Principal Component Analysis

Guest Lecture: David Inouye

Instructors: Pradeep Ravikumar, Ziv Bar-Joseph

Machine Learning 10-701

Some Slides Courtesy Barnabas Poczos, Karl Booksh Research group, Tom Mitchell, Ron Parr

Overview

I. Data visualization as motivating example

II. PCA definition and properties
   1. Minimize reconstruction error
   2. Maximize variance of projection

III. PCA algorithms
   1. Sequential
   2. Covariance decomposition
   3. Data matrix decomposition via SVD
   4. Clever reduced decomposition (eigenfaces)

IV. PCA applications
   1. Noise reduction / invariance (eigenfaces)
   2. Data compression (image compression)

V. PCA shortcomings and conclusion
   1. Ignores labels (i.e. unsupervised)
   2. Only captures linear variation

Data Visualization

Example:

- **53** blood measurements (features) from **65** people
- How can we visualize the measurements?

Data Visualization

- Matrix format (65x53)

Instances

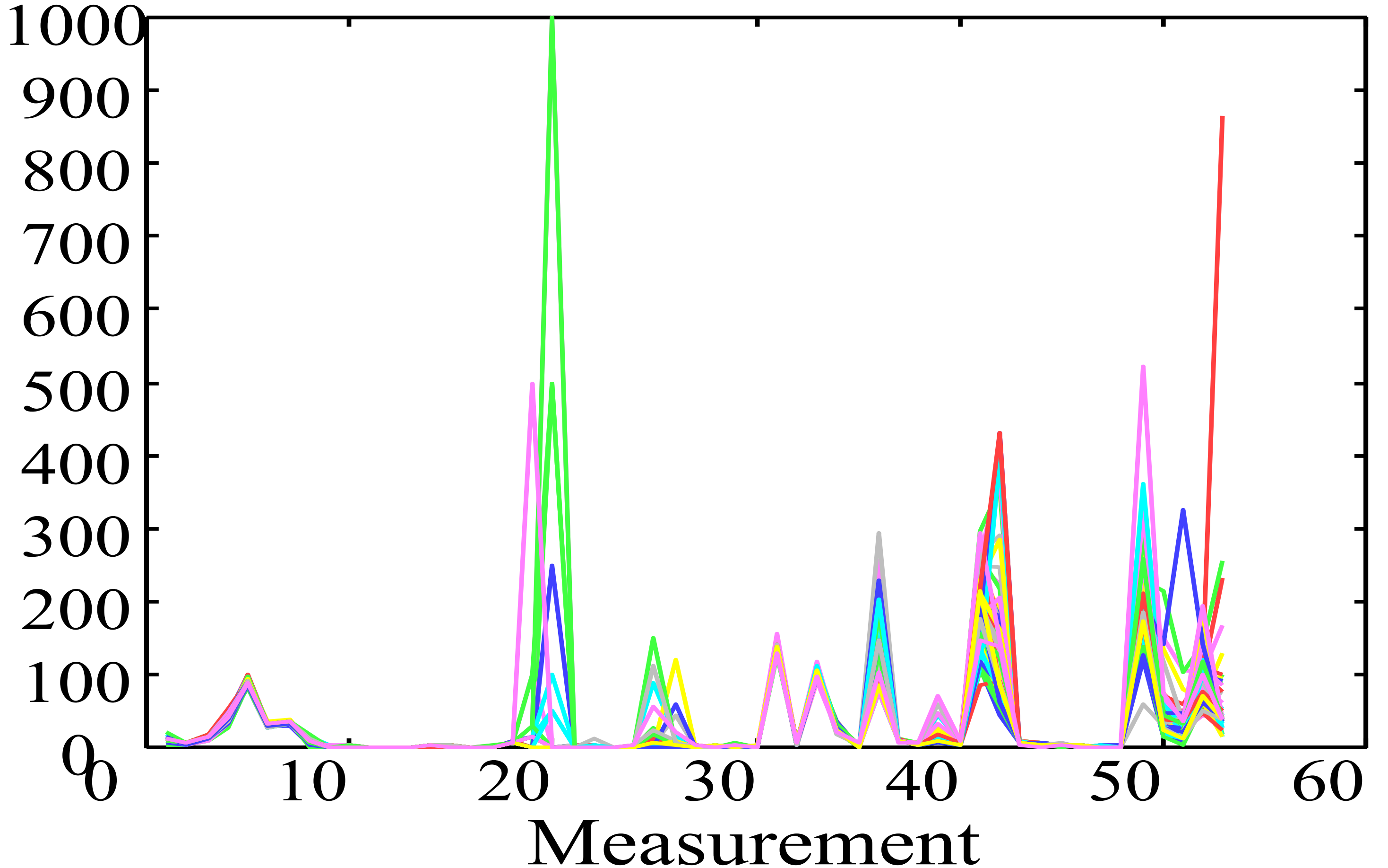
	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000	84.0000	28.0000	33.0000
A6	6.9000	4.8600	16.0000	47.0000	97.0000	33.0000	34.0000
A7	7.8000	4.6800	14.7000	43.0000	92.0000	31.0000	34.0000
A8	8.6000	4.8200	15.8000	42.0000	88.0000	33.0000	37.0000
A9	5.1000	4.7100	14.0000	43.0000	92.0000	30.0000	32.0000

Features

Difficult to see the correlations between the features...

Data Visualization

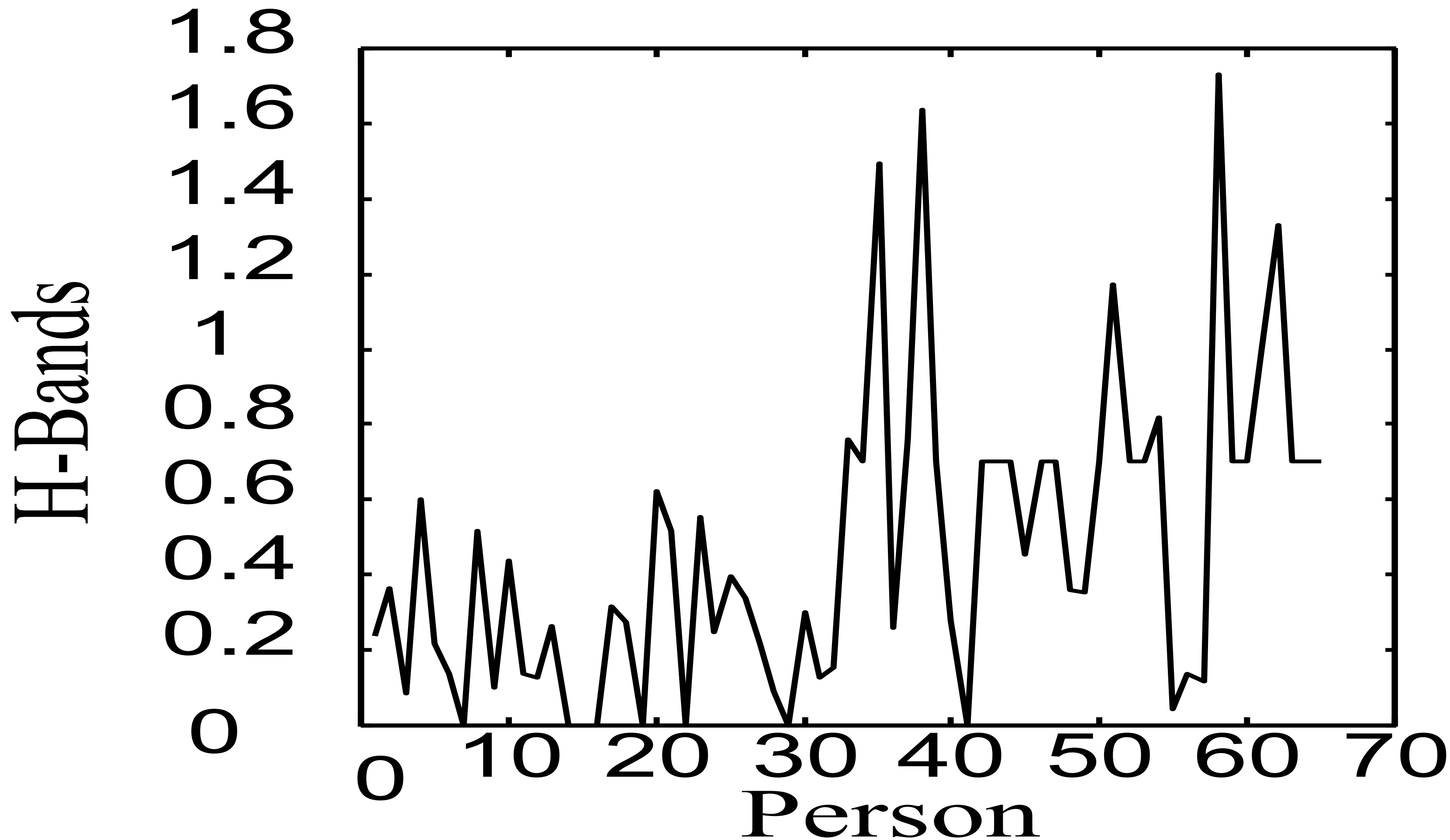
- Curves (65 curves, one for each person)



Difficult to compare the different patients...

Data Visualization

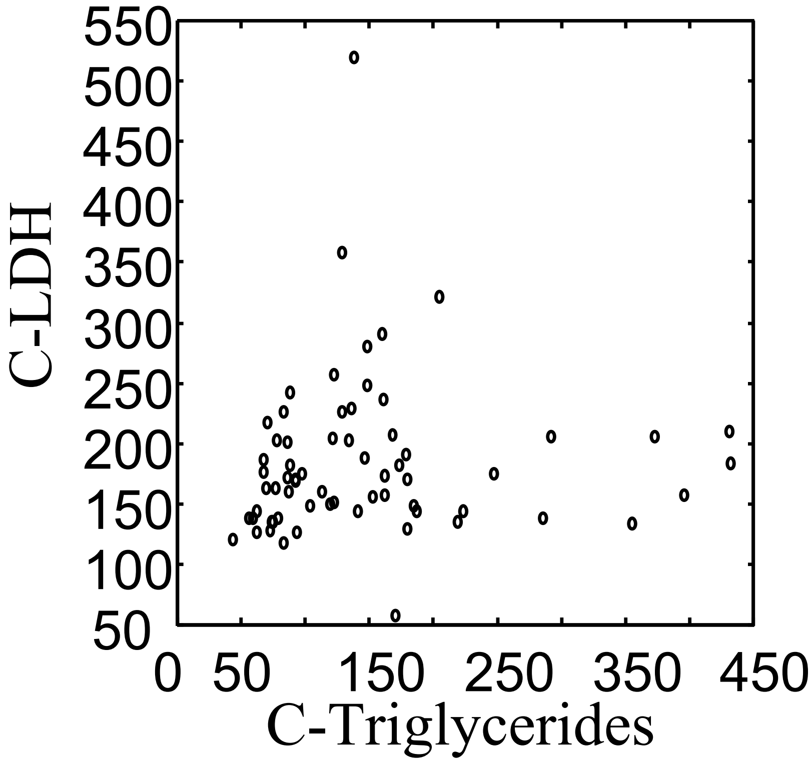
- Curves (53 pictures, one for each feature)



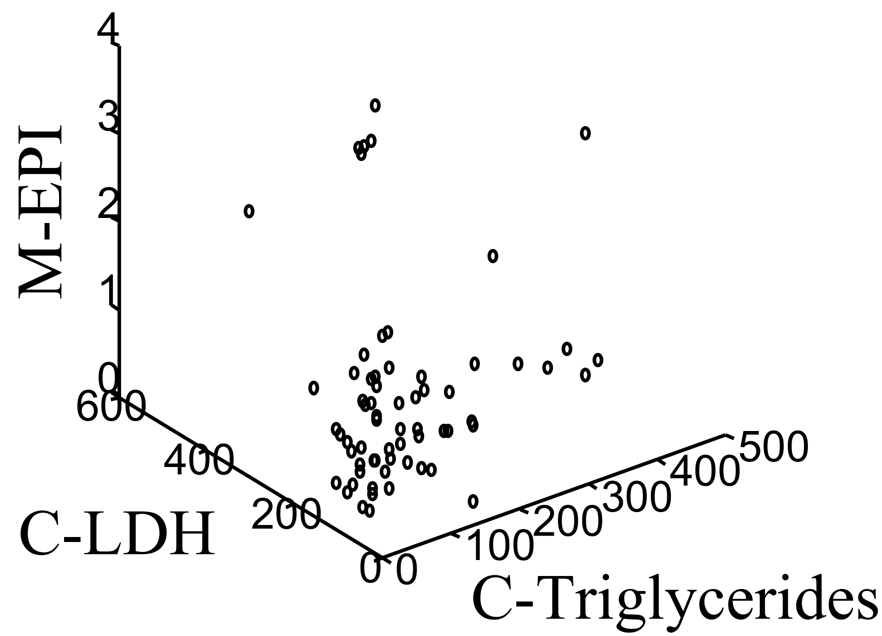
Difficult to see the correlations between the features...

