, \bar{y}_1, \ldots, \bar{x}_{68}, \bar{y}_{68})$



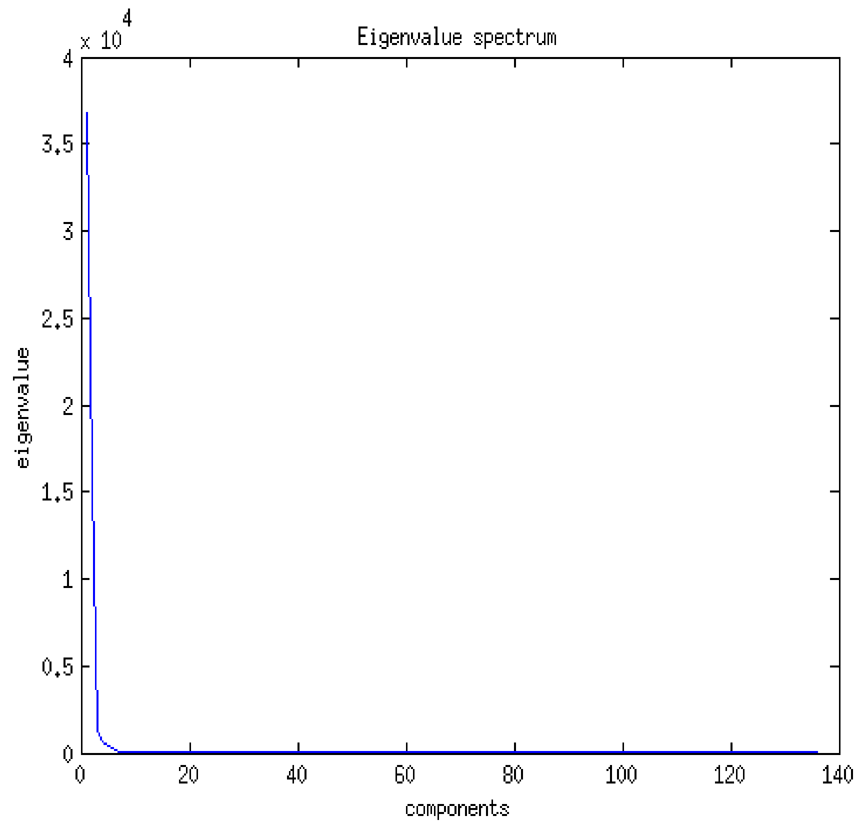
Application 1: Visualization of high dimensional data sets

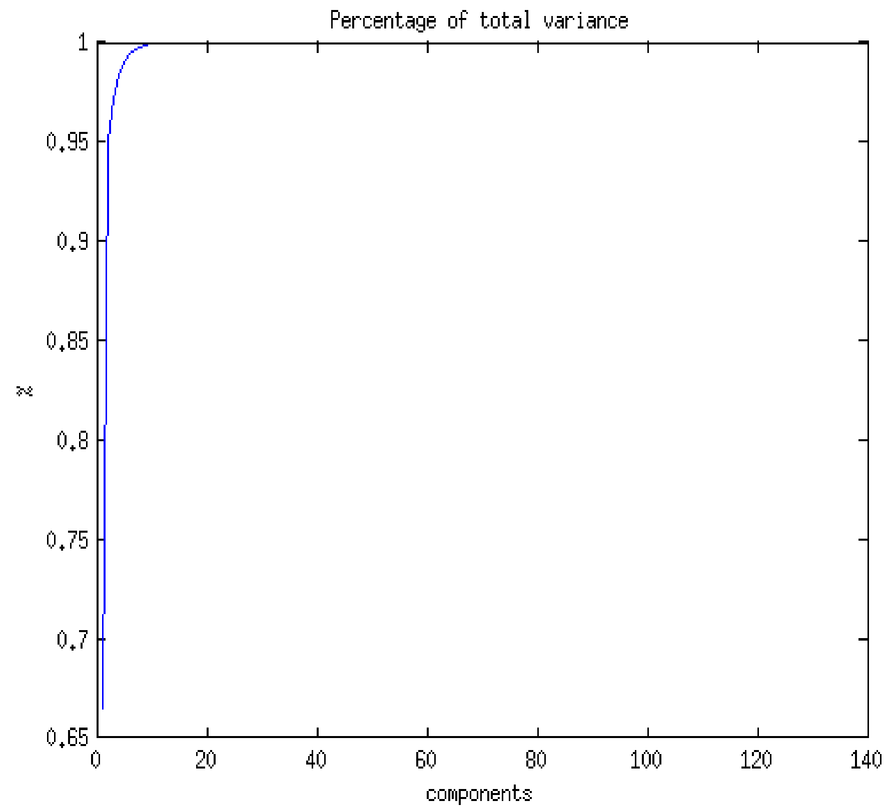
Color = video frame



How many eigenvectors should you use for a good representation of data?

Eigenvalues and total variance





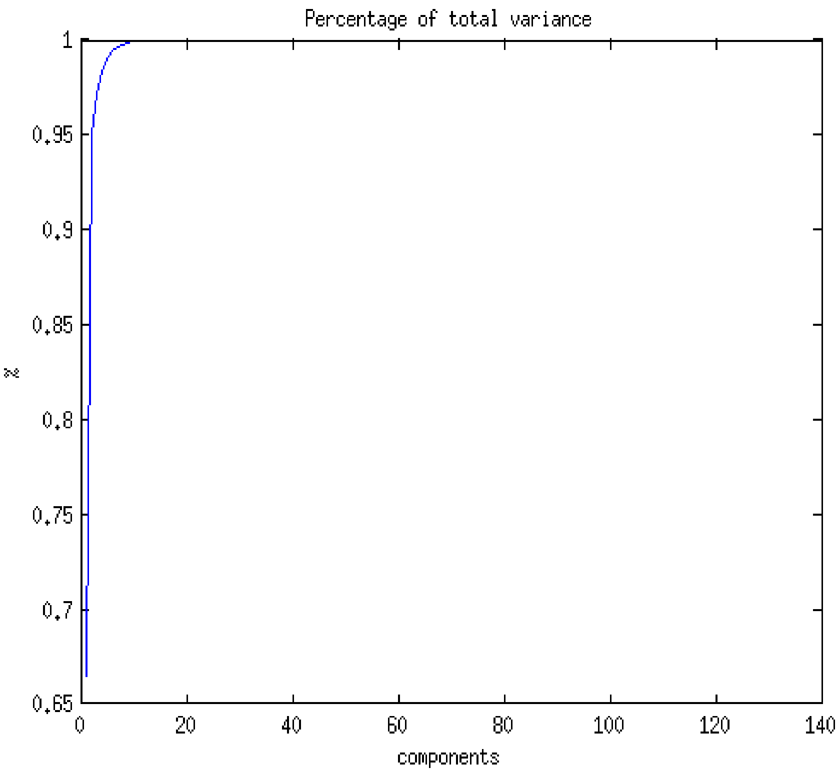
Total variation captured by M PCs

$$\sum_{i=1}^{M} \lambda_i$$

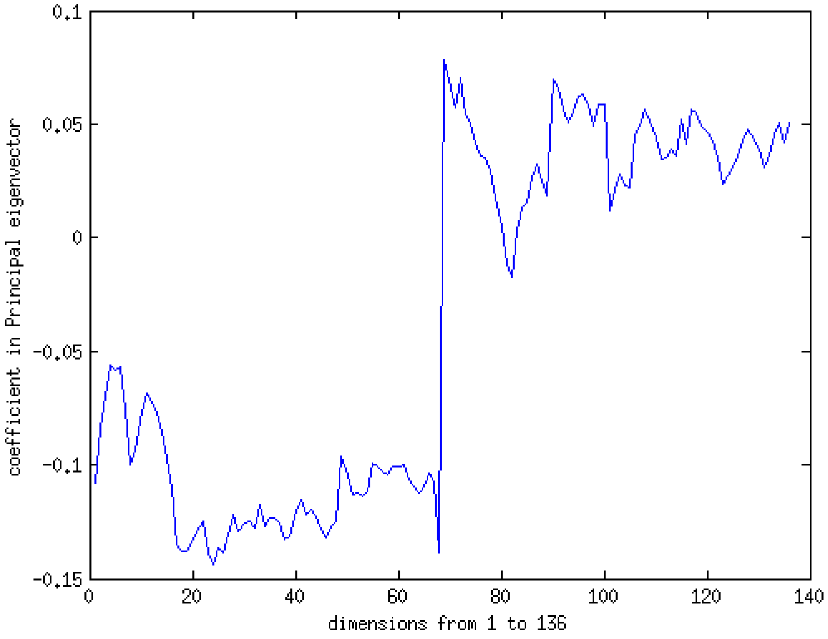
First two eigenvectors capture 95% of variation
First three capture 97%

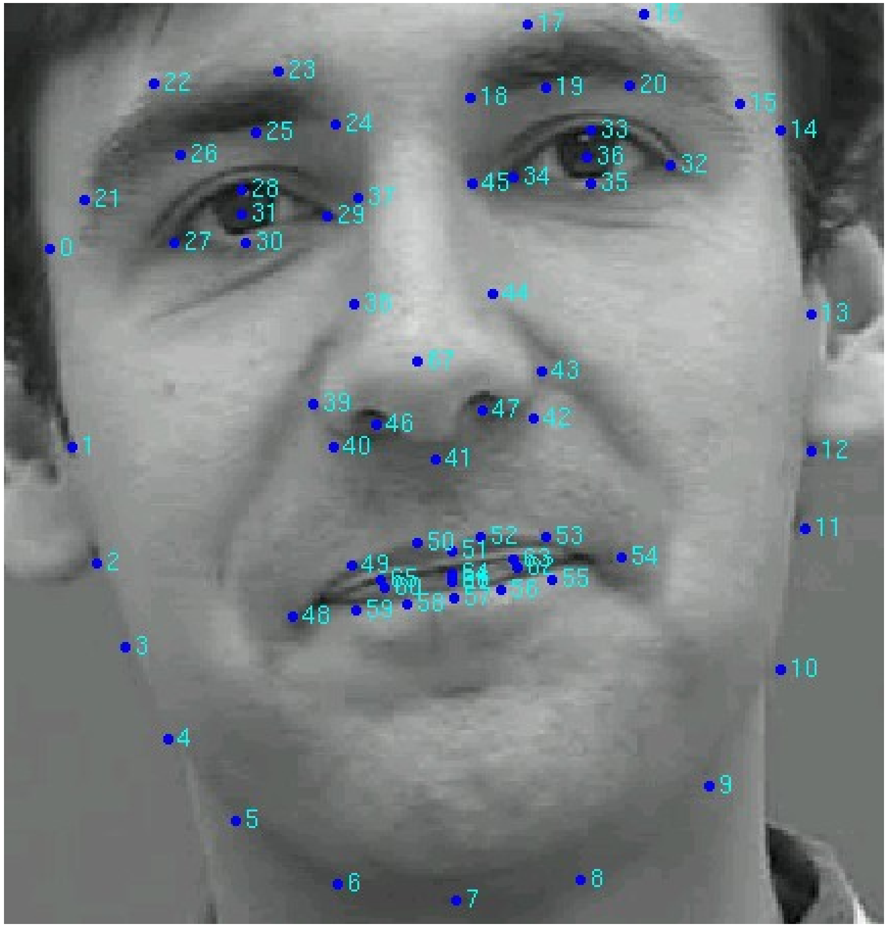
Application 2: Coding

- One face: 68 landmarks = 136 coordinates
- Faces along PC1 encoded with a single scalar value
- Faces within first M PCs encoded with M scalar values
- How many scalar values do you need to describe any face?



