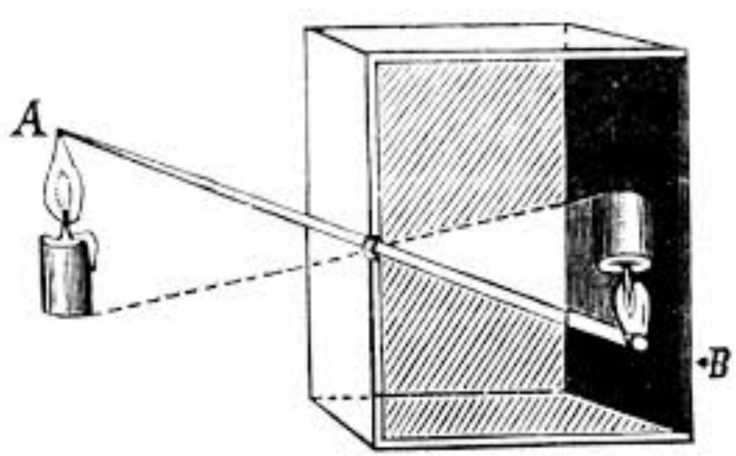
David R. Thompson
Jet Propulsion Laboratory, California Institute of Technology

Linear Dimensionality Reduction

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

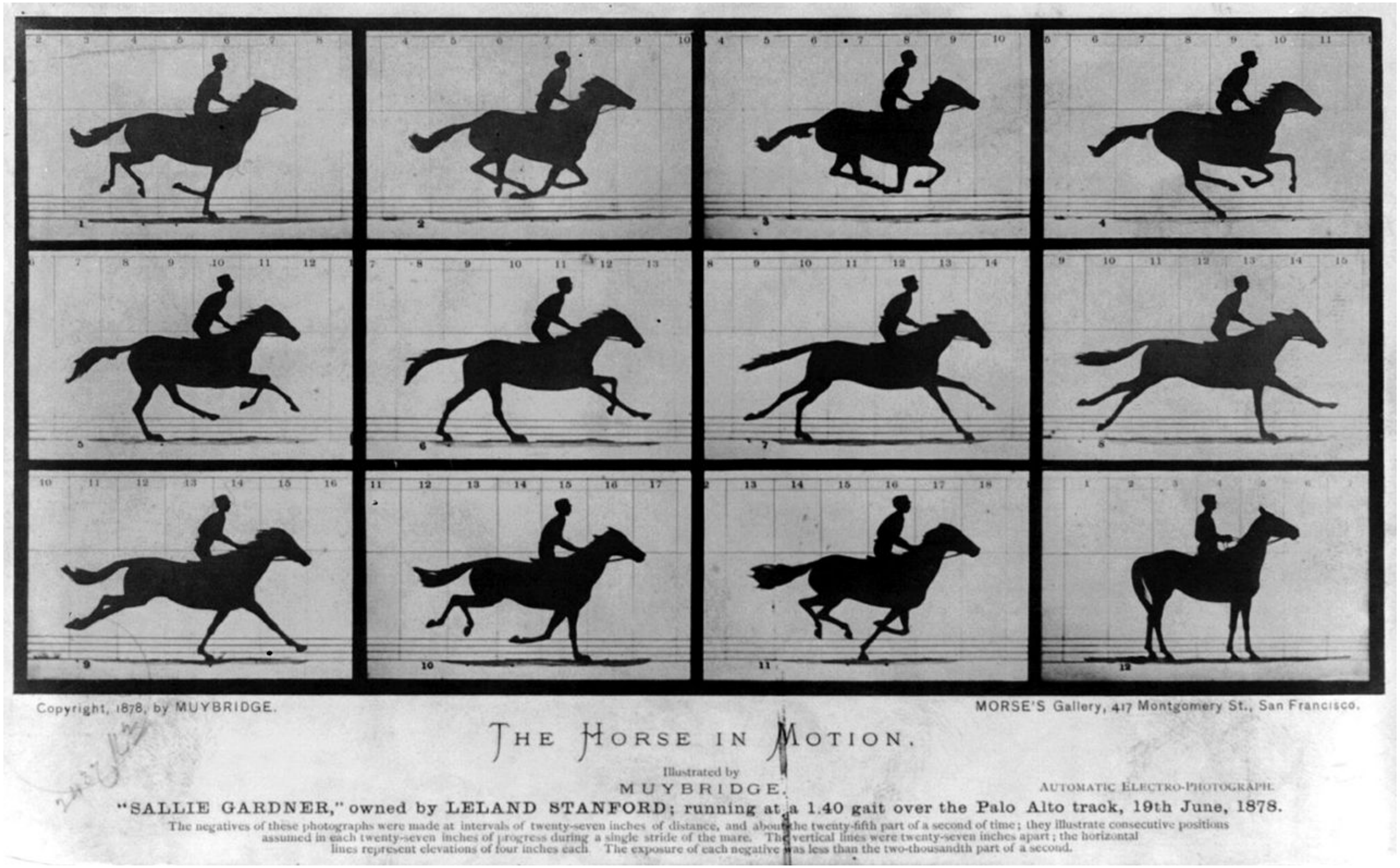
Objectives

1. Introduction to dimensionality reduction and its relationship with feature selection
2. Understand Principal Component Analysis and its relation to SVD
3. Find your eigenface representation