Data Visualization

Bi-variate



Tri-variate



How can we visualize the other variables???

… difficult to see in 4 or higher dimensional spaces...

Data Visualization

- Is there a representation better than the coordinate axes?

- Is it really necessary to show all the 53 dimensions?
  - … what if there are strong correlations between the features?

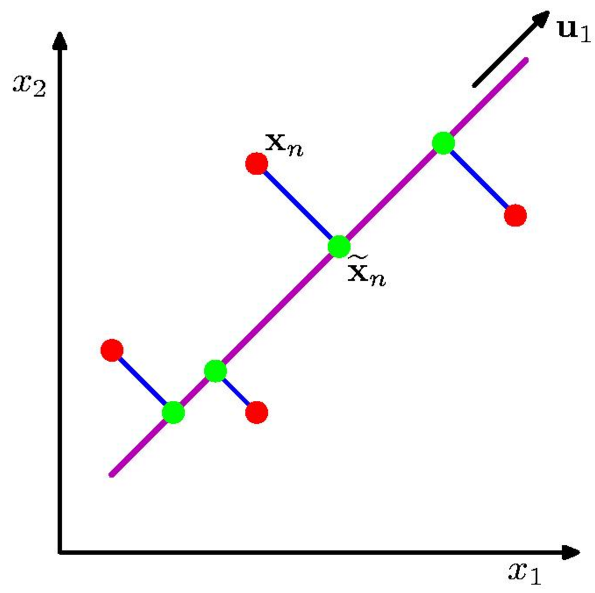
- How could we find
  the *smallest* subspace of the 53-D space that
  keeps the *most information* about the original data?

- A solution: **Principal Component Analysis**

Overview

I. Data visualization as motivating example
II. PCA definition and properties
    1. Minimize reconstruction error
    2. Maximize variance of projection
III. PCA algorithms
    1. Sequential
    2. Covariance decomposition
    3. Data matrix decomposition via SVD
    4. Clever reduced decomposition (eigenfaces)
IV. PCA applications
    1. Noise reduction / invariance (eigenfaces)
    2. Data compression (image compression)
V. PCA shortcomings and conclusion
    1. Ignores labels (i.e. unsupervised)
    2. Only captures linear variation

Principal Component Analysis



PCA:

- Orthogonal projection of the data onto a lower-dimension linear space that <u>equivalently</u>...
  1. *minimizes* the mean squared distance between
     - data points (red points) and projections (green points)
     - i.e. sum of squares of blue line lengths
  2. *maximizes* variance of projected data (green points)

Principal Component Analysis

Idea:

- ❑ Given data points in a N-dimensional space, project them into a lower dimensional space while preserving as much information as possible.
  - Find best planar approximation of 3D data
  - Find best 12-D approximation of 10,000-D data

- ❑ In particular, choose **linear** projection that minimizes *squared error* in reconstructing the original data.

Principal Component Analysis

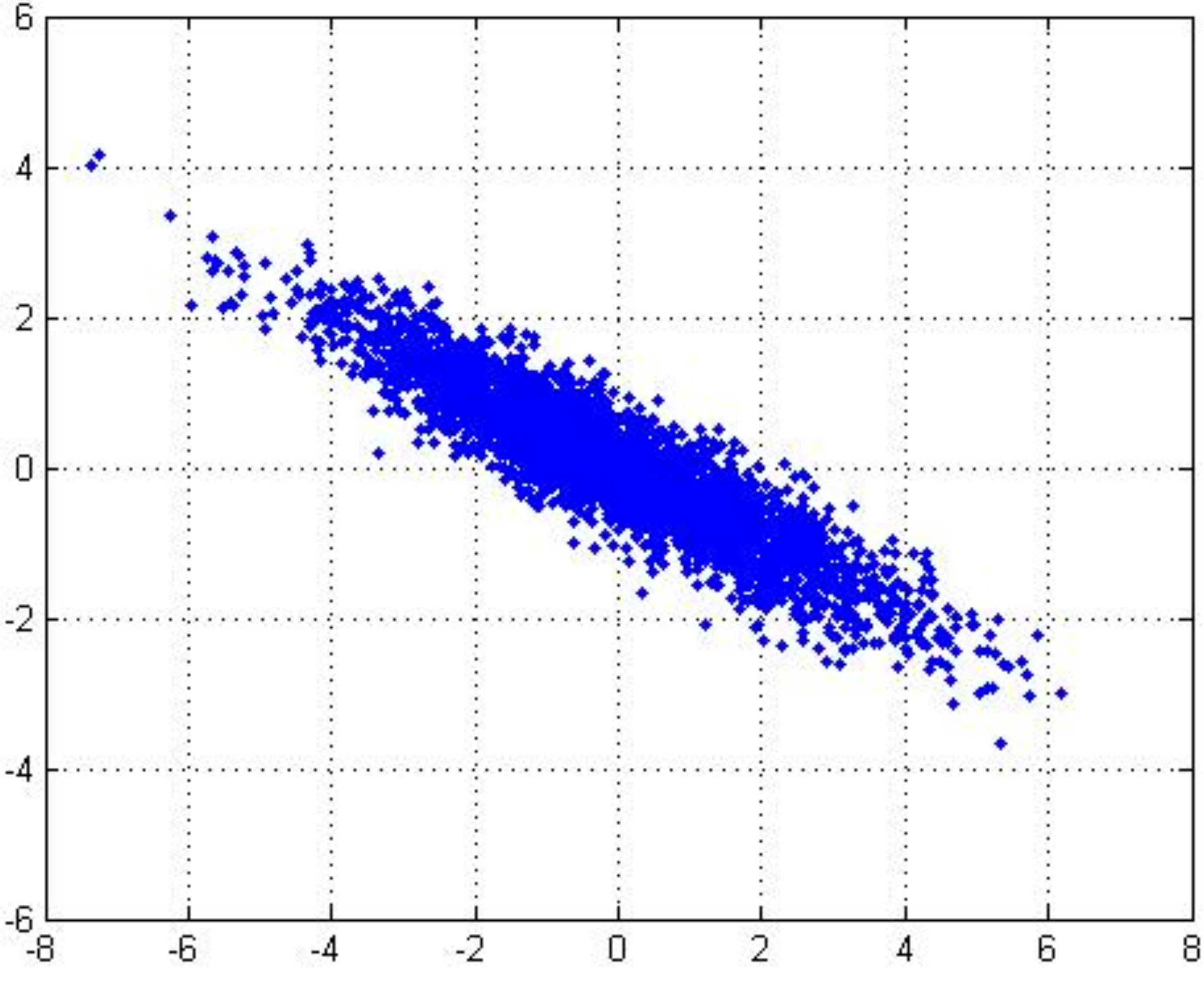
Properties:

- ❑ **PCA vectors** originate from the center of mass.

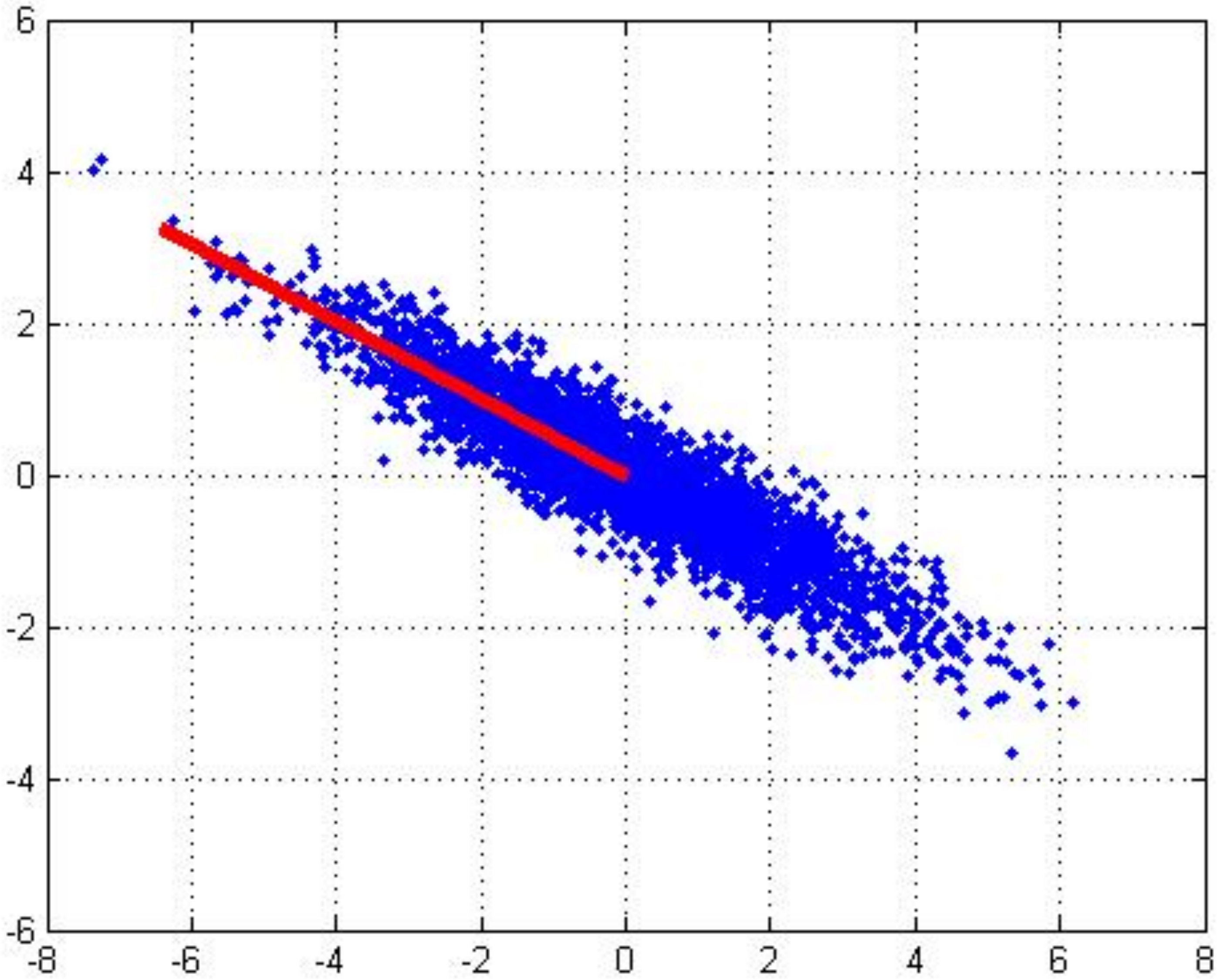
- ❑ Principal component #1: points in the direction of the **largest variance**.