Visualizing the 1st principal eigenvector

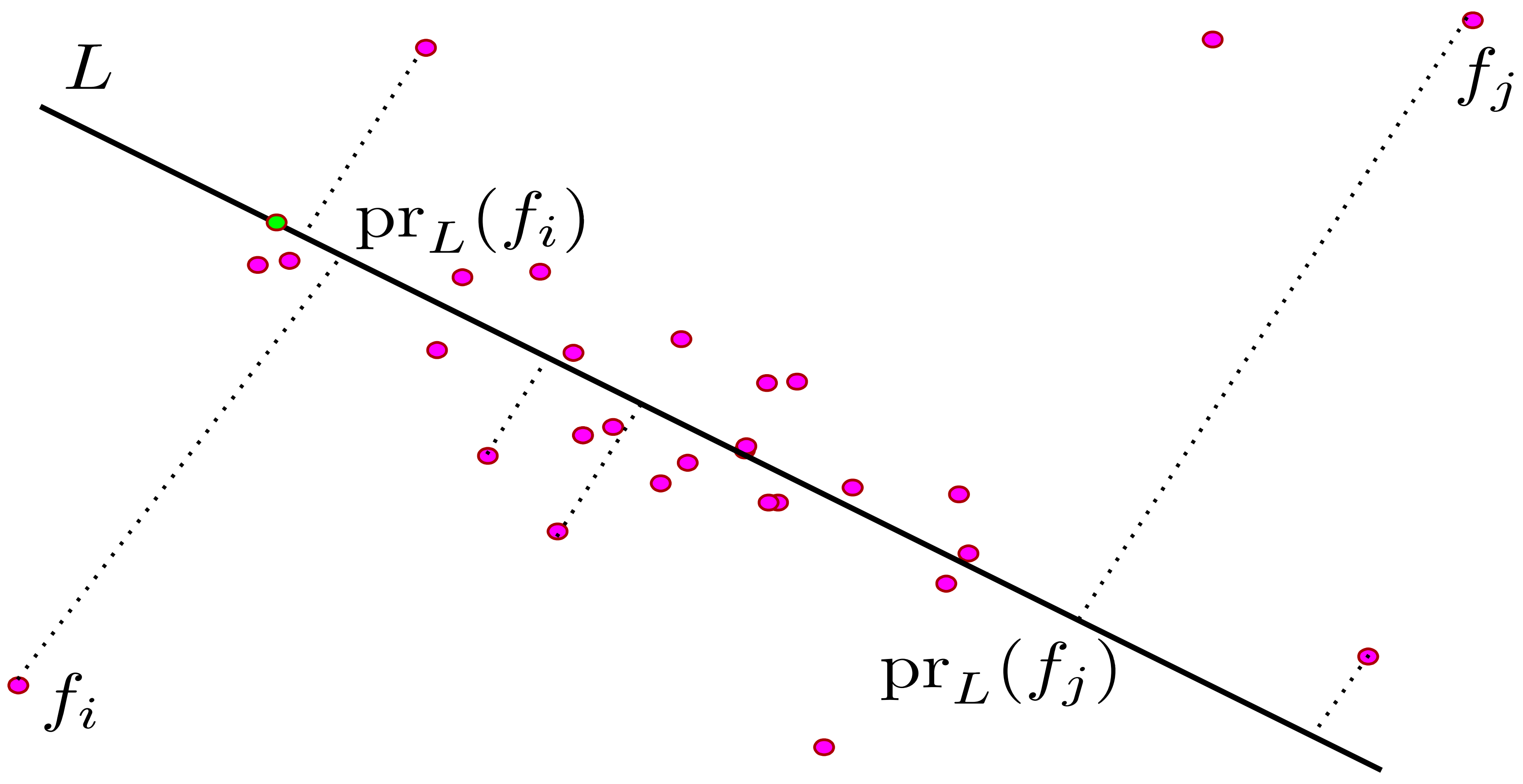




Look back at movie

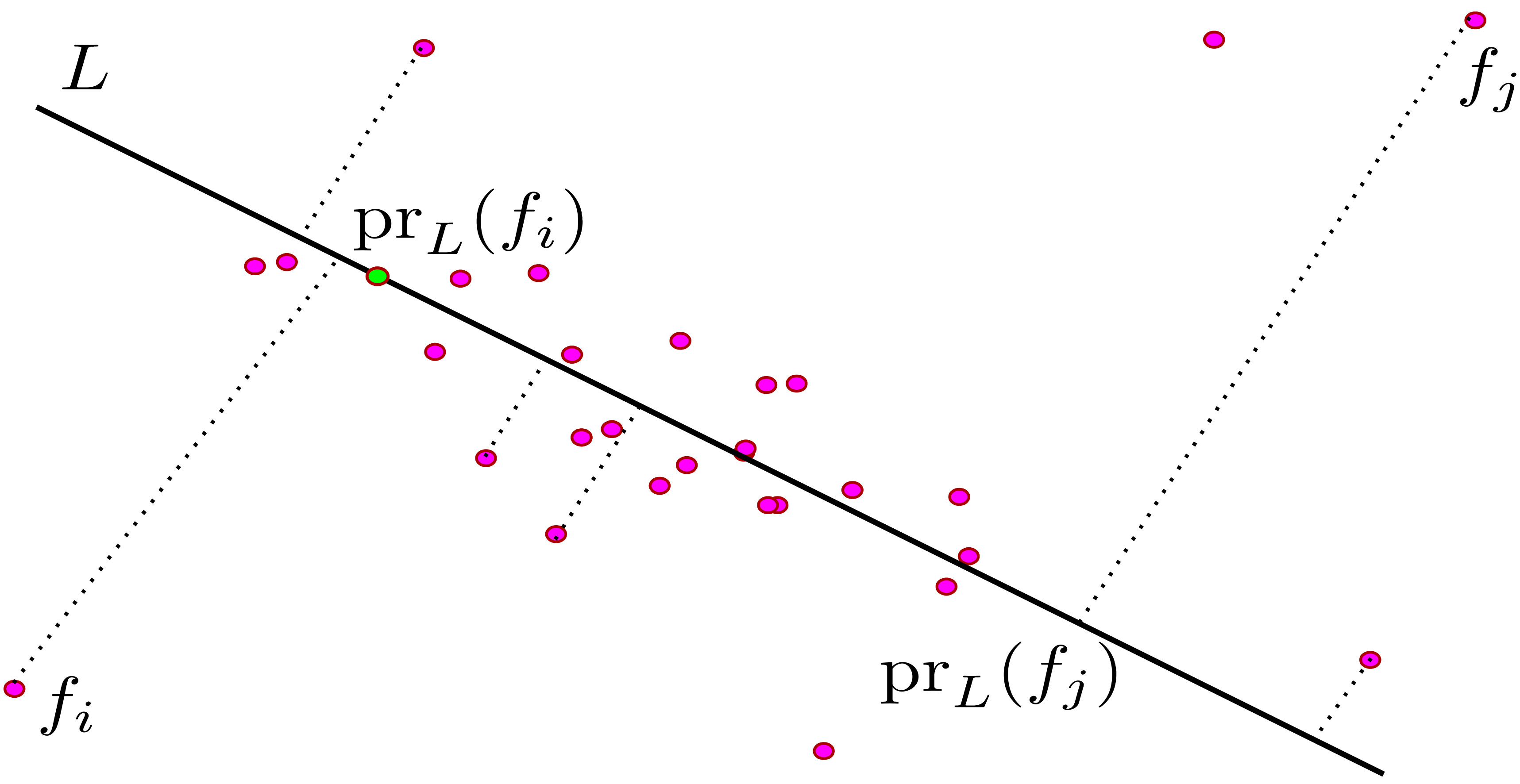
Application 3: Eigenfaces – visualizing data variance

- Look at points along PC1 – they are faces



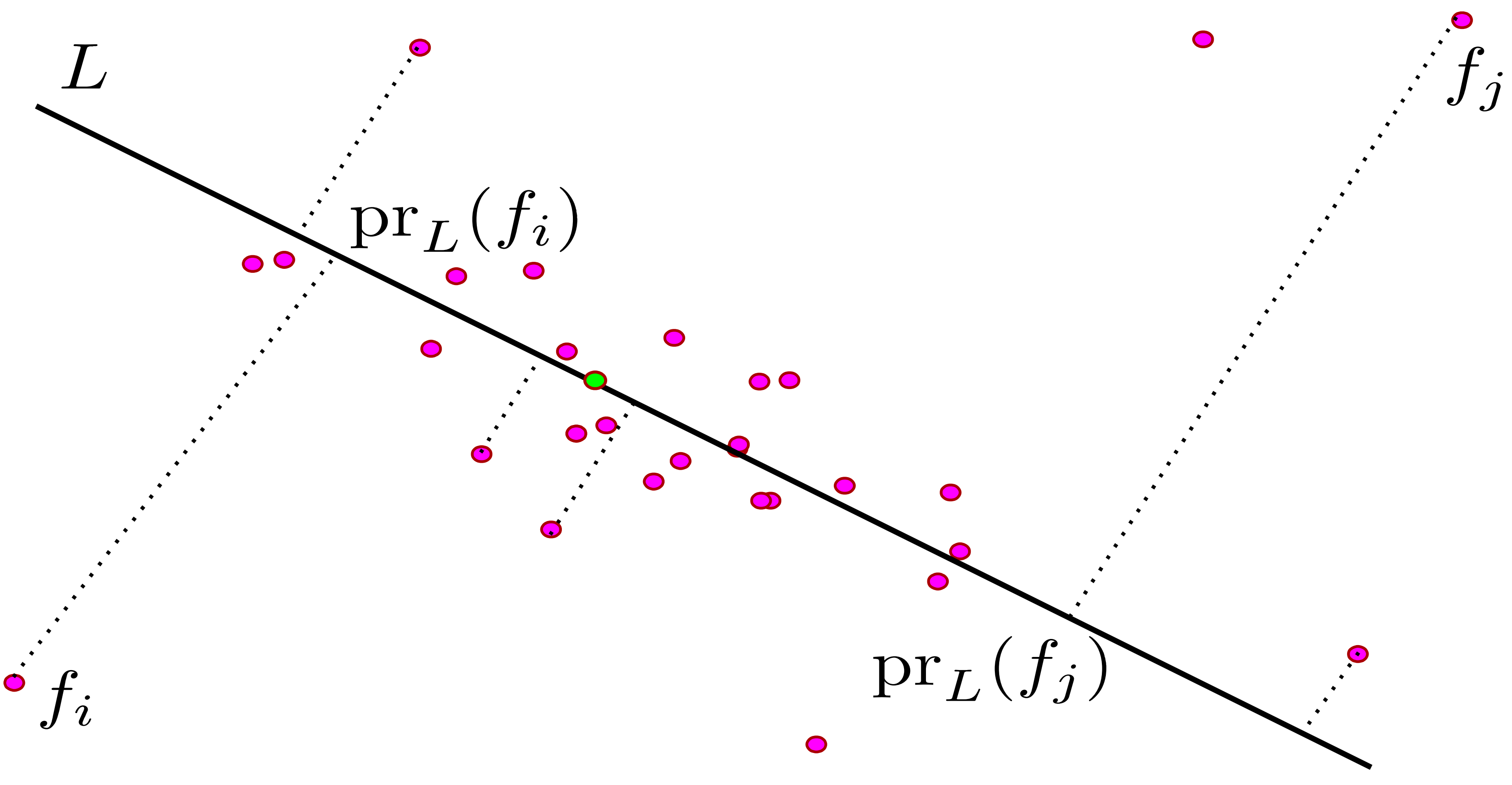
Application 3: Eigenfaces – visualizing data variance

- Look at points along PC1 – they are faces



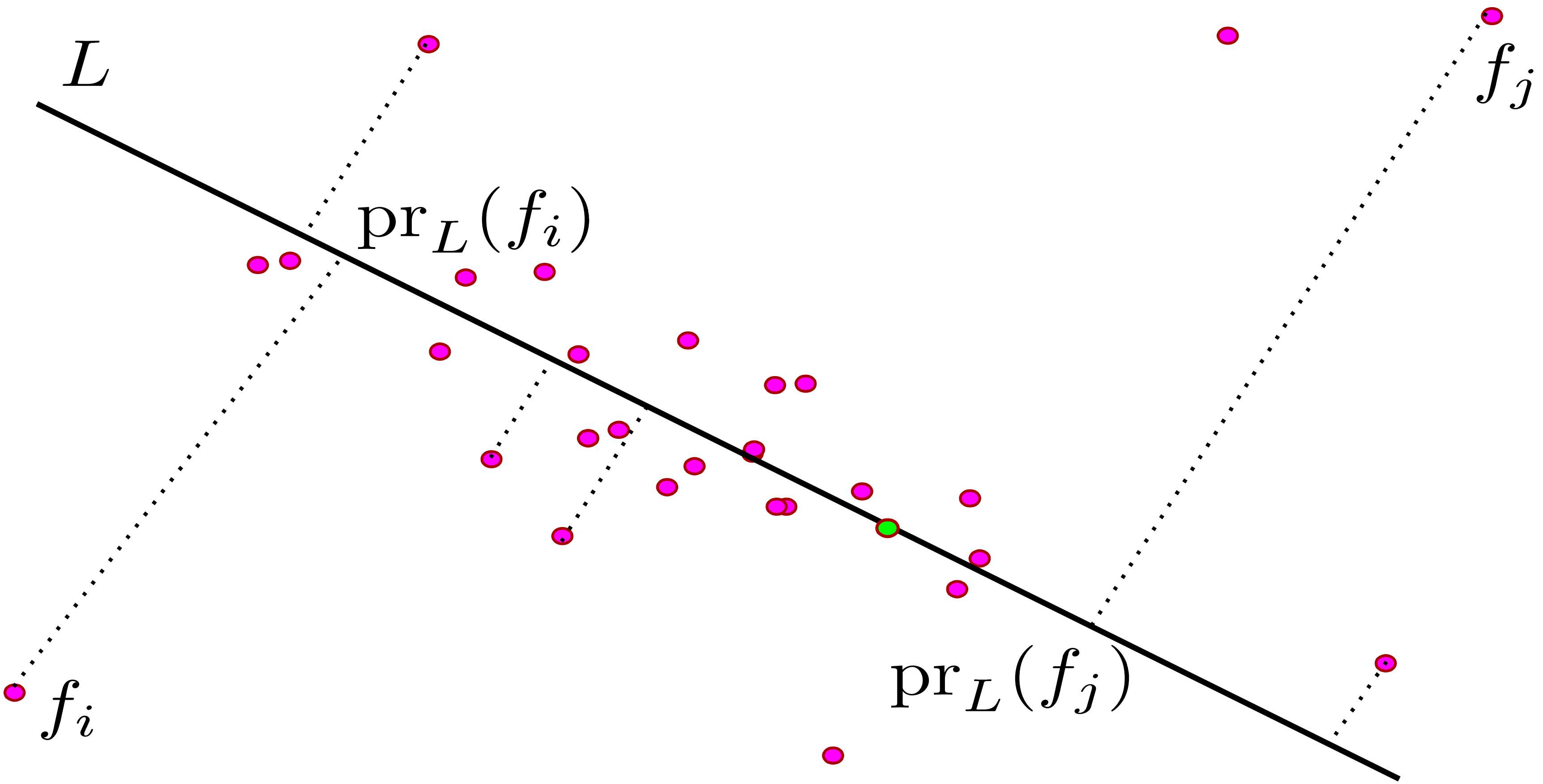
Application 3: Eigenfaces – visualizing data variance

- Look at points along PC1 – they are faces



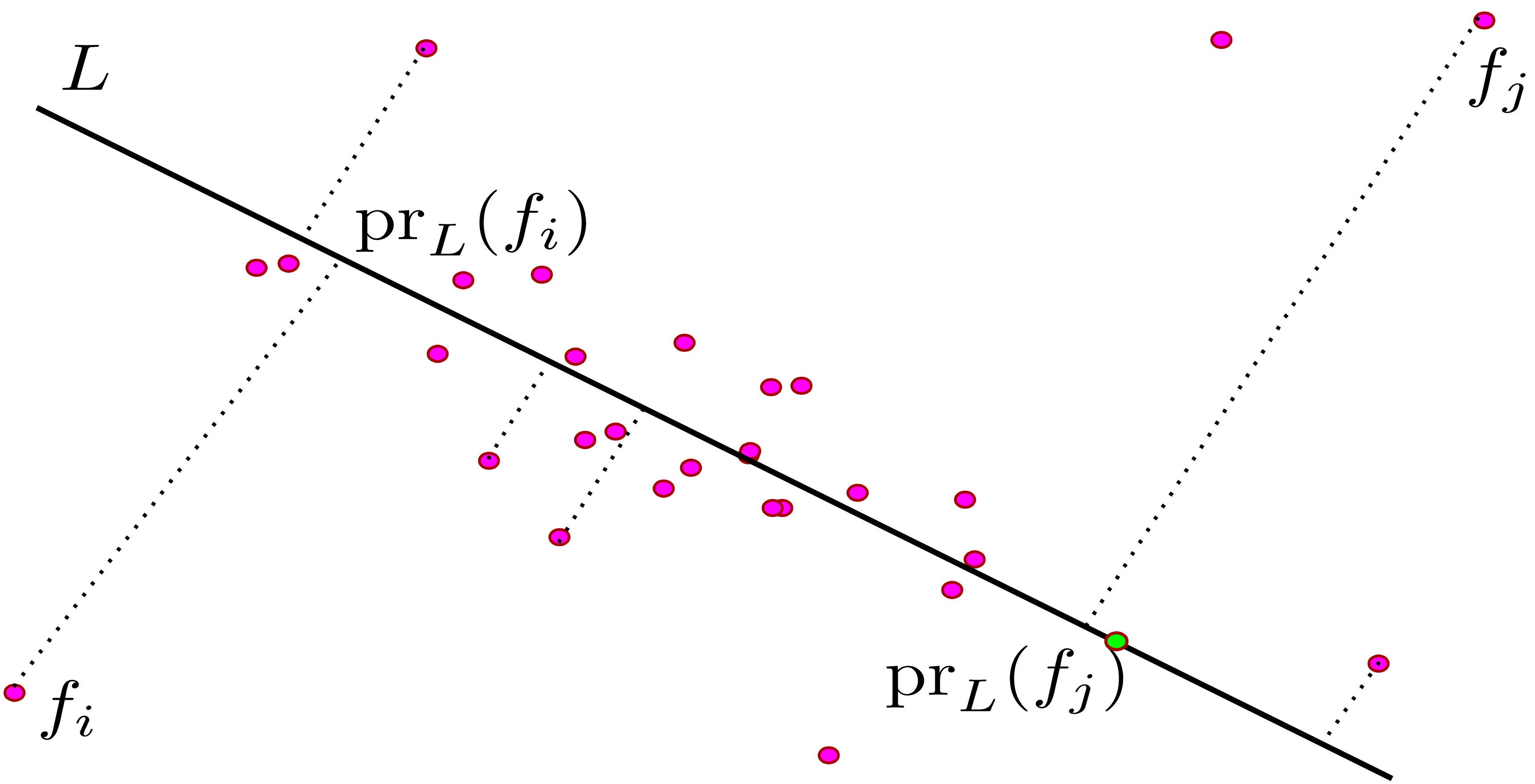
Application 3: Eigenfaces – visualizing data variance

- Look at points along PC1 – they are faces



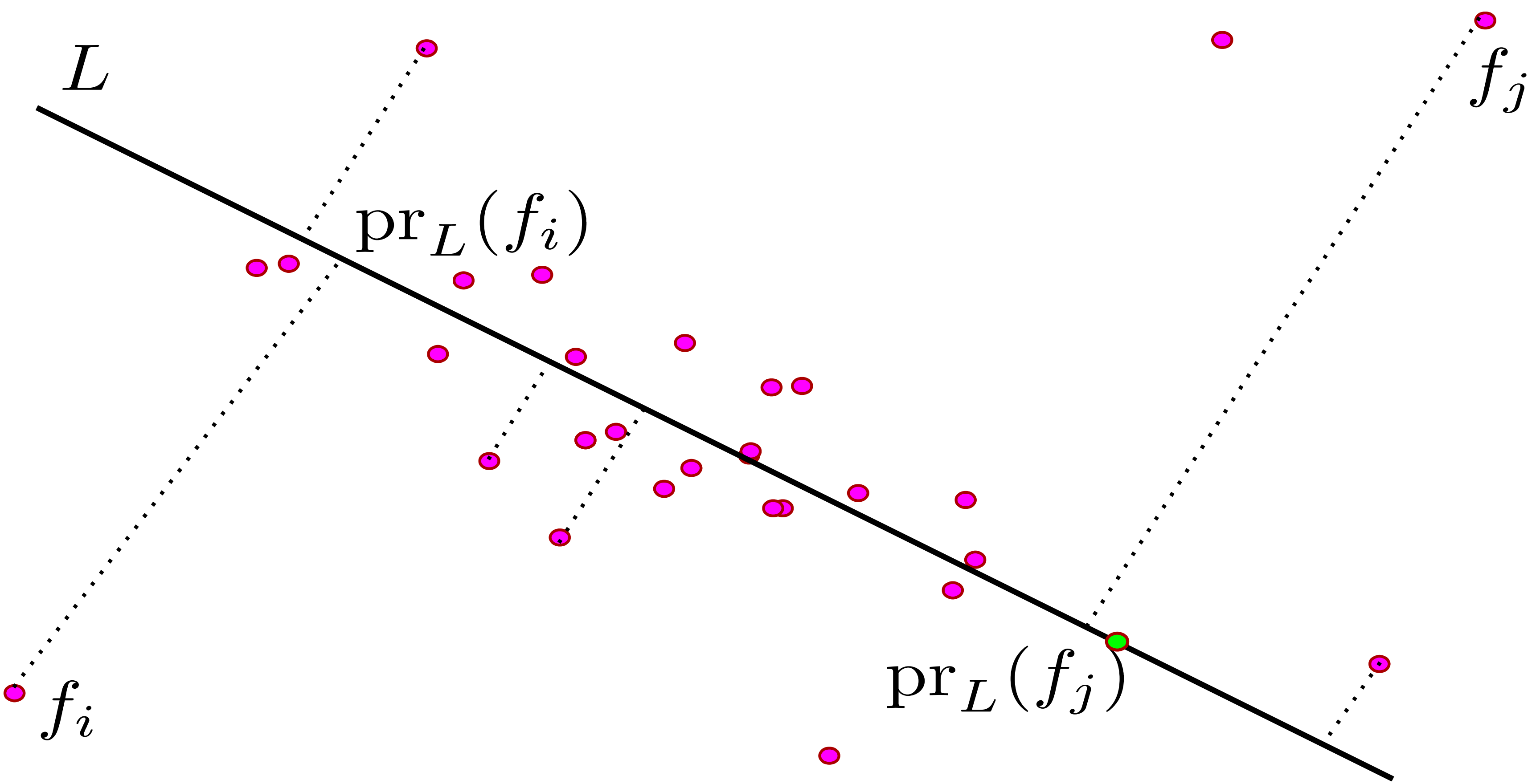
Application 3: Eigenfaces – visualizing data variance

- Look at points along PC1 – they are faces



Application 3: Eigenfaces – visualizing data variance

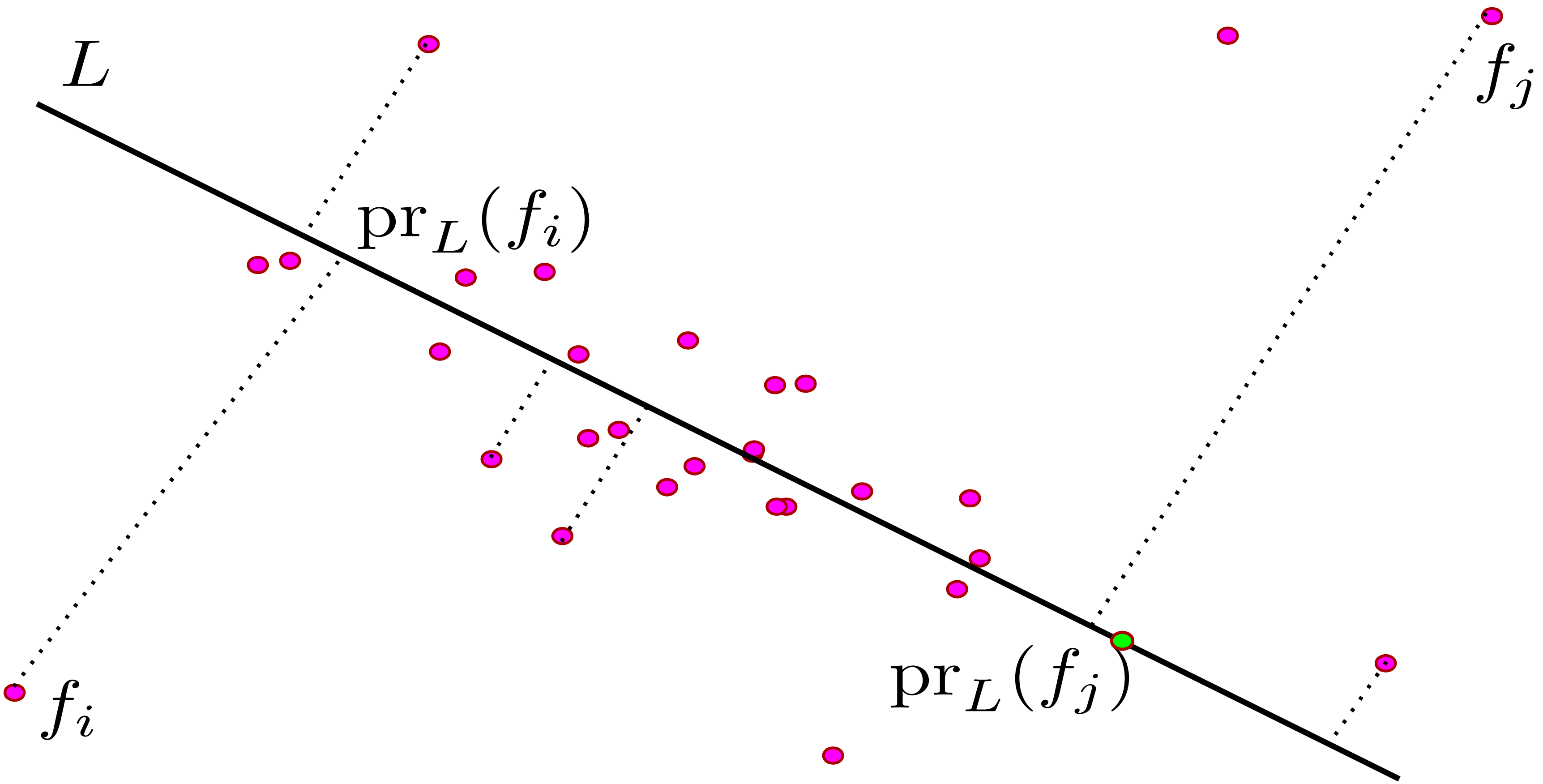
- Look at points along PC1 – they are faces



Play movie!