- ❑ Each subsequent principal component
  - is **orthogonal** to the previous ones, and
  - points in the directions of the **largest variance of the residual subspace**

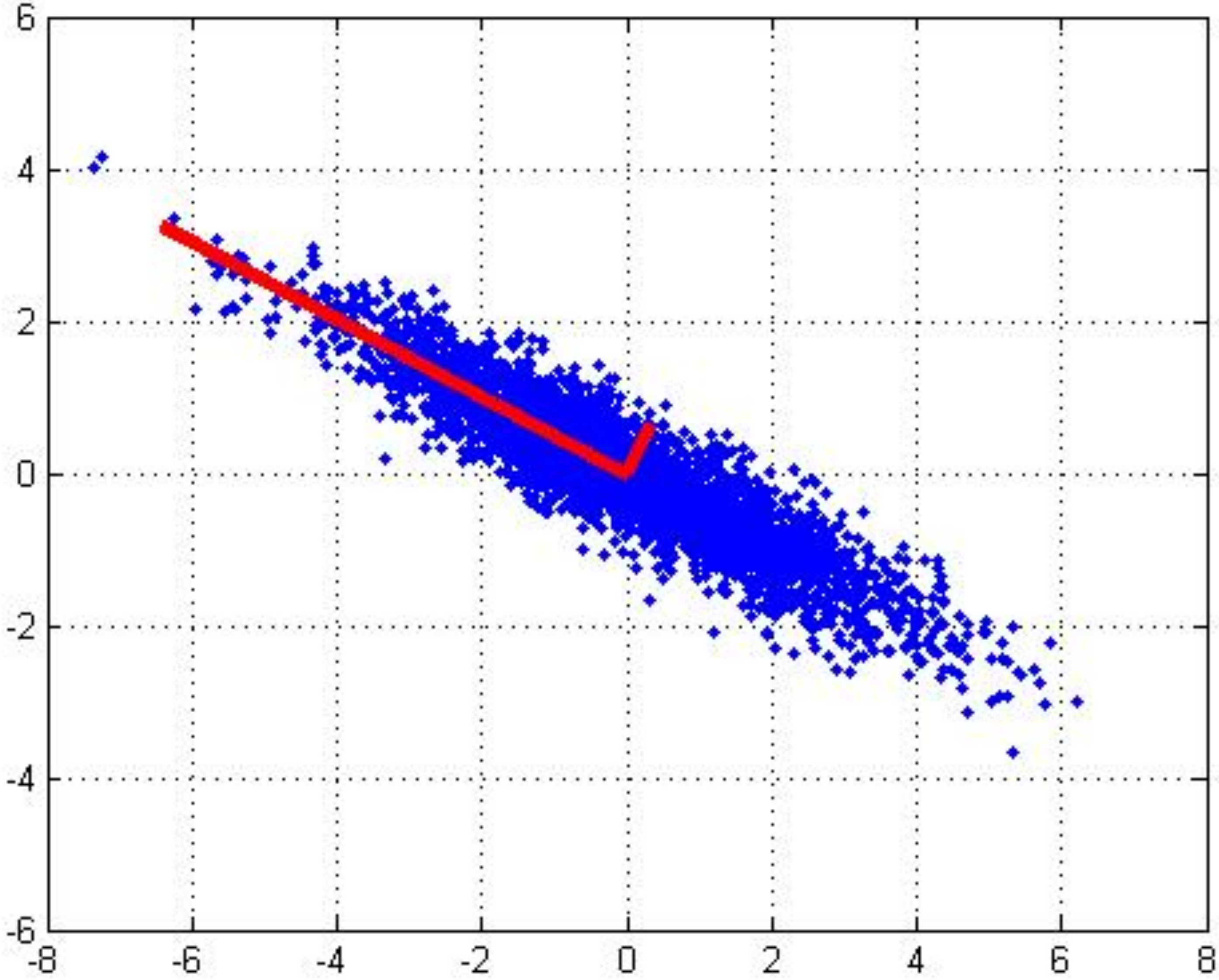
2D Gaussian dataset



1st PCA axis



2^{nd} PCA axis



Overview

I. Data visualization as motivating example
II. PCA definition and properties
   1. Minimize reconstruction error
   2. Maximize variance of projection
III. PCA algorithms
   1. Sequential
   2. Covariance decomposition
   3. Data matrix decomposition via SVD
   4. Clever reduced decomposition (eigenfaces)
IV. PCA applications
   1. Noise reduction / invariance (eigenfaces)
   2. Data compression (image compression)
V. PCA shortcomings and conclusion
   1. Ignores labels (i.e. unsupervised)
   2. Only captures linear variation

PCA algorithm I (sequential)

Given the **centered** data {$\mathbf{x}_1, \ldots, \mathbf{x}_m$}, compute the principal vectors:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m} \sum_{i=1}^{m} \{(\mathbf{w}^T \mathbf{x}_i)^2\}$$ **1st PCA vector**

To find $\mathbf{w_1}$, maximize the variance of projection of $\mathbf{x}$

PCA algorithm I (sequential)

Given the **centered** data $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, compute the principal vectors:

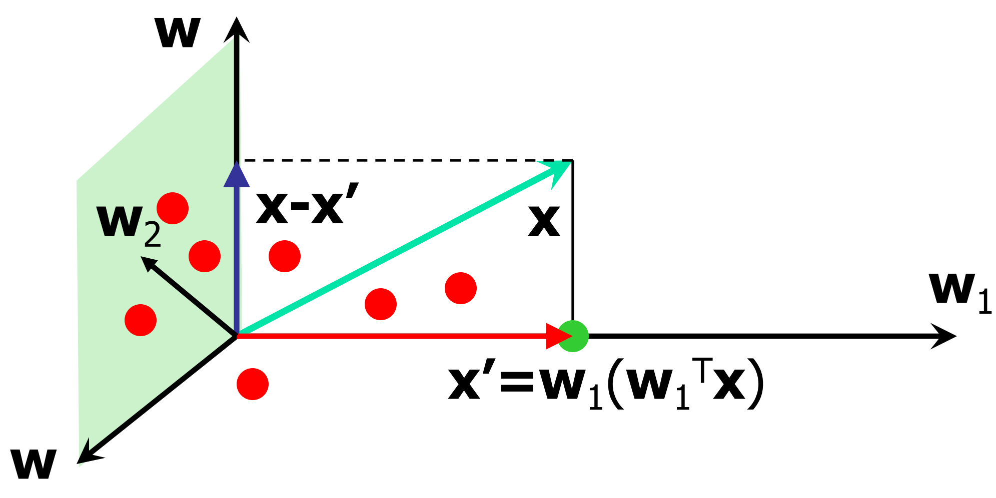
$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m}\sum_{i=1}^{m}\{(\mathbf{w}^T\mathbf{x}_i)^2\}$$ 1st PCA vector

To find $\mathbf{w}_1$, maximize the variance of projection of $\mathbf{x}$

$$\mathbf{w}_2 = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m}\sum_{i=1}^{m}\{[\mathbf{w}^T(\mathbf{x}_i - \underbrace{\mathbf{w}_1\mathbf{w}_1^T\mathbf{x}_i}_{\mathbf{x}'})]^2\}$$ 2nd PCA vector

x′ projection onto w_1

To find $\mathbf{w}_2$, we maximize the **variance** of the projection in the **residual** subspace



PCA algorithm I (sequential)

Given $\mathbf{w_1}, \ldots, \mathbf{w_{k-1}}$, we calculate $\mathbf{w_k}$ principal vector as before:

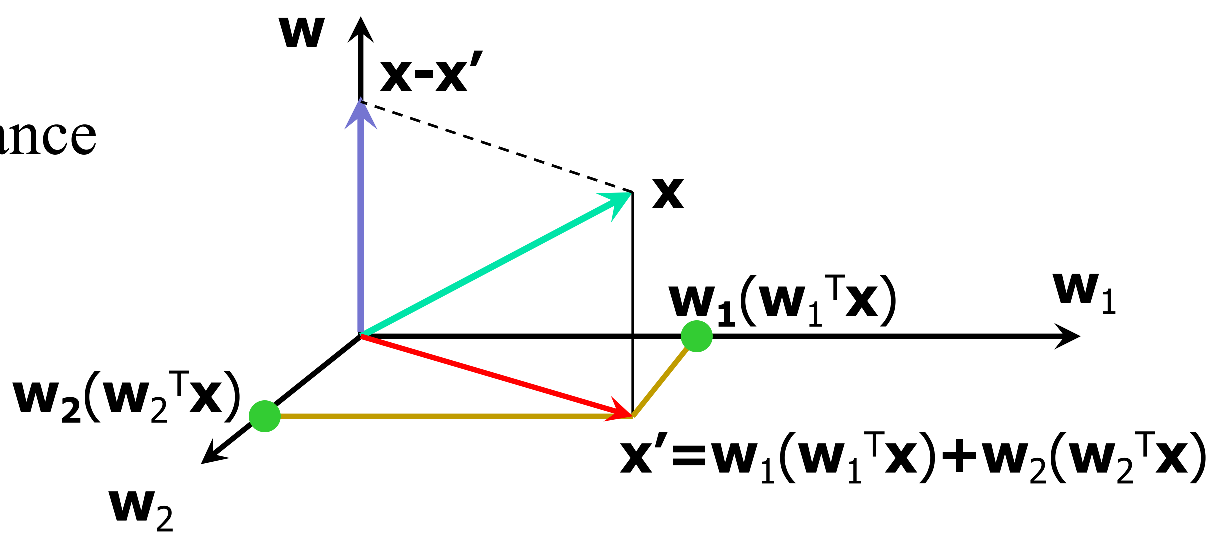
Maximize the variance of projection of $\mathbf{x}$

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m} \sum_{i=1}^{m} \{[\mathbf{w}^T(\mathbf{x}_i - \underbrace{\sum_{j=1}^{k-1} \mathbf{w}_j \mathbf{w}_j^T \mathbf{x}_i}_{\mathbf{x}'})]^2\}$$

k^{th} PCA vector

$\mathbf{x}'$ projection onto previous directions

We maximize the variance of the projection in the residual subspace



PCA algorithm II (sample covariance matrix)

- Given data $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, compute covariance matrix Σ

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T$$

where

$$\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$$

- **PCA** basis vectors = the eigenvectors of Σ

- Larger eigenvalue $\Rightarrow$ more important eigenvectors

PCA algorithm II (sample covariance matrix)

PCA algorithm($\mathbf{X}$, k): top k eigenvalues/eigenvectors

% $\underline{\mathbf{X}}$ = N × m data matrix, N is number of features

% … each data point $\mathbf{x}_i$ = column vector, i=1..m

- $\underline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$
- $\mathbf{X} \leftarrow$ subtract mean $\underline{\mathbf{x}}$ from each column vector $\mathbf{x}_i$ in $\underline{\mathbf{X}}$
- $\Sigma \leftarrow \mathbf{X}\mathbf{X}^T$ … covariance matrix of $\mathbf{X}$
- $\{ \lambda_i, \mathbf{u}_i \}_{i=1..N}$ = eigenvectors/eigenvalues of Σ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$
- Return $\{ \lambda_i, \mathbf{u}_i \}_{i=1..k}$
  % top k PCA components

PCA algorithm III (SVD of the data matrix)