Application 3: Eigenfaces – visualizing data variance

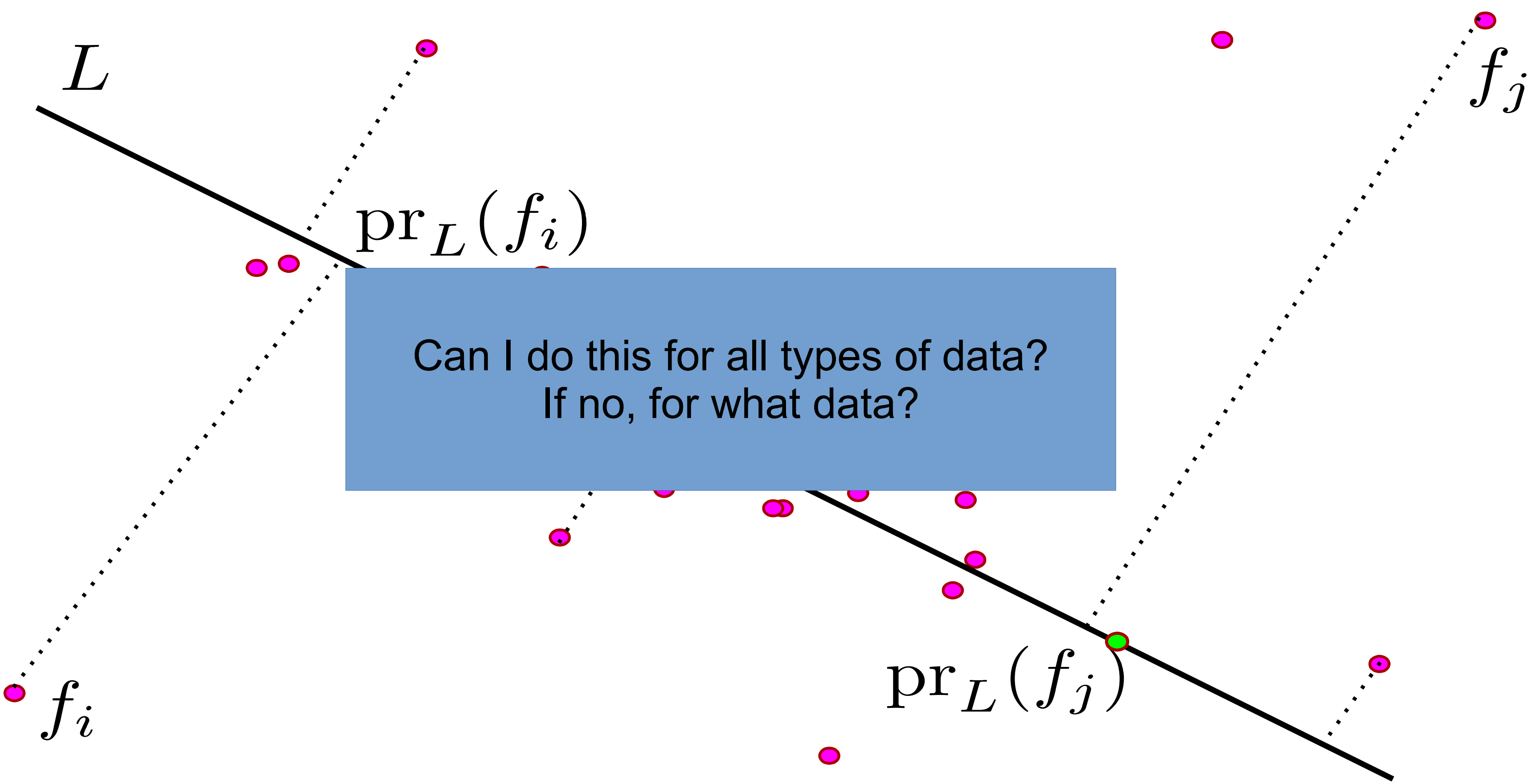
- Look at points along higher PCs



Play movie!

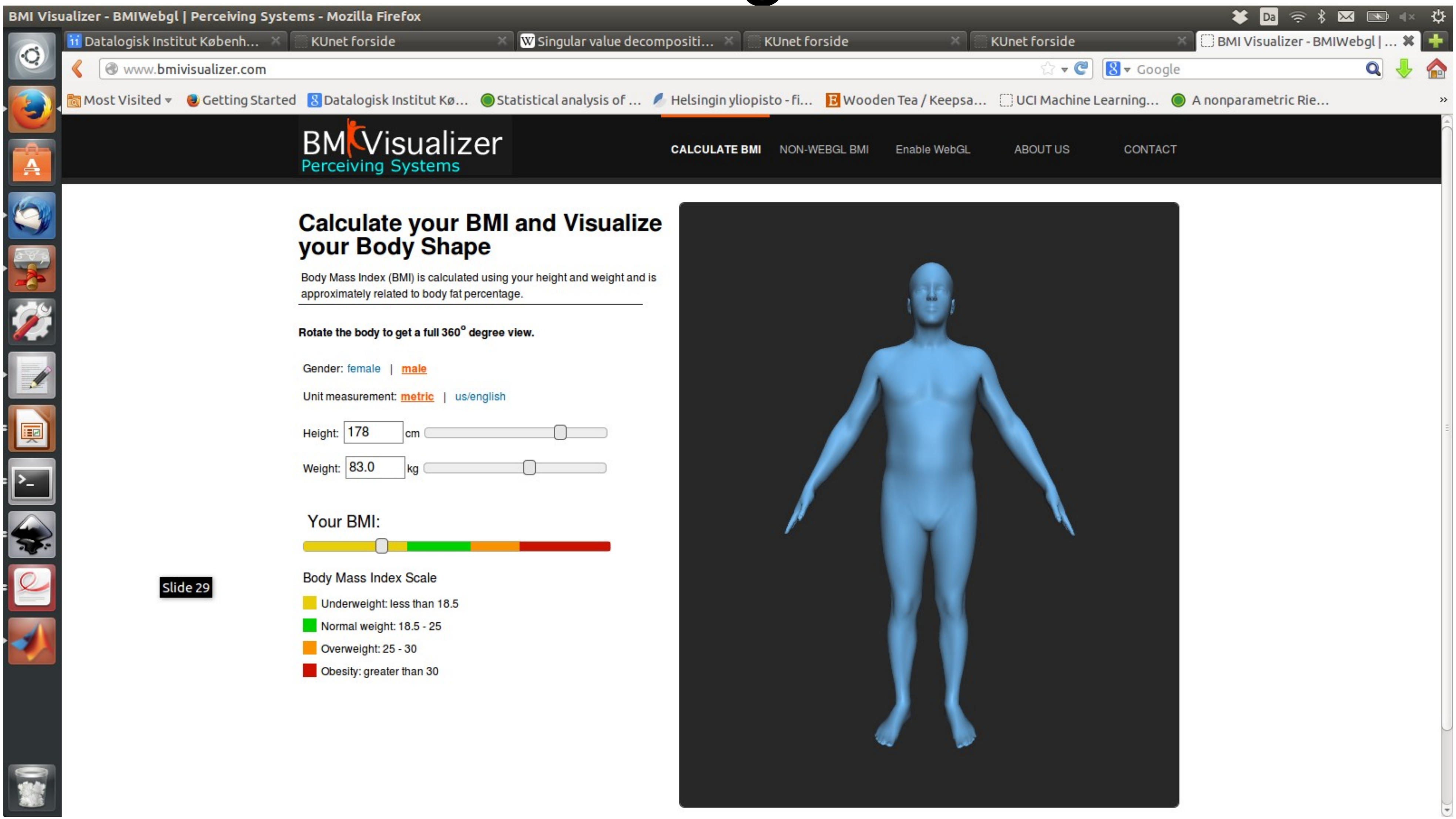
Application 3: Eigenfaces – visualizing data variance

- Look at points along higher PCs



Play movie!

Case 2: BMIvisualizer: PCA + regression



Application 4: PCA in preprocessing

Standardization

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

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

- Covariance matrix becomes the correlation matrix

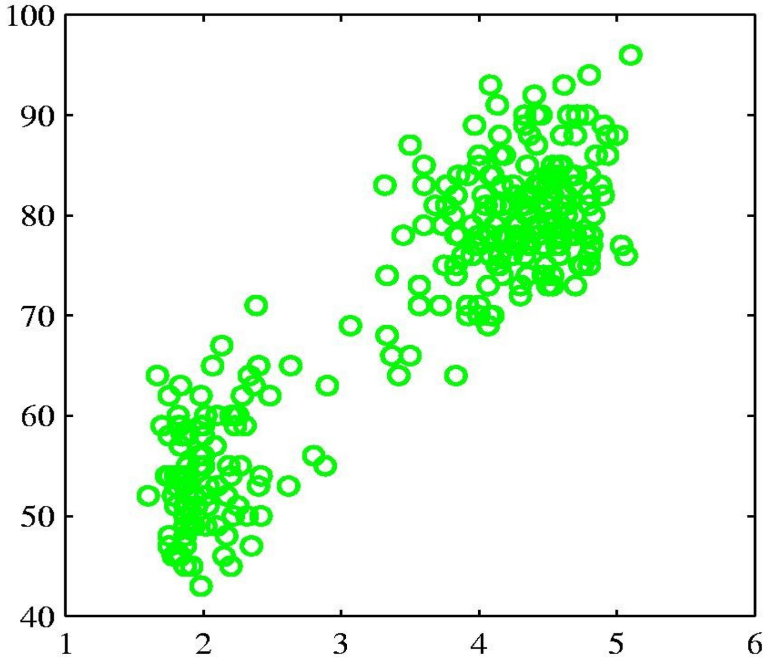
$$\rho_{ij} = \tilde{\mathbf{S}}_{ij} = \frac{1}{N} \sum_{n=1}^{N} \frac{(x_{ni} - \bar{x}_i)}{\sigma_i} \frac{(x_{nj} - \bar{x}_j)}{\sigma_j}$$

- But PCA does more than this!
  We can decorelate variables as we will see in a minute.

Application 4: PCA in preprocessing

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

Application 4: PCA in preprocessing

PCA in preprocessing: Whitening

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

$$\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u_D}) \quad \text{and} \quad \mathbf{L} = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ & \ddots & & \\ 0 & 0 & \ldots & \lambda_D \end{pmatrix}$$

Translate, rotate, and scale the data into the coordinate system of the PCs: $\mathbf{y_n} = \mathbf{L^{-1/2}U^T(x_n - x)}$

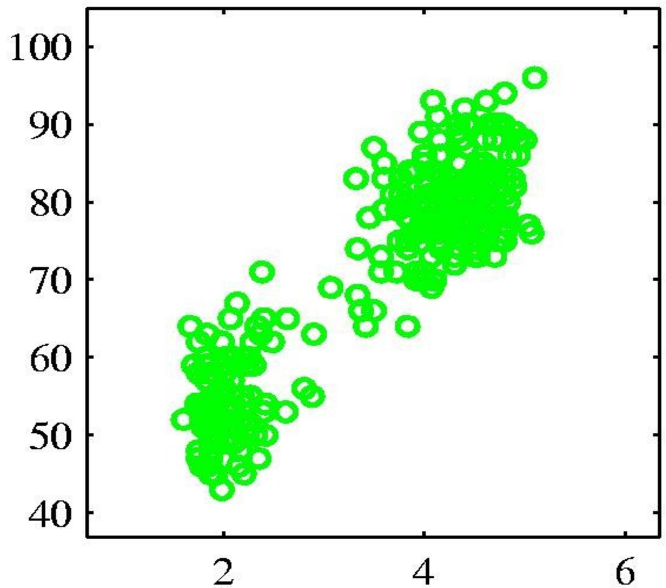
In this coordinate system the data is zero mean and have identity covariance

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

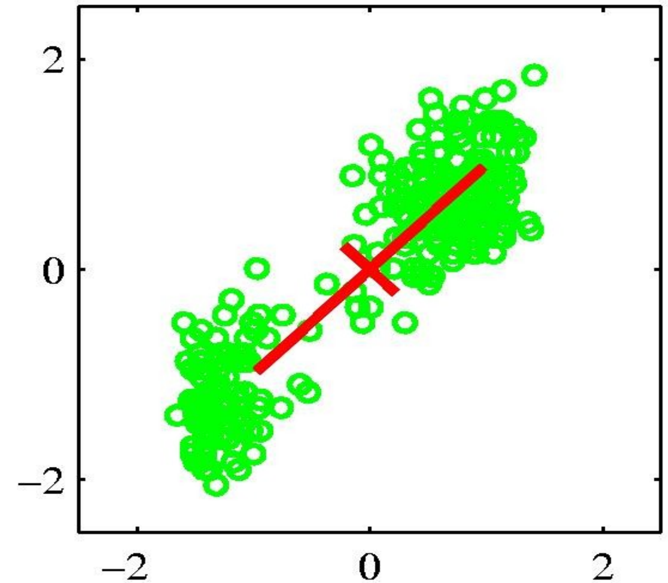
$$= \mathbf{L}^{-1/2}\mathbf{U}^T\mathbf{SUL}^{-1/2} = \mathbf{L^{-1/2}LL^{-1/2}} = \mathbf{I}$$

This is referred to as whitening the data.