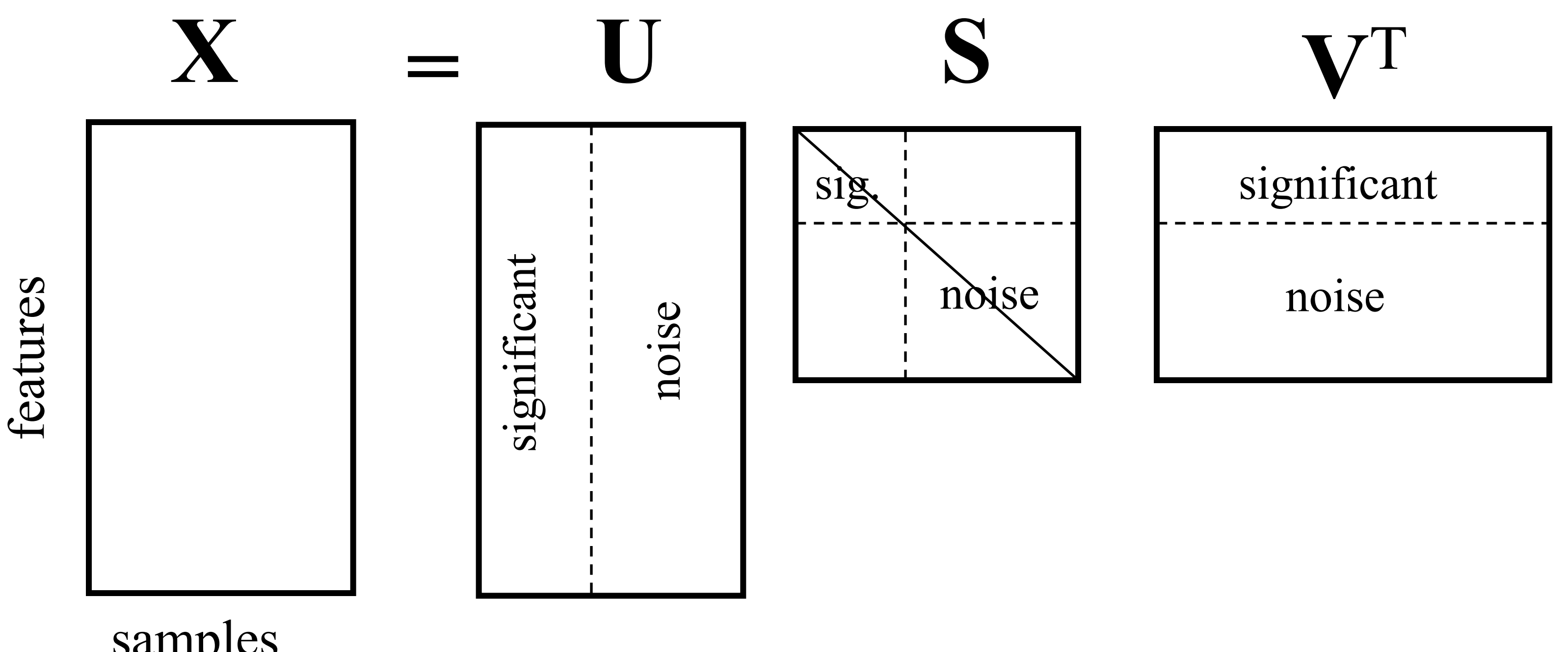
Singular Value Decomposition of the **centered** data matrix **X.**

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{N \times m},$$

m: number of instances,

N: dimension

$$\mathbf{X}_{\text{features} \times \text{samples}} = \mathbf{USV}^{T}$$



PCA algorithm III

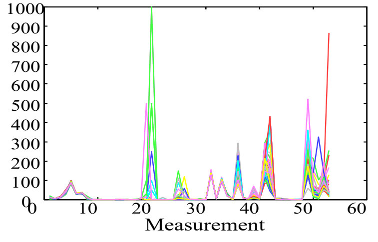
- **Columns of U**
  - The principal vectors, $\{ \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)} \}$
  - Orthogonal and has unit norm – so $U^T U = I$
  - Can reconstruct the data using linear combinations of $\{ \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)} \}$
- **Matrix S of singular values**
  - Diagonal
  - Shows importance of each singular vectors
- **Columns of V^T**
  - The coefficients for reconstructing the samples

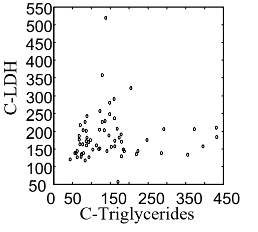
Overview

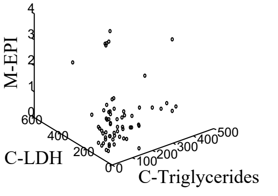
I. Data visualization as motivating example
II. PCA definition and properties
   1. Minimize reconstruction error
   2. Maximize variance of projection
III. PCA algorithms
   1. Sequential
   2. Covariance decomposition
   3. Data matrix decomposition via SVD
   4. Clever reduced decomposition (eigenfaces)
IV. PCA applications
   1. Noise reduction / invariance (eigenfaces)
   2. Data compression (image compression)
V. PCA shortcomings and conclusion
   1. Ignores labels (i.e. unsupervised)
   2. Only captures linear variation

Motivation: PCA Applications

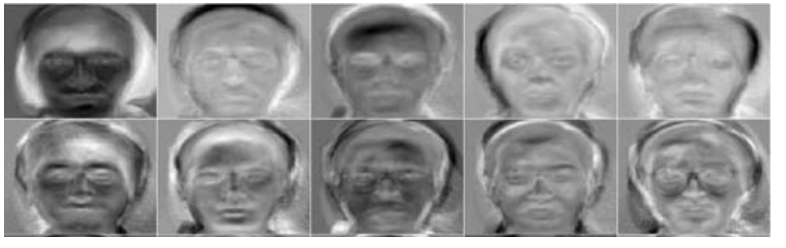
1. Data visualization (blood example)







2. Noise reduction (eigenfaces)



3. Data compression (image example)

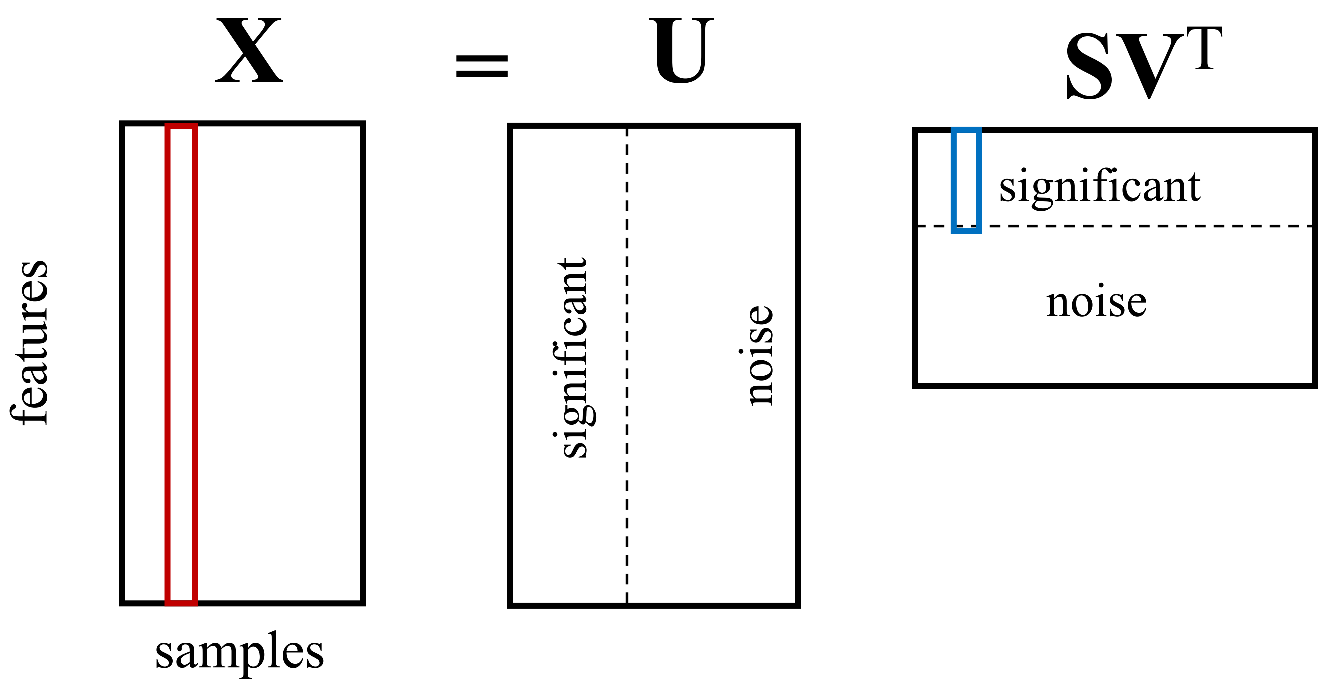
Face Recognition

- ❑ Want to identify specific person, based on facial image
- ❑ Robust to glasses, lighting, facial expression,...
  - ⇒ Can't just use the given 256 x 256 pixels

Applying PCA: Eigenfaces

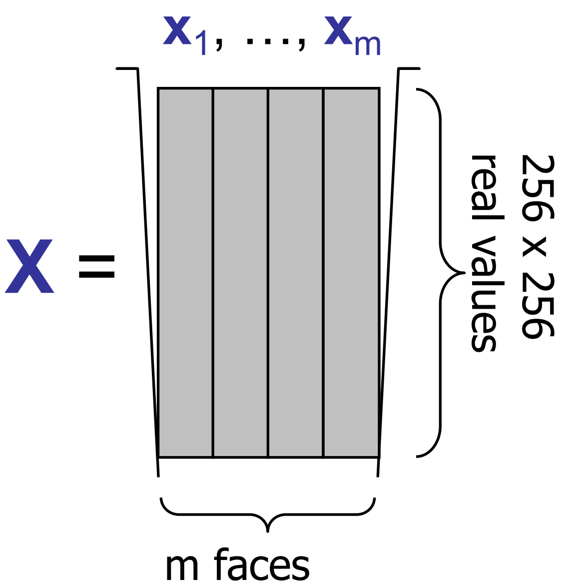
Method: Use PCA on the *whole dataset* to get "principal component" images ("eigenfaces") (U of SVD),

Then classify based on projection weights onto these principal component images (i.e. blue, V^T of SVD)



Applying PCA: Eigenfaces

- Example data set: Images of faces
  - Eigenface approach
    [Turk & Pentland], [Sirovich & Kirby]
- Each face $\mathbf{x}$ is …
  - 256×256 values (luminance at location)
  - $\mathbf{x}$ in $\Re^{256\times256}$ (view as 64K dim vector)
- Form $\mathbf{X} = [\ \mathbf{x}_1, \ldots, \mathbf{x}_m\ ]$ **centered** data matrix
- Compute $\Sigma = \mathbf{X}\mathbf{X}^T$
- Problem: Σ is $64K \times 64K$ … HUGE!!!
  (34 GB in memory)



Computational Complexity

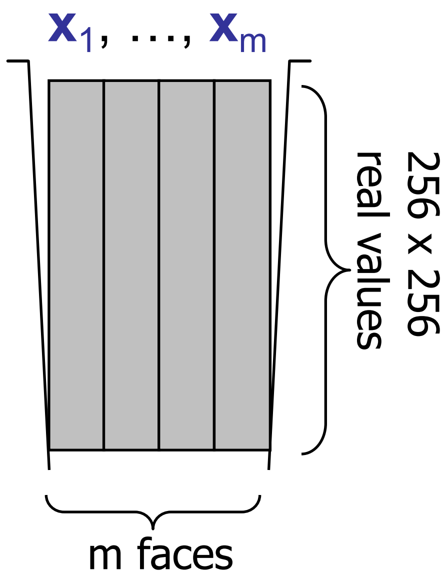
- Suppose m instances, each of size N
  - Eigenfaces: $m=500$ faces, each of size $N=64K$
- Given $N \times N$ covariance matrix Σ, can compute
  - all N eigenvectors/eigenvalues in $O(N^3)$
  - first k eigenvectors/eigenvalues in $O(k\ N^2)$

- But if $N=64K$, EXPENSIVE!

A Clever Workaround

- Note that $m << 64K$
- Use $\mathbf{L = X^T X}$ instead of $\mathbf{\Sigma = X X^T}$
- If $\mathbf{v}$ is eigenvector of $\mathbf{L}$
  then $\mathbf{Xv}$ is eigenvector of Σ
- $O(Nm^2) + O(km^2)$
- 64M vs 42,000M operations

$\mathbf{X} =$



Proof:

$$\mathbf{L\ v} = \gamma\ \mathbf{v}$$

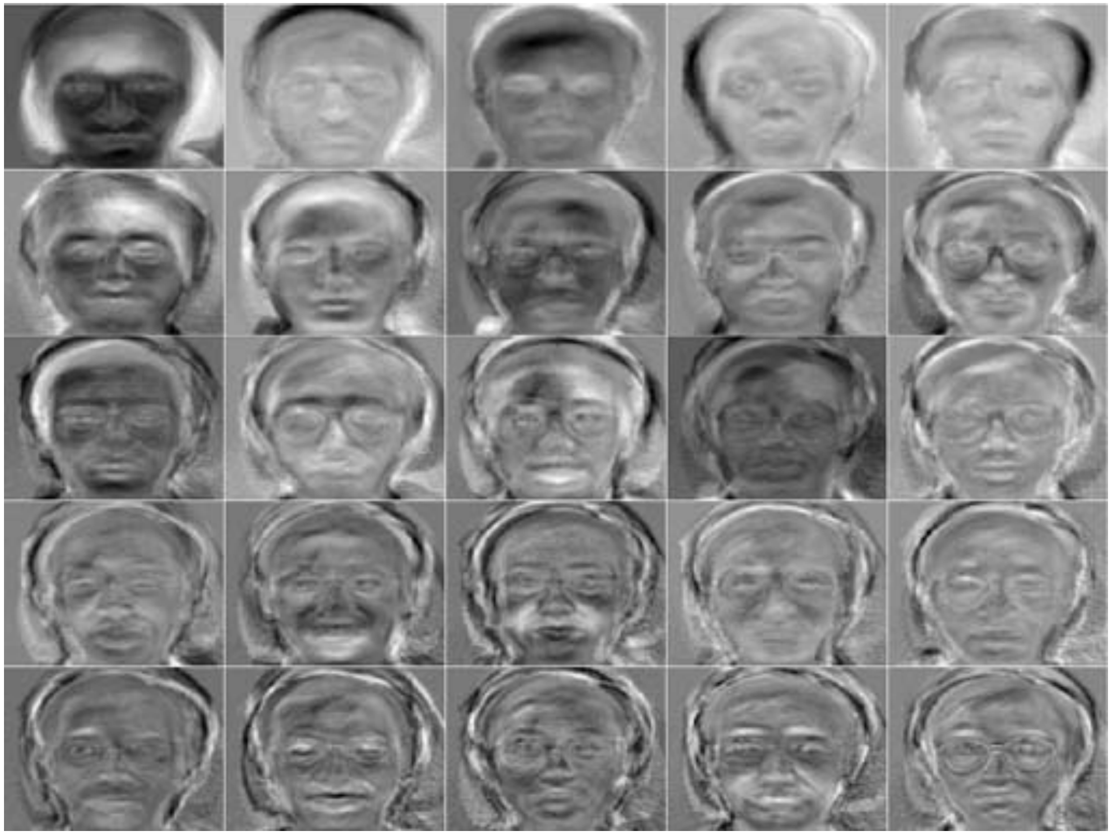
$$\mathbf{X^T X\ v} = \gamma\ \mathbf{v}$$

$$\mathbf{X\ (X^T X\ v)} = \mathbf{X}(\gamma\ \mathbf{v}) = \gamma\ \mathbf{Xv}$$

$$\mathbf{(X X^T) X\ v} = \gamma\ (\mathbf{Xv})$$

$$\Sigma\ (\mathbf{Xv}) = \gamma\ (\mathbf{Xv})$$

Principal Components



Reconstructing…

- Reconstructing only using the top principal components enables a facial representation without finer details ("noise" in this context) such as lighting, glasses and facial expression.

Shortcomings

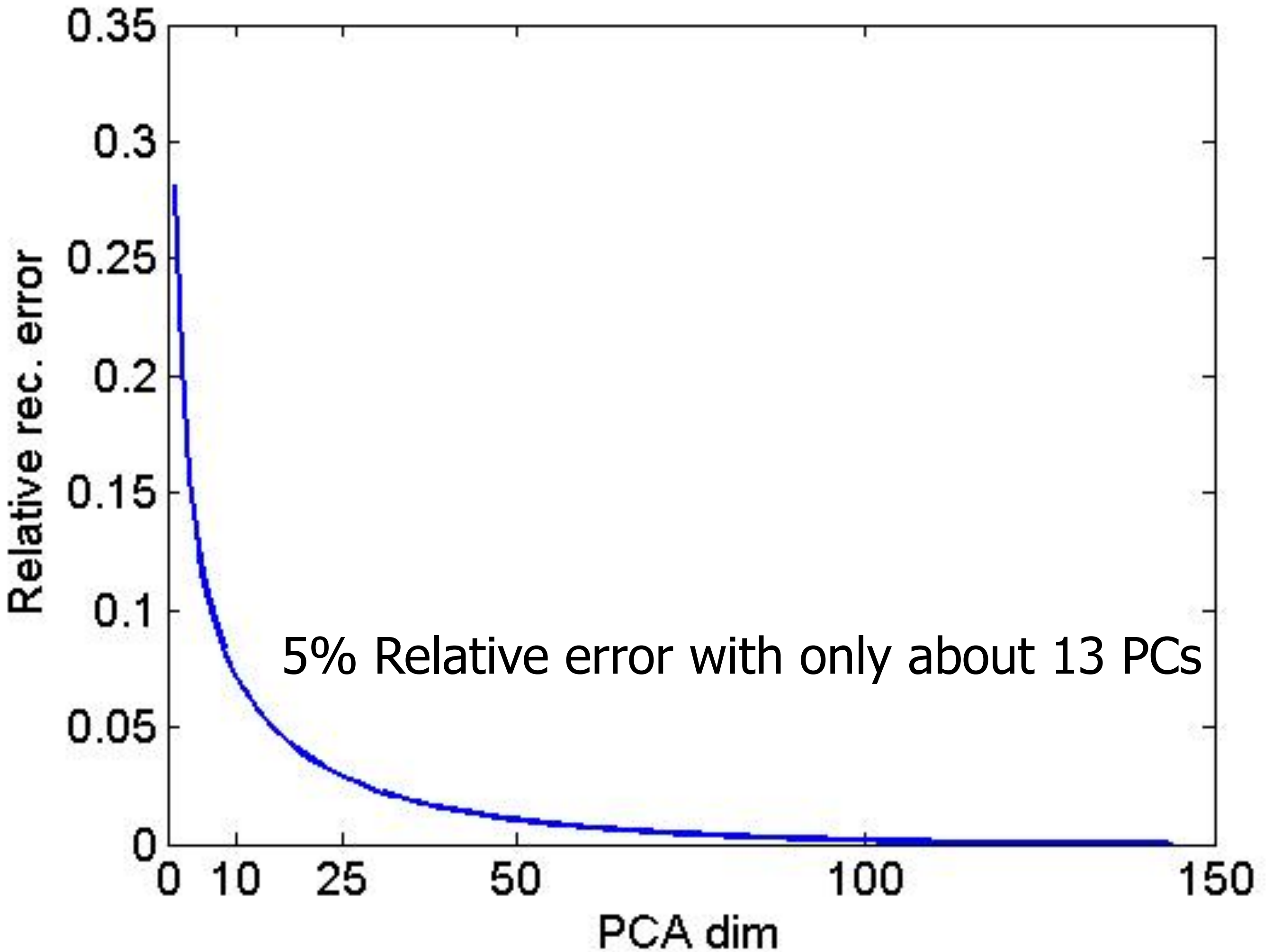
- Requires carefully controlled data:
  - All faces centered in frame
  - Same size
  - Some sensitivity to angle

- Method is completely knowledge free
  - (sometimes this is good!)
  - Doesn't know that faces are wrapped around 3D objects (heads)
  - Makes no effort to preserve class distinctions

Image Compression

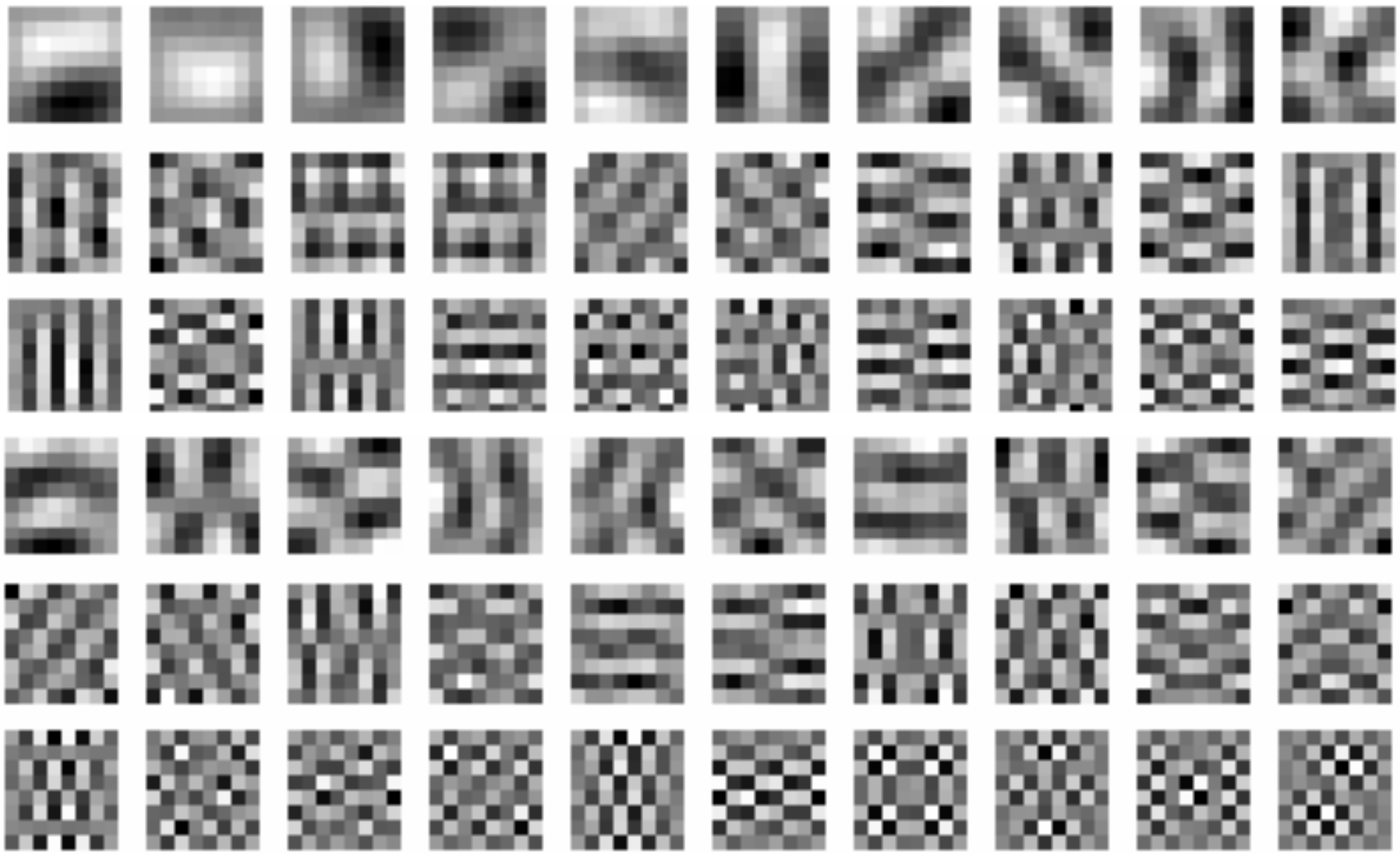
- ❑ Divide the original 372x492 image into patches:
  - Each patch is an instance that contains 12x12 pixels on a grid
- ❑ Consider each as a 144-D vector