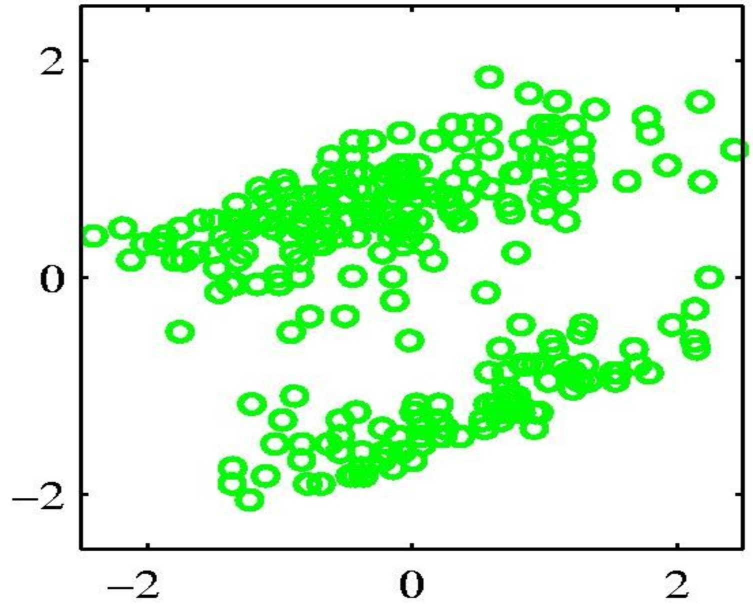
Application 4: PCA in preprocessing



orginal



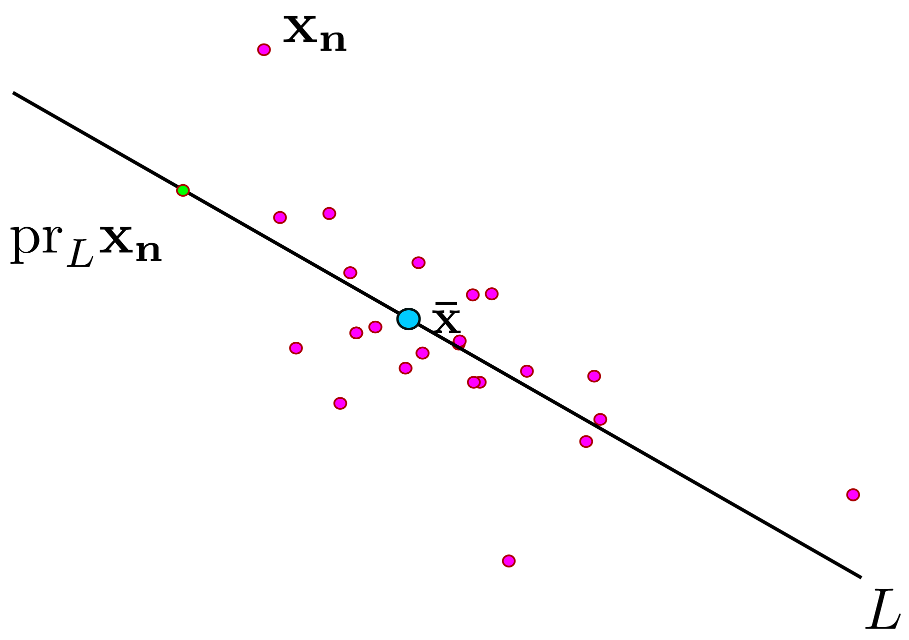
standardized



whitened

Equivalence of error minimization and variance maximization

Task: Show that $\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$
gives the same solution as $\mathrm{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$
when the second L is constrained to pass through $\bar{\mathbf{x}}$



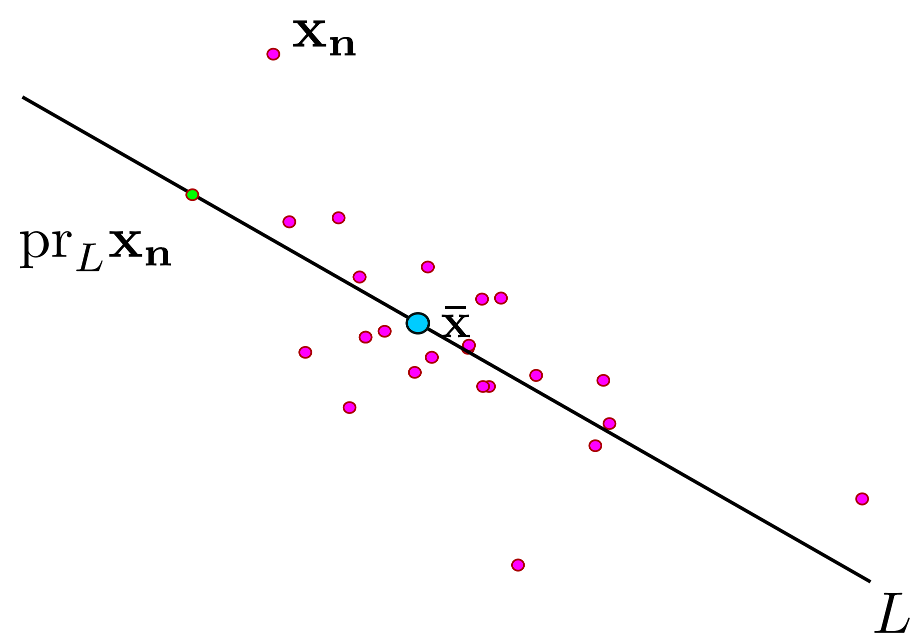
Equivalence of error minimization and variance maximization

Task: Show that $\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_{\mathbf{L}}\mathbf{x_n}\|^2$
gives the same solution as $\operatorname{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$
when the second L is constrained to pass through $\bar{\mathbf{x}}$

Claim 1: The solution L of

$$\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_{\mathbf{L}}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)



Equivalence of error minimization and variance maximization

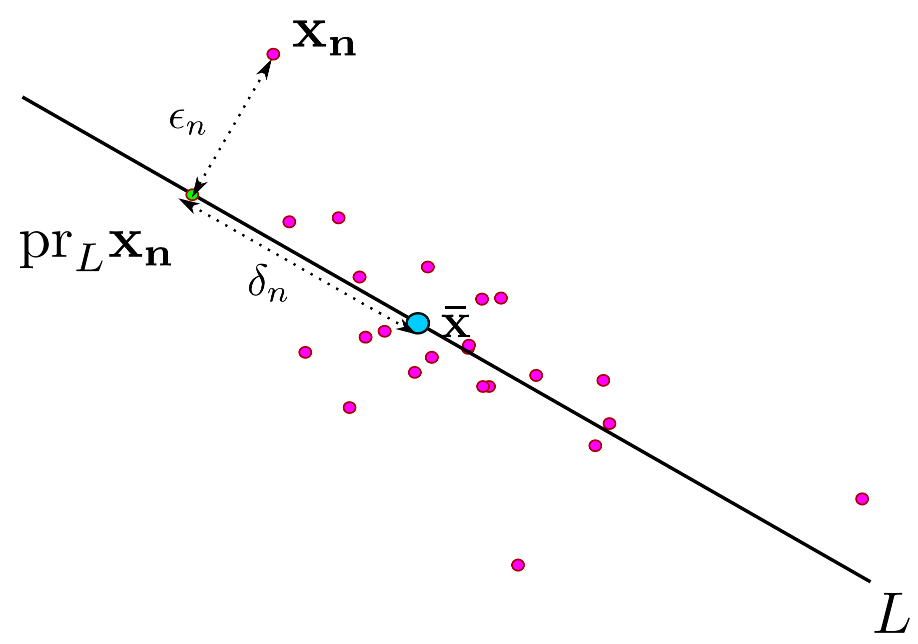
Task: Show that $\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$
gives the same solution as $\text{argmax}_L \sum_{n=1}^{N} \|\text{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$
when the second L is constrained to pass through $\bar{\mathbf{x}}$

$$\epsilon_n = \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|$$
$$\delta_n = \|\text{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$$

Claim 1: The solution L of

$$\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)



Equivalence of error minimization and variance maximization

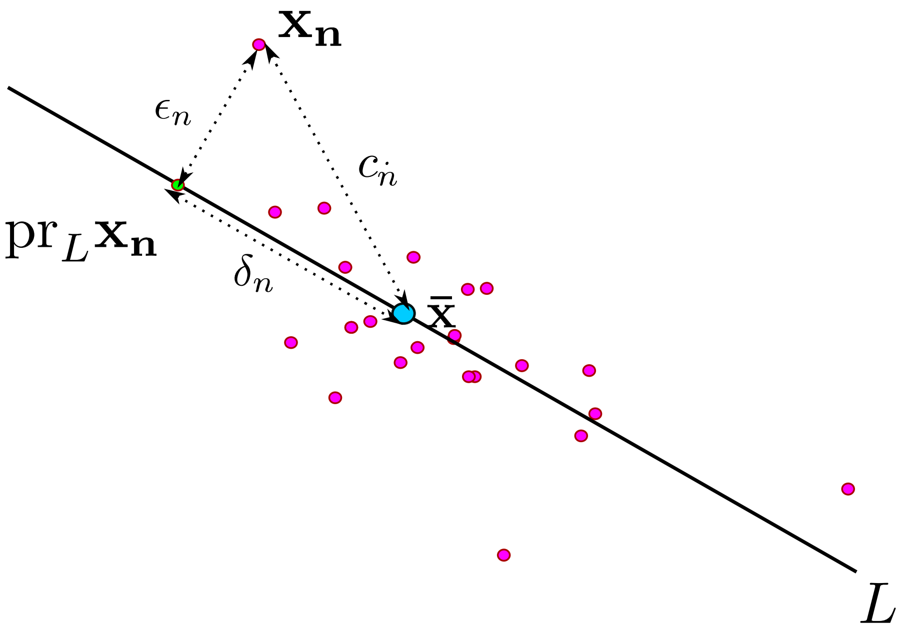
Task: Show that $\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$ gives the same solution as $\text{argmax}_L \sum_{n=1}^{N} \|\text{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$ when the second L is constrained to pass through $\bar{\mathbf{x}}$

Claim 1: The solution L of

$$\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)

$$\epsilon_n = \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|$$
$$\delta_n = \|\text{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$$
$$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\|$$



Equivalence of error minimization and variance maximization

Task: Show that $\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$ gives the same solution as $\text{argmax}_L \sum_{n=1}^{N} \|\text{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$ when the second L is constrained to pass through $\bar{\mathbf{x}}$

Claim 1: The solution L of

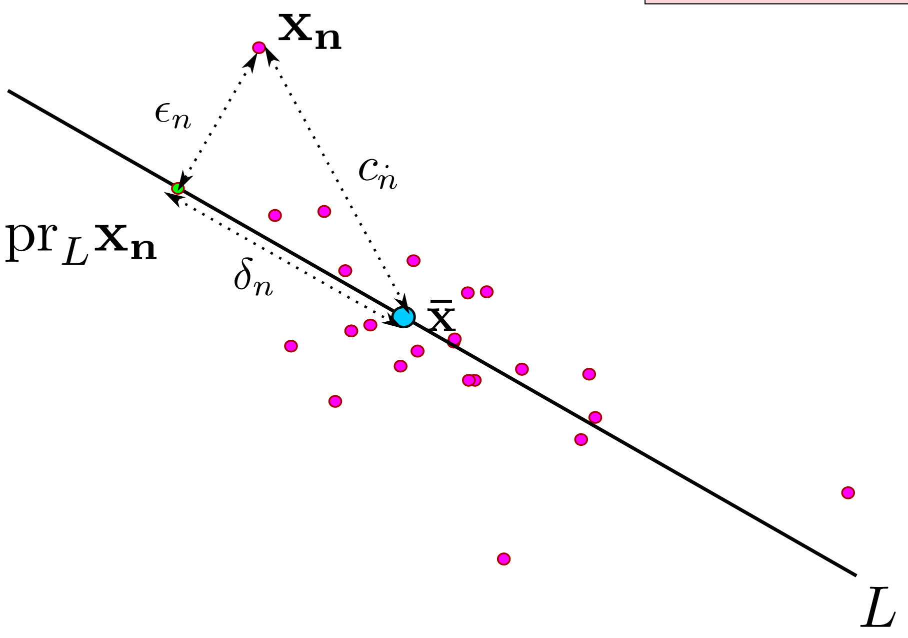
$$\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)

$\epsilon_n = \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|$

$\delta_n = \|\text{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$

$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\|$ Not dependent on L!



Equivalence of error minimization and variance maximization

Task: Show that $\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$
gives the same solution as $\mathrm{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$
when the second L is constrained to pass through $\bar{\mathbf{x}}$

Claim 1: The solution L of

$$\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

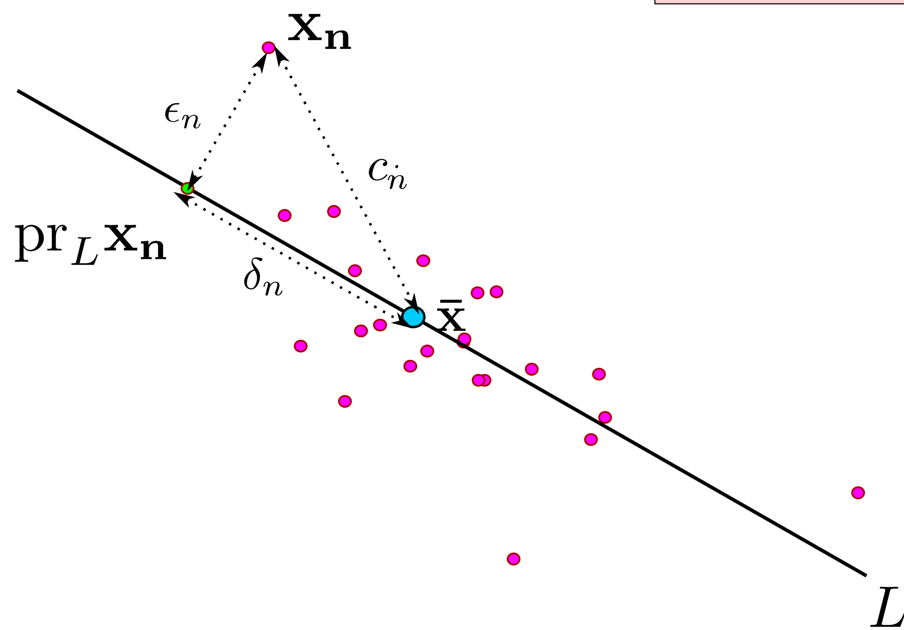
contains $\bar{\mathbf{x}}$ (we show this in a minute)

Pythagoras gives $\epsilon_n^2 + \delta_n^2 = c_n^2$ for all $n = 1, \ldots, N$

$\epsilon_n = \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|$

$\delta_n = \|\mathrm{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$

$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\|$ Not dependent on L!