L_2 Reconstruction Error



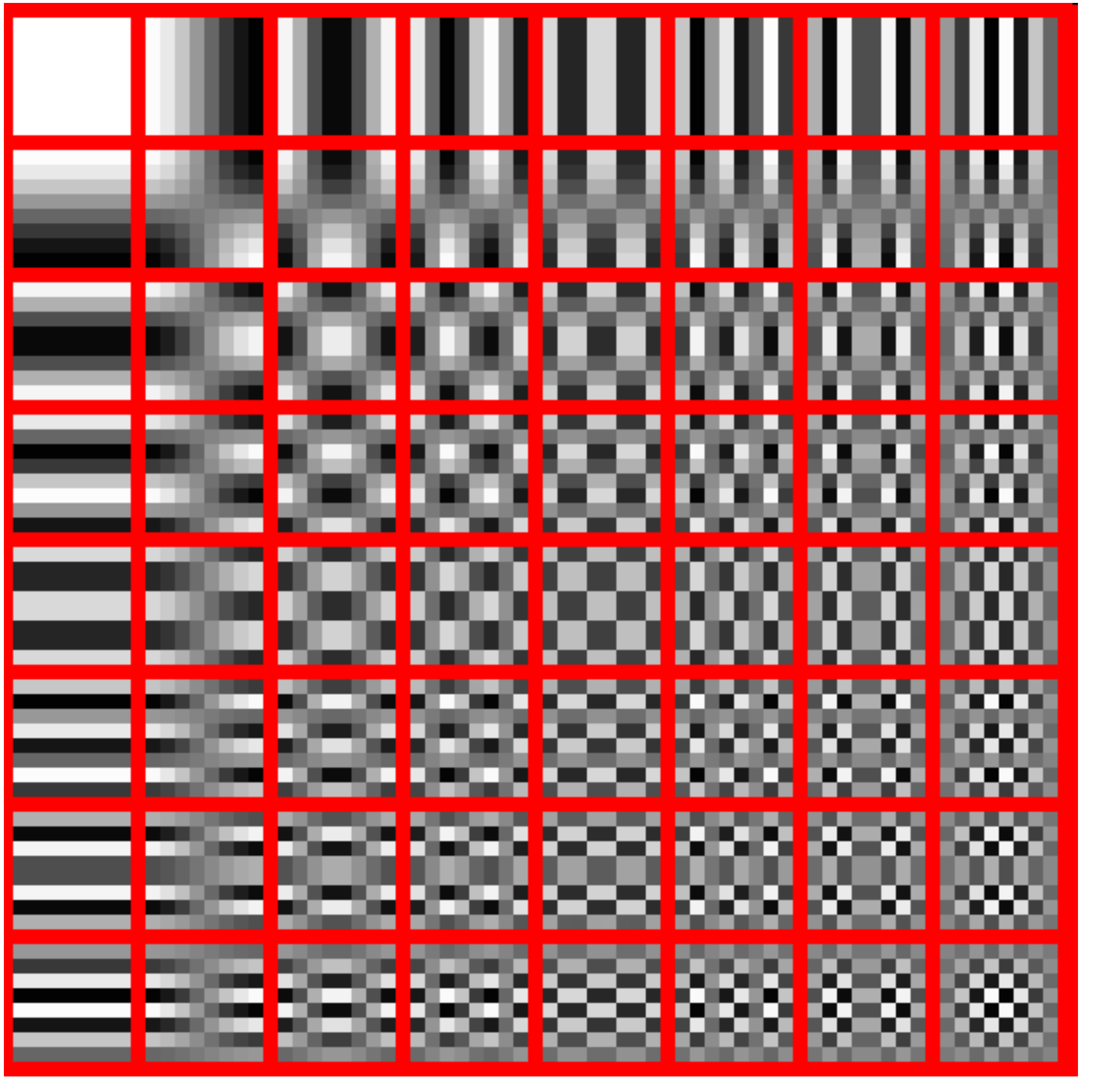
Most information is in the first PCA vectors…

60 most important eigenvectors



Looks like the discrete cosine bases of JPG!...

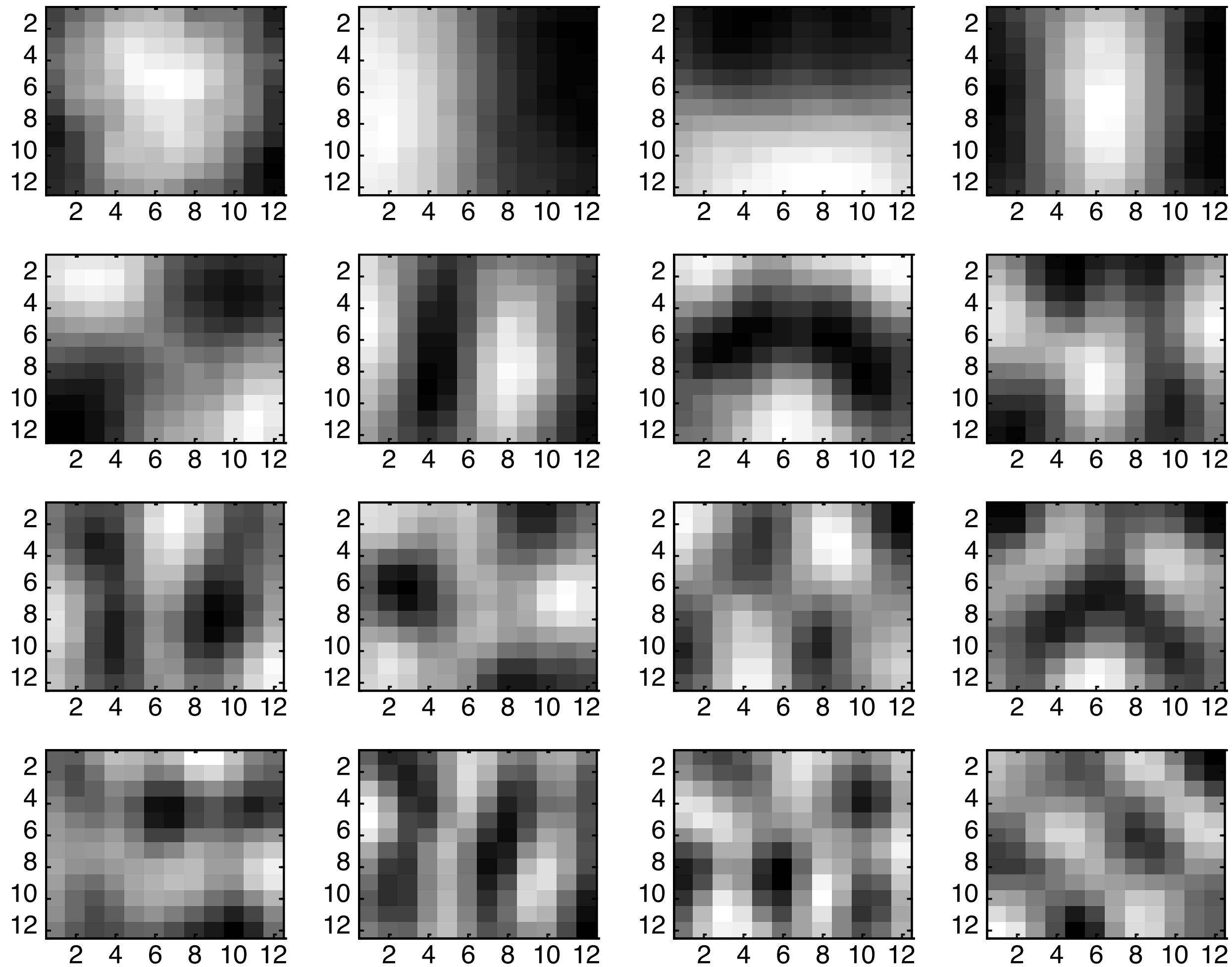
2D Discrete Cosine Basis



http://en.wikipedia.org/wiki/Discrete_cosine_transform

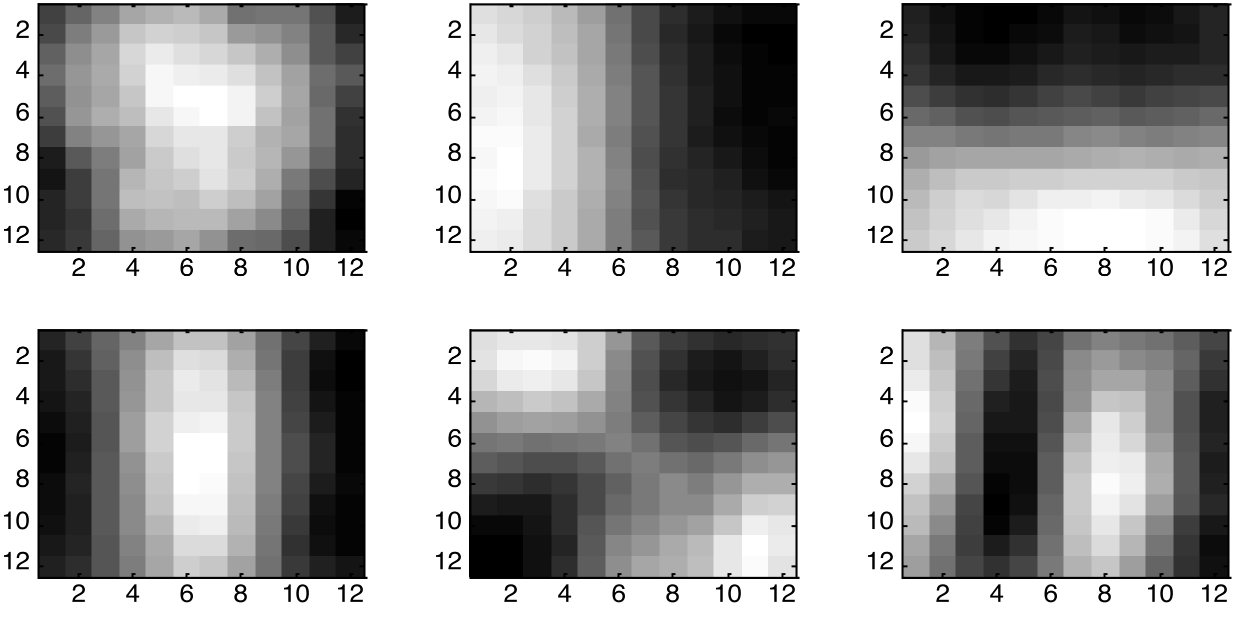
PCA compression: 144D → 60D

16 most important eigenvectors



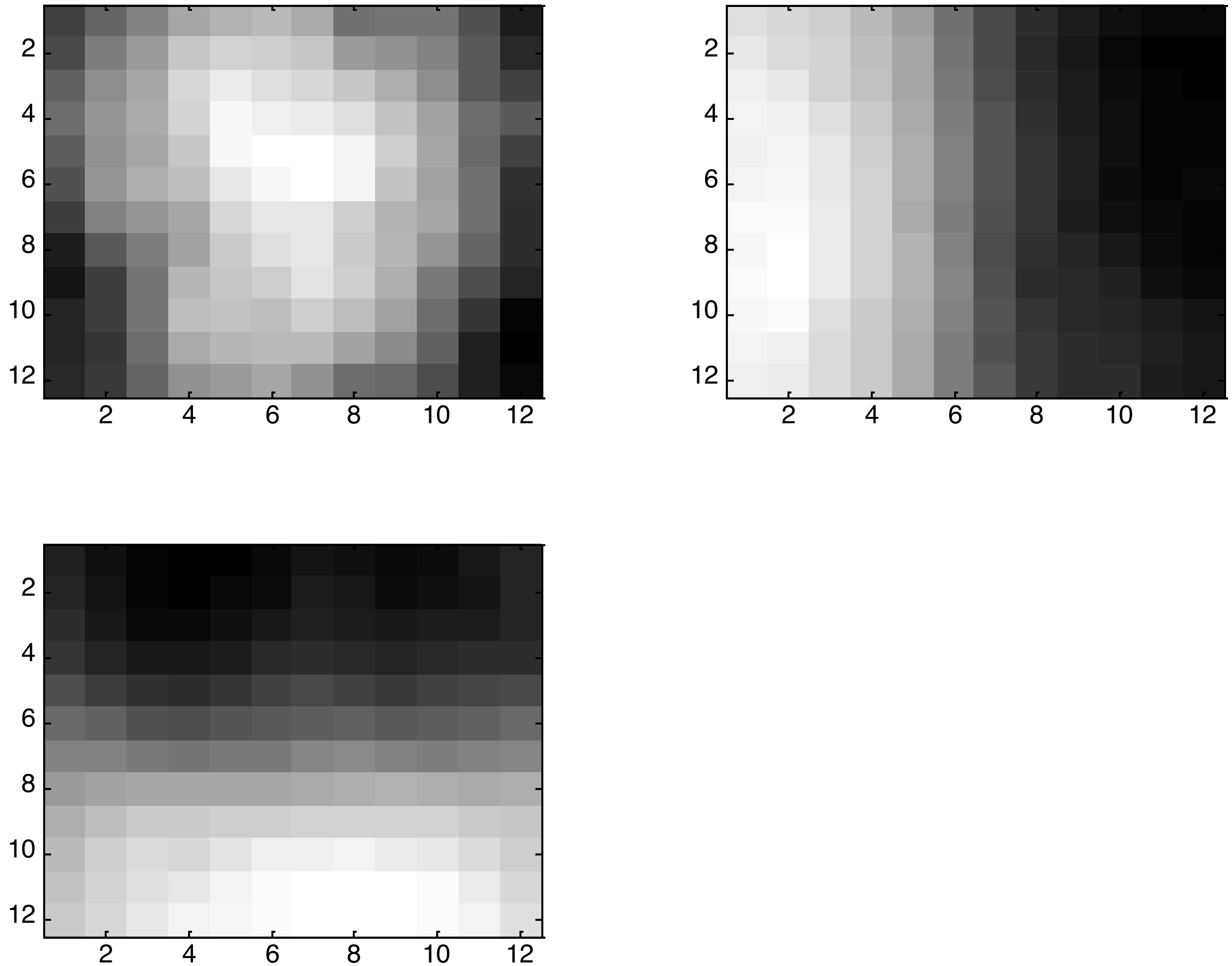
PCA compression: 144D → 16D

6 most important eigenvectors



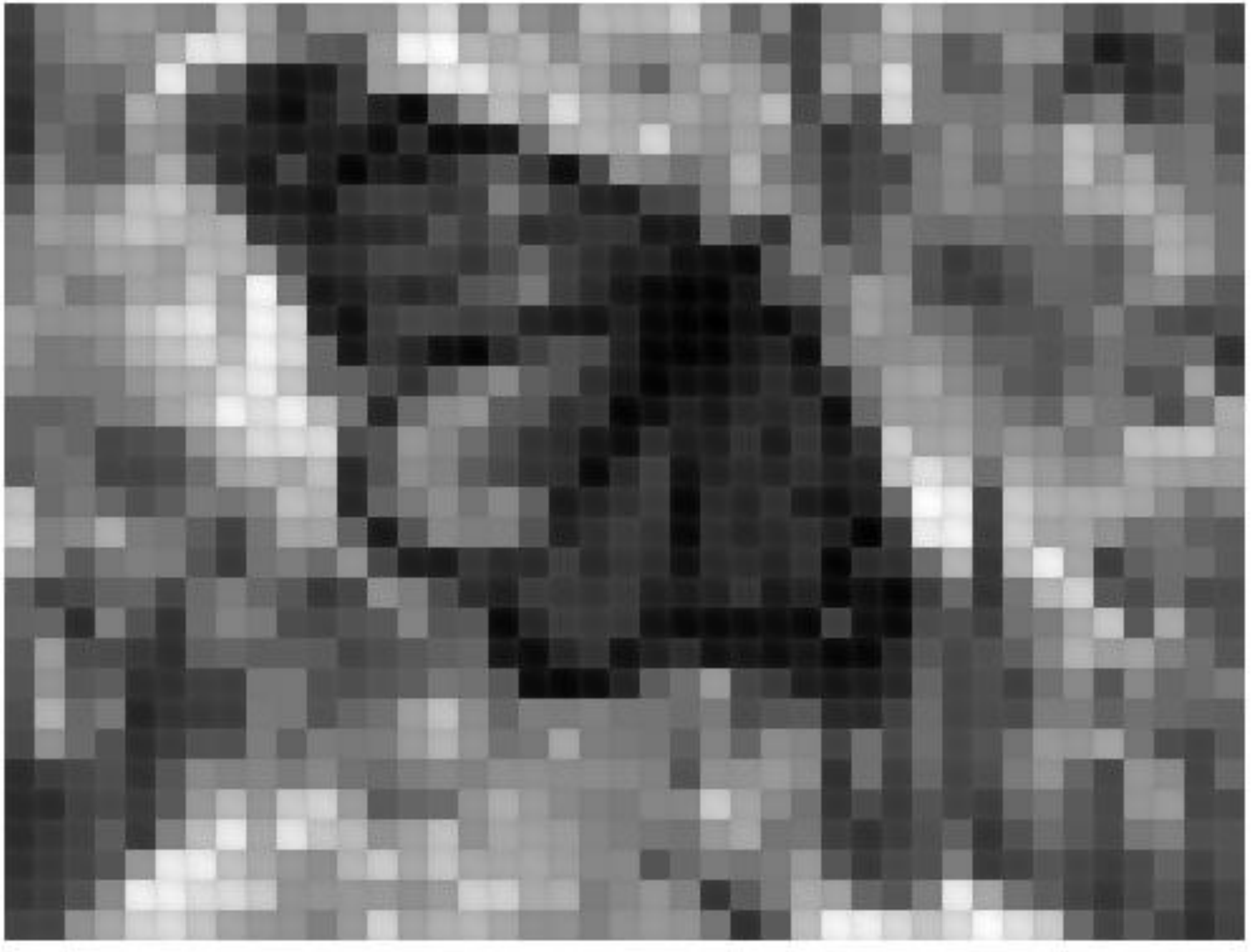
PCA compression: 144D → 6D

3 most important eigenvectors



PCA compression: 144D → 3D

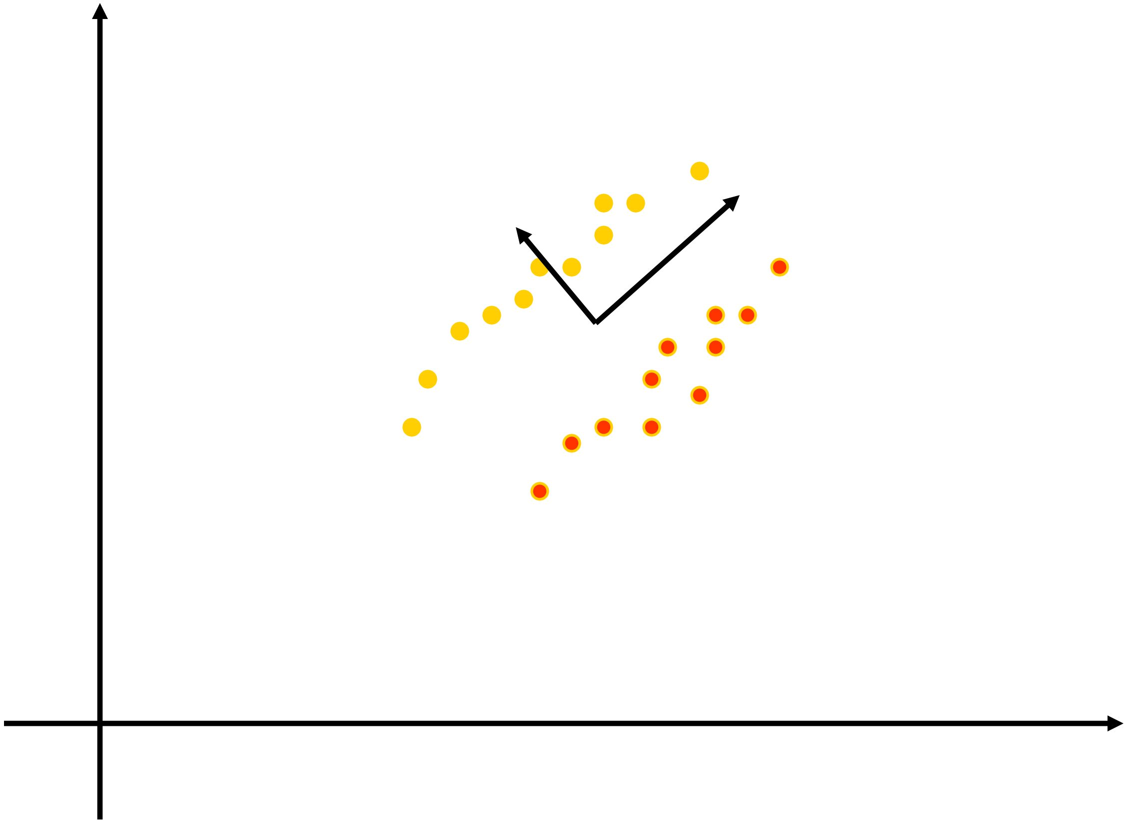
PCA compression: 144D → 1D



Overview

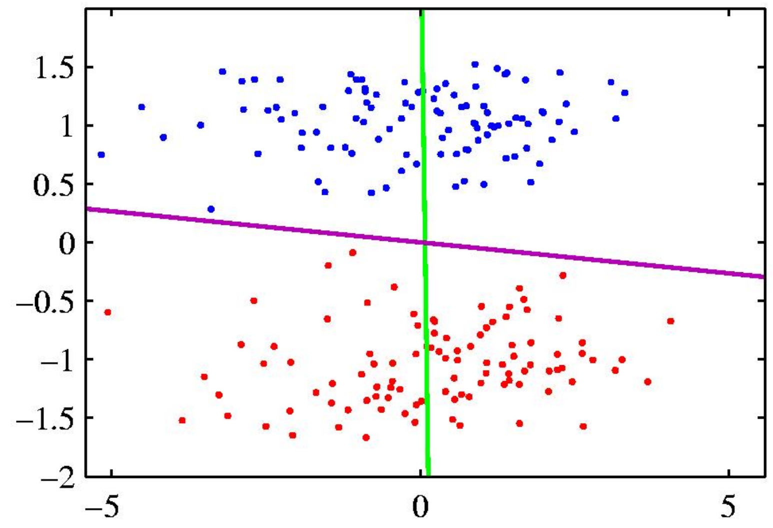
I. Data visualization as motivating example
II. PCA definition and properties
    1. Minimize reconstruction error
    2. Maximize variance of projection
III. PCA algorithms
    1. Sequential
    2. Covariance decomposition
    3. Data matrix decomposition via SVD
    4. Clever reduced decomposition (eigenfaces)
IV. PCA applications
    1. Noise reduction / invariance (eigenfaces)
    2. Data compression (image compression)
V. PCA shortcomings and conclusion
    1. Ignores labels (i.e. unsupervised)
    2. Only captures linear variation