Equivalence of error minimization and variance maximization

Task: Show that $\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$ gives the same solution as $\mathrm{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$ when the second L is constrained to pass through $\bar{\mathbf{x}}$

$\epsilon_n = \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|$

$\delta_n = \|\mathrm{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$

$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\|$ Not dependent on L!

Claim 1: The solution L of

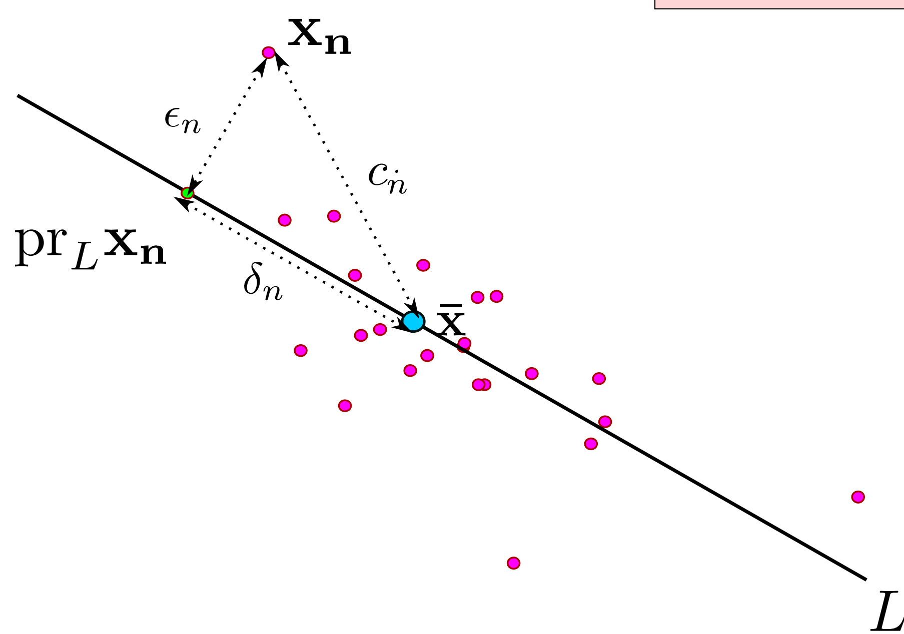
$$\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)

Pythagoras gives $\epsilon_n^2 + \delta_n^2 = c_n^2$ for all $n = 1, \ldots, N$

From that we see:

$$\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_L\mathbf{x}_n\|^2 = \mathrm{argmin}_L \sum_{n=1}^{N} \epsilon_n^2$$



Equivalence of error minimization and variance maximization

Task: Show that $\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$ gives the same solution as $\text{argmax}_L \sum_{n=1}^{N} \|\text{pr}_L \mathbf{x_n} - \bar{\mathbf{x}}\|^2$ when the second L is constrained to pass through $\bar{\mathbf{x}}$

Claim 1: The solution L of

$$\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)

Pythagoras gives $\epsilon_n^2 + \delta_n^2 = c_n^2$ for all $n = 1, \ldots, N$

From that we see:

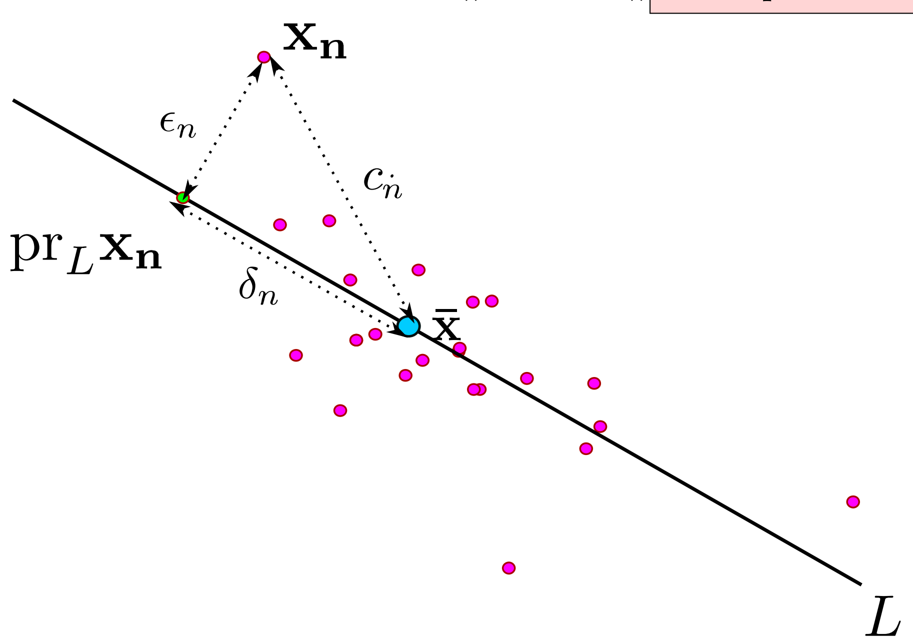
$$\text{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \text{pr}_L \mathbf{x}_n\|^2 = \text{argmin}_L \sum_{n=1}^{N} \epsilon_n^2$$

$$= \text{argmin}_L \sum_{n=1}^{N} (c_n^2 - \delta_n^2)$$

$$\epsilon_n = \|\mathbf{x_n} - \text{pr}_\mathbf{L}\mathbf{x_n}\|$$

$$\delta_n = \|\text{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$$

$$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\|$$ Not dependent on L!



Equivalence of error minimization and variance maximization

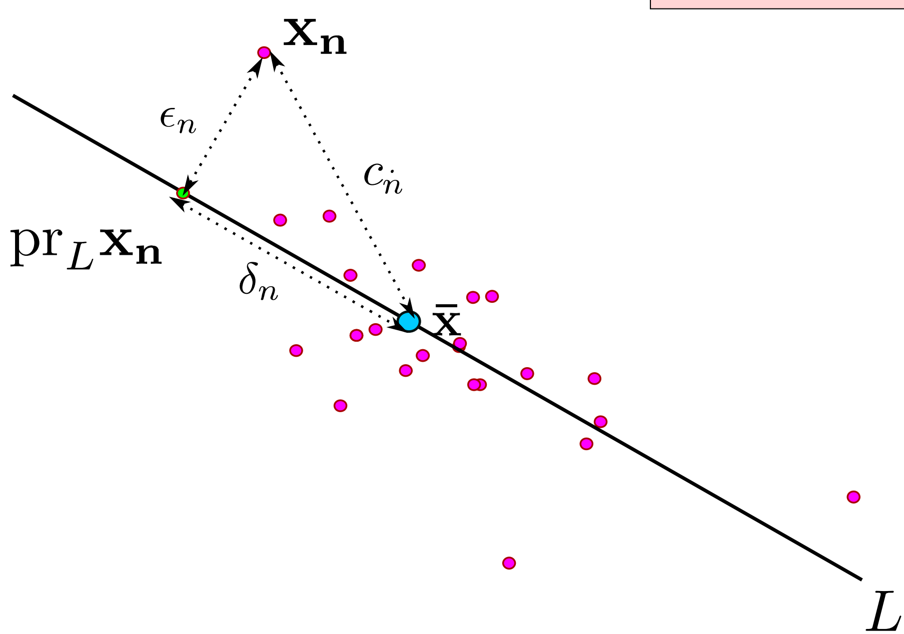
Task: Show that $\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$
gives the same solution as $\mathrm{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$
when the second L is constrained to pass through $\bar{\mathbf{x}}$

$$\epsilon_n = \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|$$
$$\delta_n = \|\mathrm{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$$
$$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\| \quad \boxed{\text{Not dependent on } L!}$$

Claim 1: The solution L of

$$\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)



Pythagoras gives $\epsilon_n^2 + \delta_n^2 = c_n^2$ for all $n = 1, \ldots, N$

From that we see:

$$\mathrm{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_L\mathbf{x}_n\|^2 = \mathrm{argmin}_L \sum_{n=1}^{N} \epsilon_n^2$$
$$= \mathrm{argmin}_L \sum_{n=1}^{N} (c_n^2 - \delta_n^2)$$
$$= \mathrm{argmin}_L - \sum_{n=1}^{N} \delta_n^2$$

Equivalence of error minimization and variance maximization

Task: Show that $\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$ gives the same solution as $\operatorname{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$ when the second L is constrained to pass through $\bar{\mathbf{x}}$

$$\epsilon_n = \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|$$
$$\delta_n = \|\mathrm{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$$
$$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\|$$ Not dependent on L!

Claim 1: The solution L of

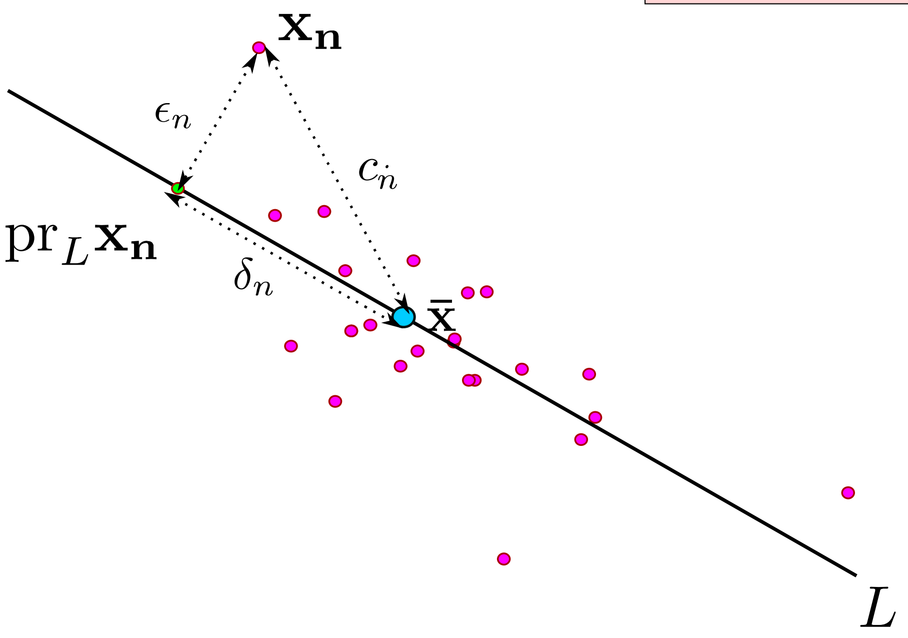
$$\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)

Pythagoras gives $\epsilon_n^2 + \delta_n^2 = c_n^2$ for all $n = 1, \ldots, N$

From that we see:

$$\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_L\mathbf{x}_n\|^2 = \operatorname{argmin}_L \sum_{n=1}^{N} \epsilon_n^2$$
$$= \operatorname{argmin}_L \sum_{n=1}^{N} (c_n^2 - \delta_n^2)$$
$$= \operatorname{argmin}_L - \sum_{n=1}^{N} \delta_n^2$$
$$= \operatorname{argmax}_L \sum_{n=1}^{N} \delta_n^2$$



Equivalence of error minimization and variance maximization

Task: Show that $\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$ gives the same solution as $\operatorname{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x_n} - \bar{\mathbf{x}}\|^2$ when the second L is constrained to pass through $\bar{\mathbf{x}}$

Claim 1: The solution L of

$$\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|^2$$

contains $\bar{\mathbf{x}}$ (we show this in a minute)

Pythagoras gives $\epsilon_n^2 + \delta_n^2 = c_n^2$ for all $n = 1, \ldots, N$

From that we see:

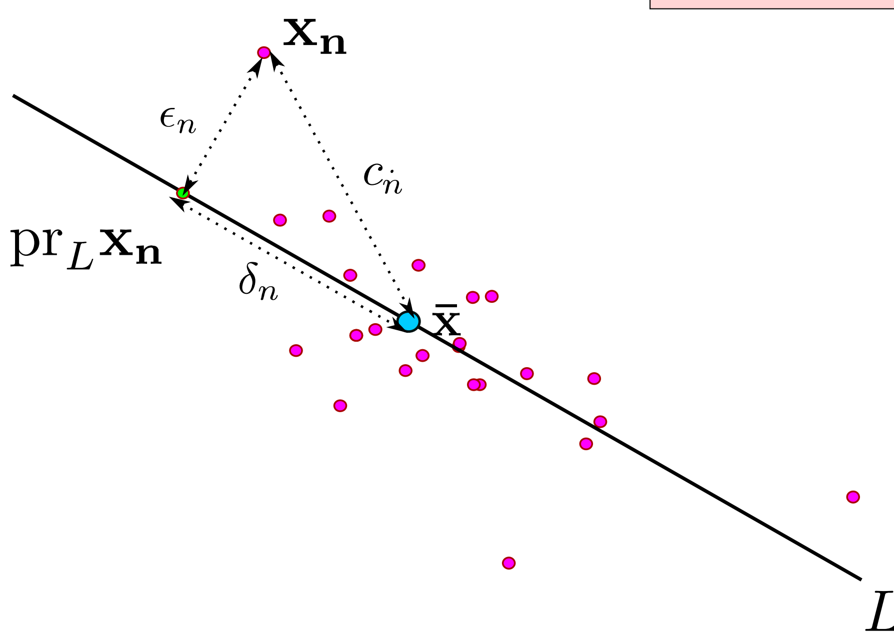
$$\operatorname{argmin}_L \sum_{n=1}^{N} \|\mathbf{x_n} - \mathrm{pr}_L\mathbf{x}_n\|^2 = \operatorname{argmin}_L \sum_{n=1}^{N} \epsilon_n^2$$

$$= \operatorname{argmin}_L \sum_{n=1}^{N} (c_n^2 - \delta_n^2)$$
$$= \operatorname{argmin}_L - \sum_{n=1}^{N} \delta_n^2$$
$$= \operatorname{argmax}_L \sum_{n=1}^{N} \delta_n^2$$
$$= \operatorname{argmax}_L \sum_{n=1}^{N} \|\mathrm{pr}_L\mathbf{x}_n - \bar{\mathbf{x}}\|^\mathbf{2}$$

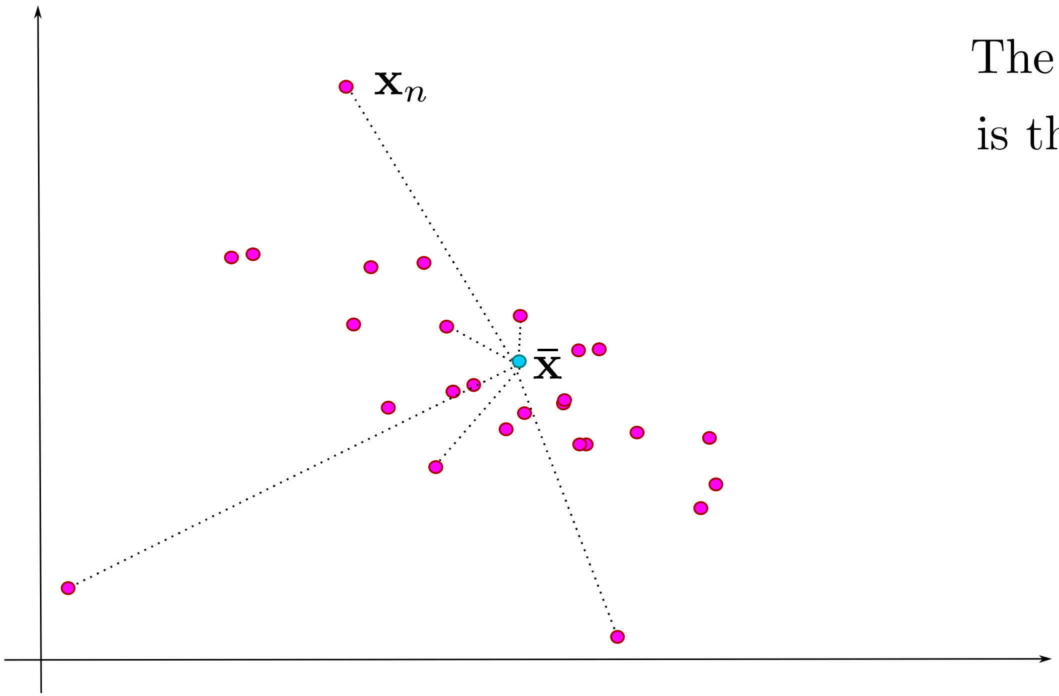
$\epsilon_n = \|\mathbf{x_n} - \mathrm{pr}_\mathbf{L}\mathbf{x_n}\|$

$\delta_n = \|\mathrm{pr}_\mathbf{L}\mathbf{x_n} - \bar{\mathbf{x}}\|$

$c_n = \|\mathbf{x_n} - \bar{\mathbf{x}}\|$ Not dependent on L!



Means and PCA

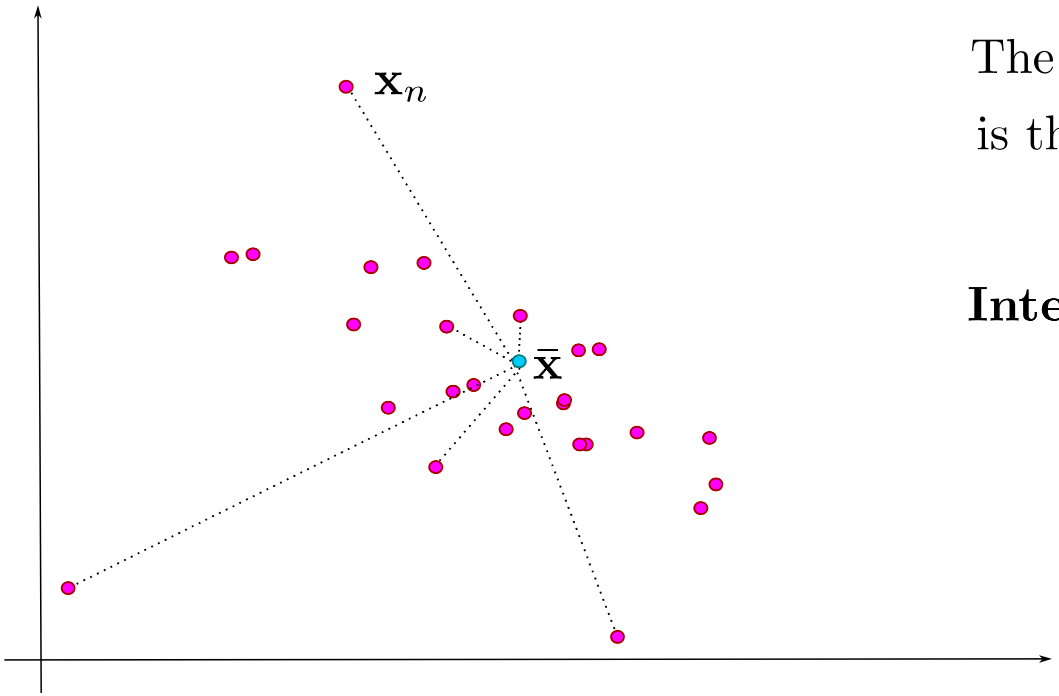


The mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$

is the minimizer of the variance function

$$var(\mathbf{x}) = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{x}\|^2$$

Means and PCA



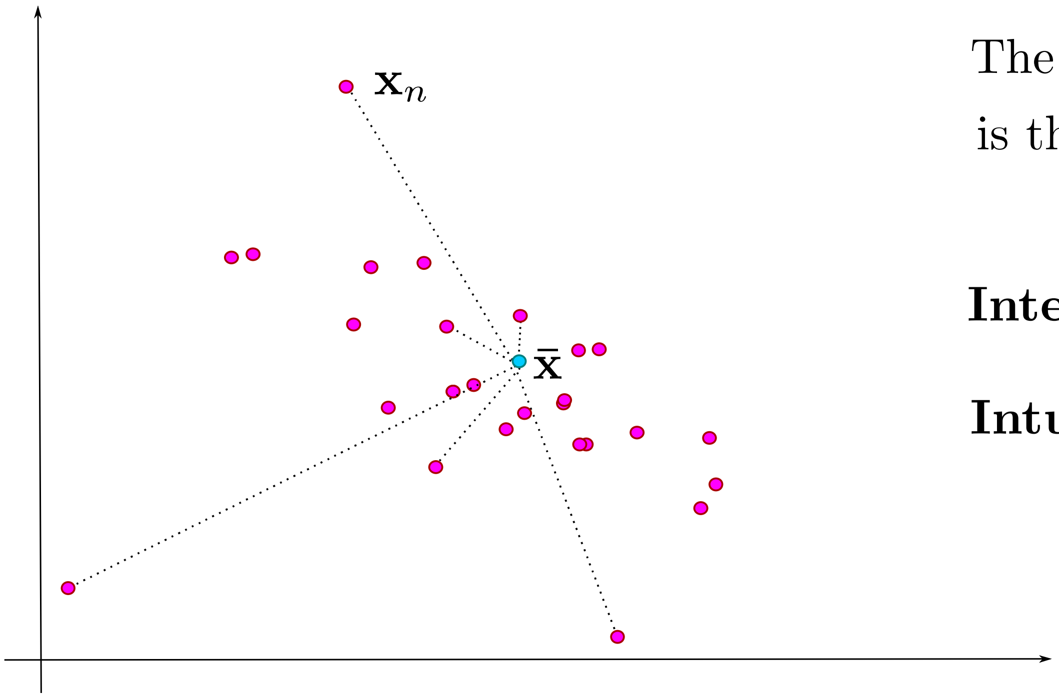
The mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$

is the minimizer of the variance function

$$var(\mathbf{x}) = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{x}\|^2$$

Interpretation: Zeroeth PC

Means and PCA



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

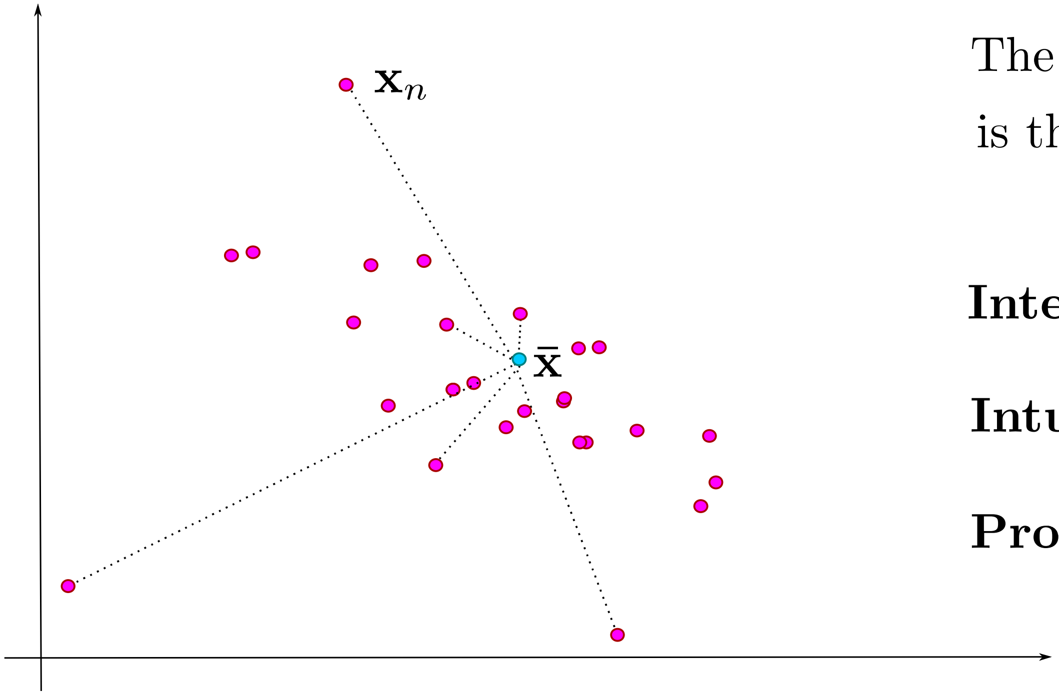
is the minimizer of the variance function

$$var(\mathbf{x}) = \sum_{n=1}^{N}\|\mathbf{x}_n - \mathbf{x}\|^2$$

Interpretation: Zeroeth PC

Intuition: Mean is the most central point

Means and PCA



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

is the minimizer of the variance function

$$var(\mathbf{x}) = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{x}\|^2$$

Interpretation: Zeroeth PC

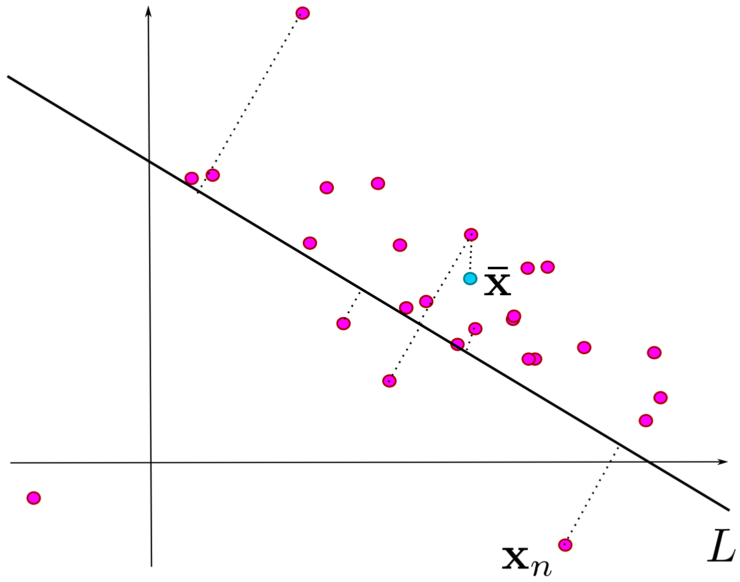
Intuition: Mean is the most central point

Proof: Set derivative of $var(\mathbf{x})$ to 0

Means and PCA

The mean $\bar{\mathbf{x}}$ lies on the principal component / subspace L defined as