Problematic Data Set for PCA



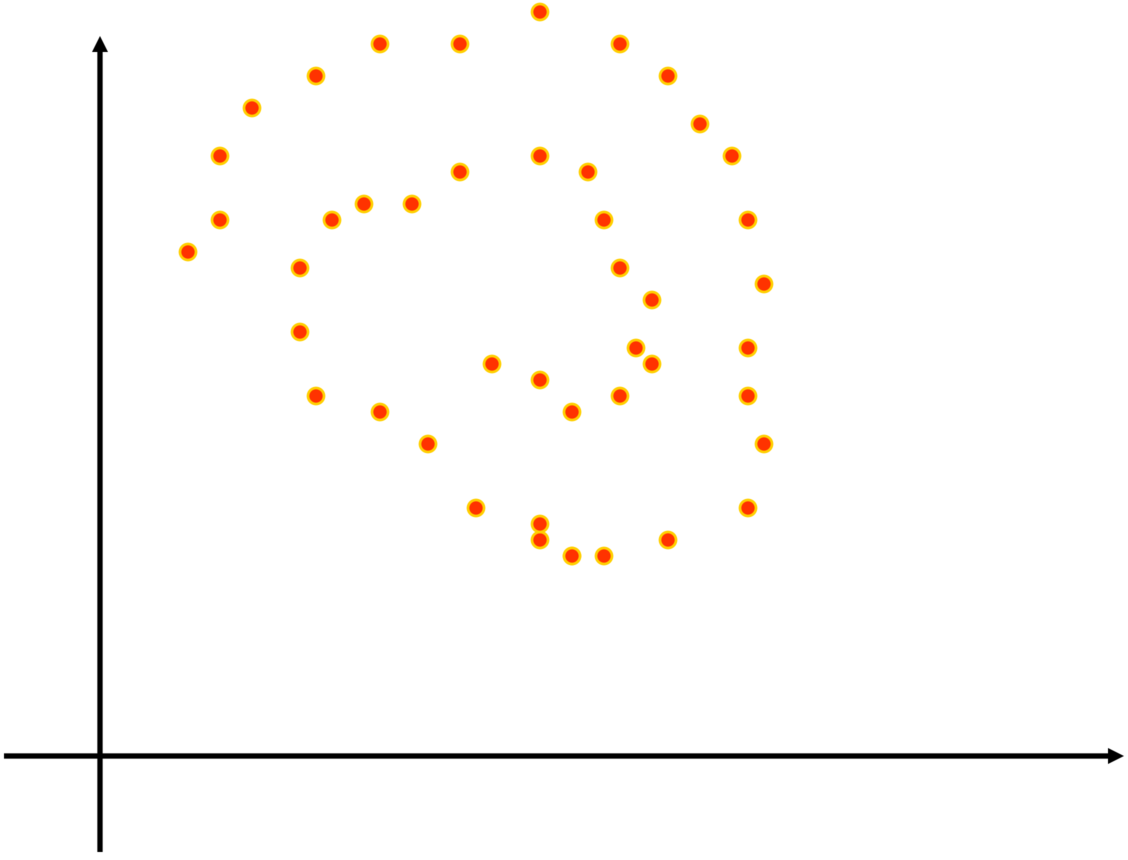
PCA doesn’t know labels!

PCA with Classes



- PCA maximizes variance, *independent of class*
  - $\Rightarrow$ magenta
- If we would want to separate classes
  - $\Rightarrow$ green line

Problematic Data Set for PCA



PCA cannot capture NON-LINEAR structure!

PCA Conclusions

- PCA
  - Finds orthonormal basis for data
  - Sorts dimensions in order of "importance"
  - Usually discard unimportant dimensions
- Applications:
  - Visualization
  - Data compression / compact representation
  - Remove noise to improve classification (hopefully)
- Not magic:
  - Doesn't know class labels
  - Can only capture **linear** variations
- One of many tricks to **reduce dimensionality**!

Kernel PCA

Kernel PCA

Performing PCA in the feature space

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{N \times m}$,

m: number of instances, N: dimension

Lemma

$\mathbf{u}$ is eigenvector of $\Sigma \Rightarrow \mathbf{u}$ is a linear combinaton of the samples

Proof:

$$\lambda \mathbf{u} = \Sigma \mathbf{u} = \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{u} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^T \mathbf{u}) \mathbf{x}_i$$

$$\Rightarrow \mathbf{u} = \sum_{i=1}^{m} \underbrace{\frac{(\mathbf{x}_i^T \mathbf{u})}{\lambda m}}_{\alpha_i} \mathbf{x}_i = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i$$

Kernel PCA

$$\mathbf{u} = \sum_{i=1}^{m} \underbrace{\frac{(\mathbf{x}_i^T \mathbf{u})}{\lambda m}}_{\alpha_i} \mathbf{x}_i = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i \quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{N \times m},$$

Lemma

To calculate $\boldsymbol{\alpha} \in \mathbb{R}^m$

- just use inner products (Gram matrix): $K_{ij} = \mathbf{x}_i^T \mathbf{x}_j$
- don't need the actual values of $\mathbf{x}_i$

Kernel PCA

Proof

$$\mathbf{\Sigma u} = \lambda \mathbf{u}, \;\; \mathbf{u} = \sum_{j=1}^{m} \alpha_j \mathbf{x}_j$$

$$\Rightarrow \mathbf{x}_i^T \mathbf{\Sigma u} = \lambda \mathbf{x}_i^T \mathbf{u}$$

$$\Rightarrow \mathbf{x}_i^T \left(\frac{1}{m} \sum_{k=1}^{m} \mathbf{x}_k \mathbf{x}_k^T \right) \left(\sum_{j=1}^{m} \alpha_j \mathbf{x}_j \right) = \lambda \mathbf{x}_i^T \left(\sum_{j=1}^{m} \alpha_j \mathbf{x}_j \right)$$

$$\Rightarrow \frac{1}{m} \sum_{k=1}^{m} \sum_{j=1}^{m} (\mathbf{x}_i^T \mathbf{x}_k)(\mathbf{x}_k^T \mathbf{x}_j) \alpha_j = \lambda \sum_{j=1}^{m} (\mathbf{x}_i^T \mathbf{x}_j) \alpha_j$$

$$\Rightarrow \frac{1}{m} \mathbf{K}^2 \boldsymbol{\alpha} = \lambda \mathbf{K} \boldsymbol{\alpha}$$ where $\mathbf{K} \in \mathbb{R}^{m \times m}$

$$\Rightarrow \mathbf{K} \boldsymbol{\alpha} = m \lambda \boldsymbol{\alpha}$$ If $\mathbf{K}$ is invertible (strictly pos def)

Kernel PCA

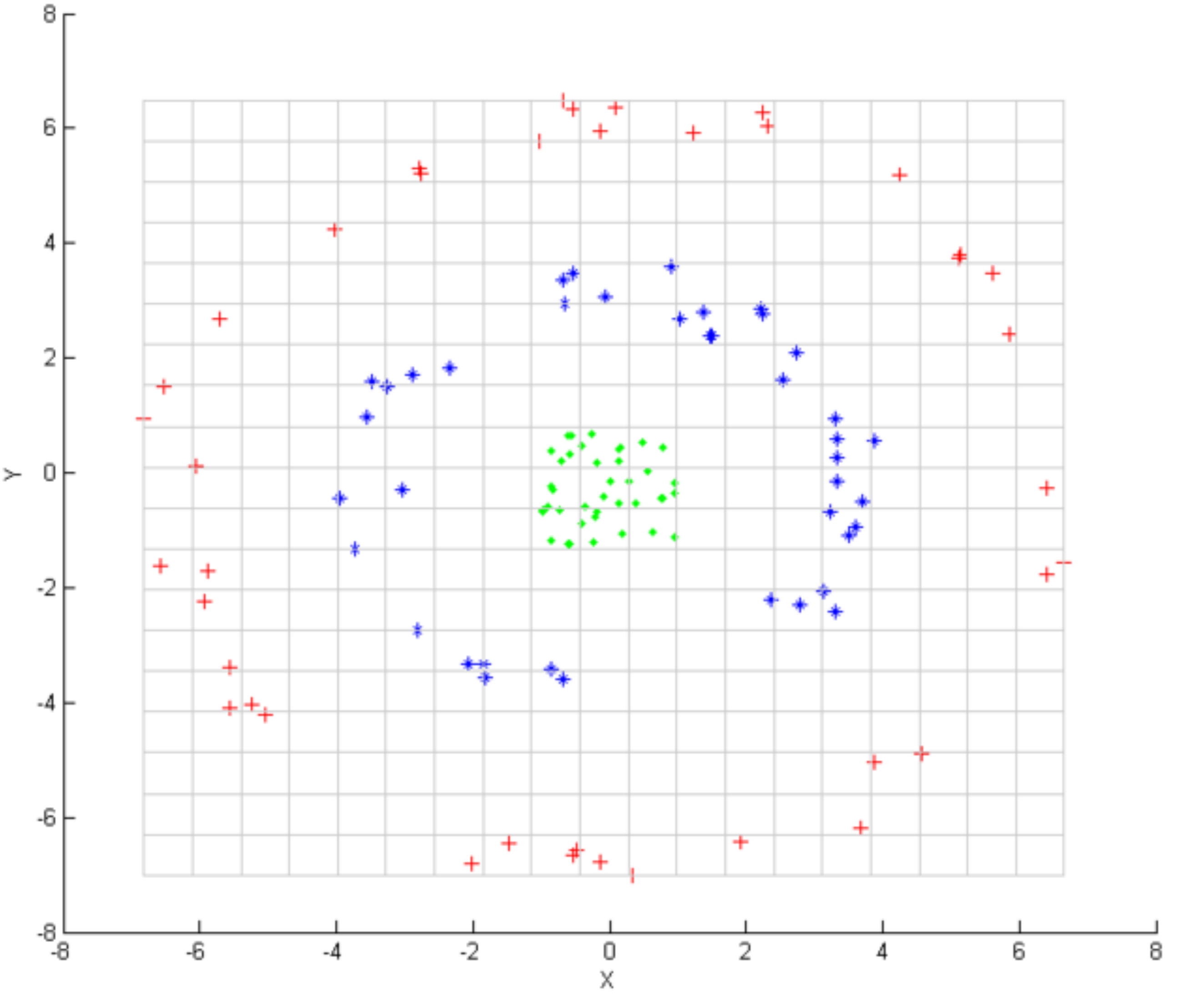
- How to use α to calculate the projection of a new sample *t?*

$$\mathbf{u}^T\mathbf{t} = (\sum_{j=1}^{m} \alpha_j \mathbf{x}_j)^T\mathbf{t} = \sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{t})$$

Again, we don't need values of $\mathbf{x}_j$!

Let $K_{i,j} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$

Input points before kernel PCA



http://en.wikipedia.org/wiki/Kernel_principal_component_analysis

Output after kernel PCA

The three groups are distinguishable using the first component only

$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} + 1)^2$