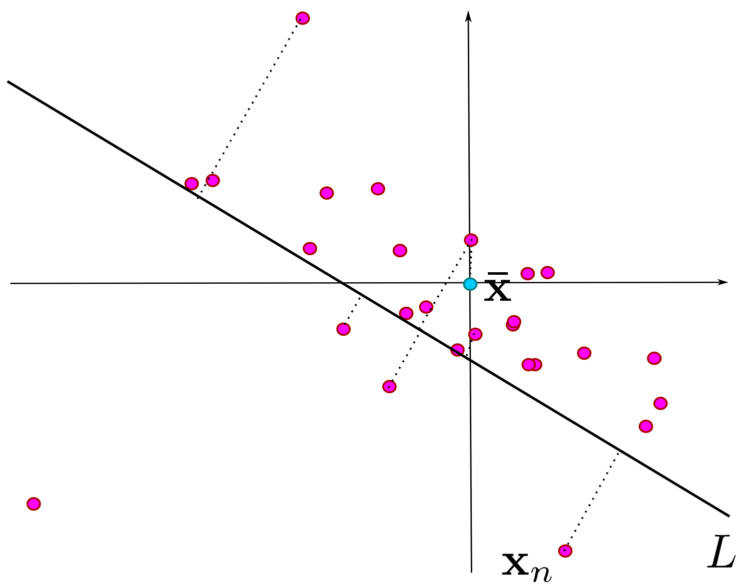
$$\mathbf{argmin}_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{pr}_L \mathbf{x}_n\|^2$$



Means and PCA

The mean $\bar{\mathbf{x}}$ lies on the principal component / subspace L defined as

$$\mathbf{argmin}_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{pr}_L \mathbf{x}_n\|^2$$

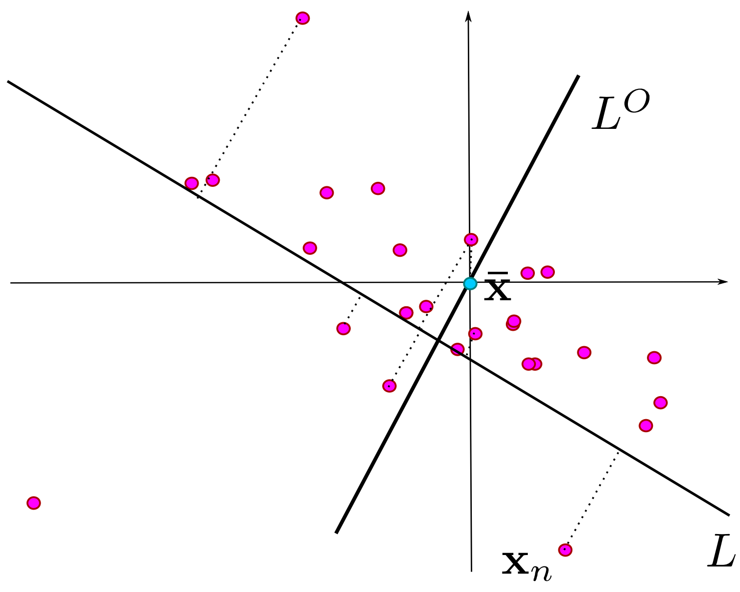


Step 1: Enough to show this for the case $\bar{\mathbf{x}} = \mathbf{0}$

Means and PCA

The mean $\bar{\mathbf{x}}$ lies on the principal component / subspace L defined as

$$\mathbf{argmin}_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{pr}_L \mathbf{x}_n\|^2$$



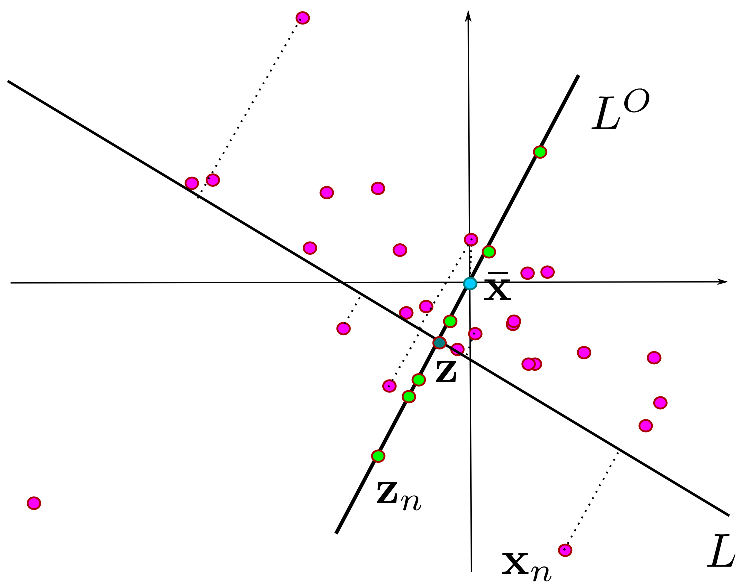
Step 1: Enough to show this for the case $\bar{\mathbf{x}} = \mathbf{0}$

Step 2: Assume L is known up to translation; let L^O be its orthogonal complement through $\mathbf{0}$

Means and PCA

The mean $\bar{\mathbf{x}}$ lies on the principal component / subspace L defined as

$$\mathbf{argmin}_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{pr}_L \mathbf{x}_n\|^2$$



Step 1: Enough to show this for the case $\bar{\mathbf{x}} = \mathbf{0}$

Step 2: Assume L is known up to translation; let L^O be its orthogonal complement through $\mathbf{0}$

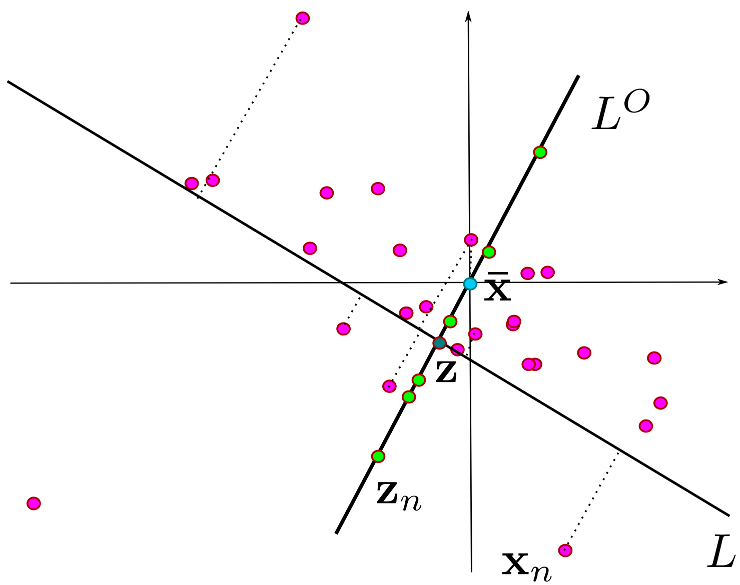
Step 3: If $\mathbf{z} = L \cap L^O$ then

$$\min_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathrm{pr}_L \mathbf{x}_n\|^2 = \min_{\mathbf{z}} \sum_{n=1}^{N} \|\mathbf{z}_n - \mathbf{z}\|^2$$

Means and PCA

The mean $\bar{\mathbf{x}}$ lies on the principal component / subspace L defined as

$$\mathbf{argmin}_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{pr}_L \mathbf{x}_n\|^2$$



Step 1: Enough to show this for the case $\bar{\mathbf{x}} = \mathbf{0}$

Step 2: Assume L is known up to translation; let L^O be its orthogonal complement through $\mathbf{0}$

Step 3: If $\mathbf{z} = L \cap L^O$ then

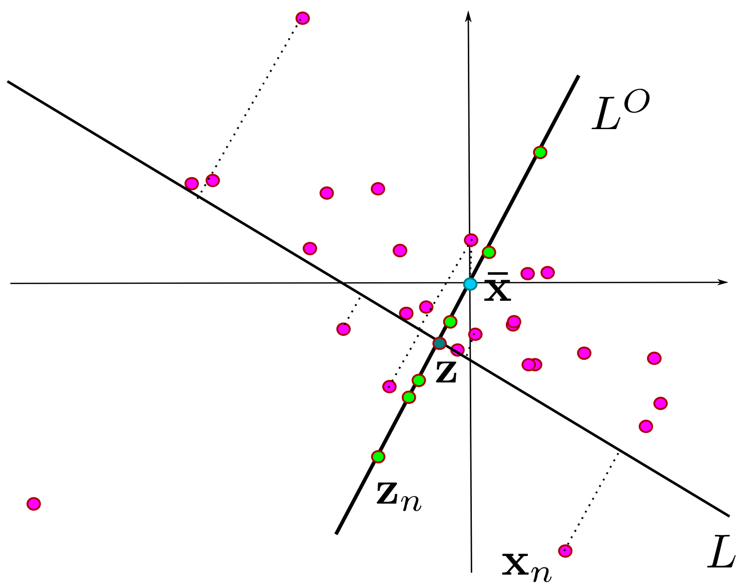
$$\min_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathrm{pr}_L \mathbf{x}_n\|^2 = \min_{\mathbf{z}} \sum_{n=1}^{N} \|\mathbf{z}_n - \mathbf{z}\|^2$$

Step 4: $\mathbf{0} \in L \Leftrightarrow \mathbf{z} = \mathbf{0}$

Means and PCA

The mean $\bar{\mathbf{x}}$ lies on the principal component / subspace L defined as

$$\mathbf{argmin}_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{pr}_L \mathbf{x}_n\|^2$$



Step 1: Enough to show this for the case $\bar{\mathbf{x}} = \mathbf{0}$

Step 2: Assume L is known up to translation; let L^O be its orthogonal complement through $\mathbf{0}$

Step 3: If $\mathbf{z} = L \cap L^O$ then

$$\min_L \sum_{n=1}^{N} \|\mathbf{x}_n - \mathrm{pr}_L \mathbf{x}_n\|^2 = \min_{\mathbf{z}} \sum_{n=1}^{N} \|\mathbf{z}_n - \mathbf{z}\|^2$$

Step 4: $\mathbf{0} \in L \Leftrightarrow \mathbf{z} = \mathbf{0}$

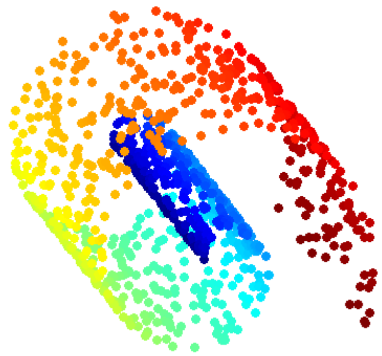
Step 5: $\mathbf{z} = \mathrm{pr}_L \bar{\mathbf{x}} = \mathbf{0}$

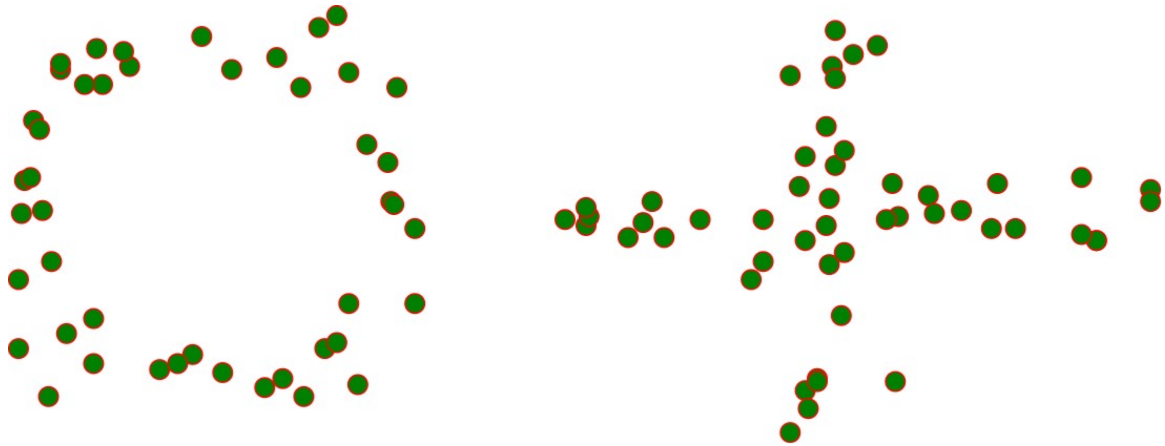
Some cases where PCA fails

- Ideas?

Some cases where PCA fails

- We will work with these on Thursday





http://www.ceremade.dauphine.fr/~peyre/

Summary

- PCA definition:
  - Error minimization = variance maximization = fitting Gaussian
- Applications:
  - Dimensionality reduction
  - Dataset visualization
  - Data variance visualization
  - Preprocessing (whitening = decorrelation)

You should now

- Know the definition of PCA
- Understand why PCA is useful for
  - Dimensionality reduction
  - Data preprocessing
  - Visualization of high dimensional data
- Be able to compute principal components for a given dataset
- Be able to use PCA for visualization of global dataset variation
- Be able to use PCA for interpretation of principal component variation for a certain class of data points including shapes
- Be able to show the equivalence between error minimization and variance maximization definitions of PCA

Until next time:

- Kernel PCA, Multidimensional Scaling, Isomap
- You should read: CB 586-590, 595-599