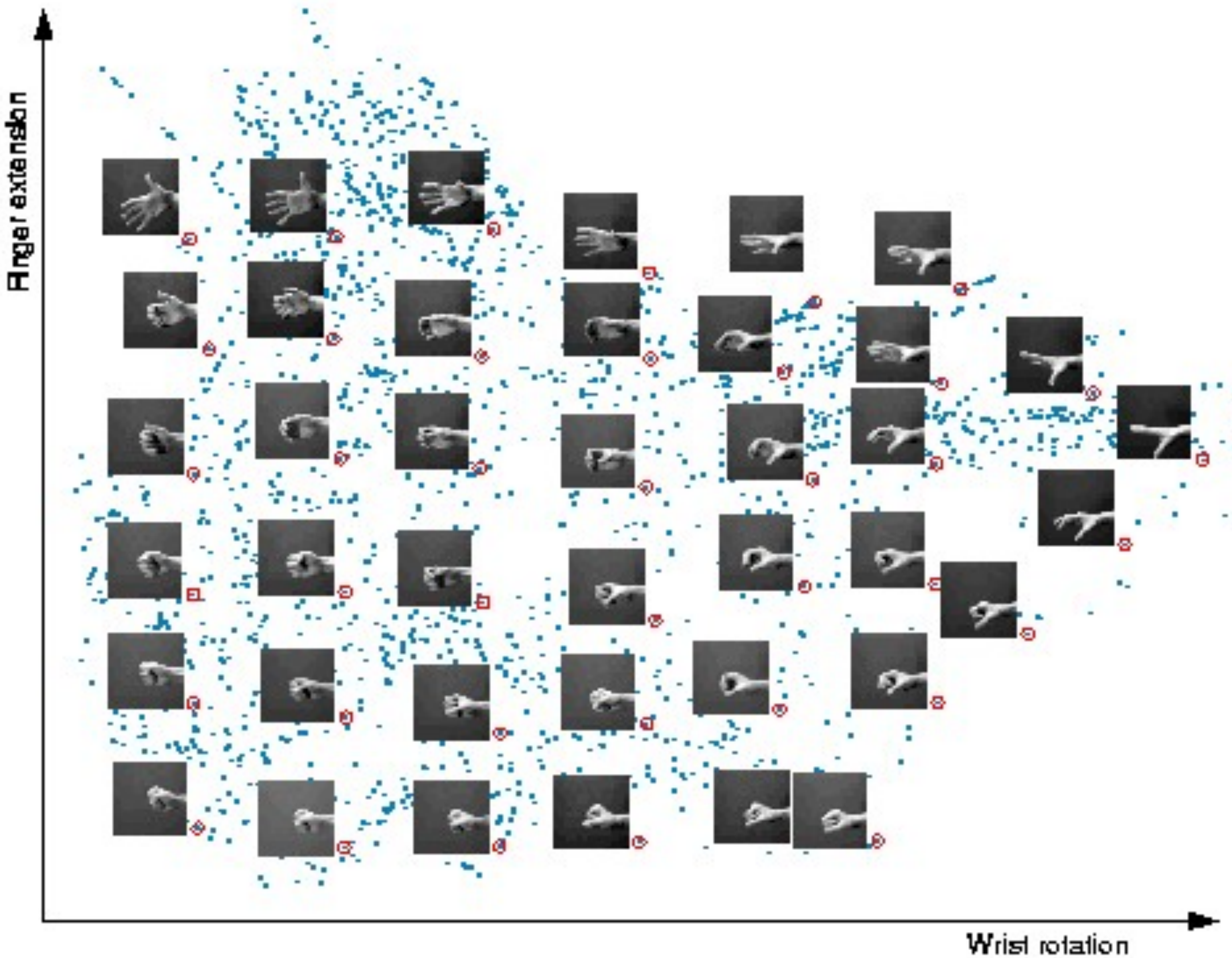
Wikipedia.nl

Dimensionality reduction



Wikipedia

Dimensionality reduction



[Tenenbaum et al., *Science* 2000]

Formal definitions

Data point as an n-dimensional column vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

Data set as a [n x d] matrix

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_d]$$

Formal definitions

Data point as an n-dimensional column vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

Projected datapoint

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix}$$

Data set as a [n x d] matrix

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_d]$$

New data set as a [m x d] matrix

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots \mathbf{y}_d]$$

Dimensionality reduction is a mapping $\mathbf{x} \mapsto \mathbf{y}$

Linear Dimensionality reduction

Define $\mathbf{x} \mapsto \mathbf{y}$ to be a *linear* mapping:

$$\mathbf{Y} = \mathbf{A}\mathbf{X}$$

[m x d] projected data matrix

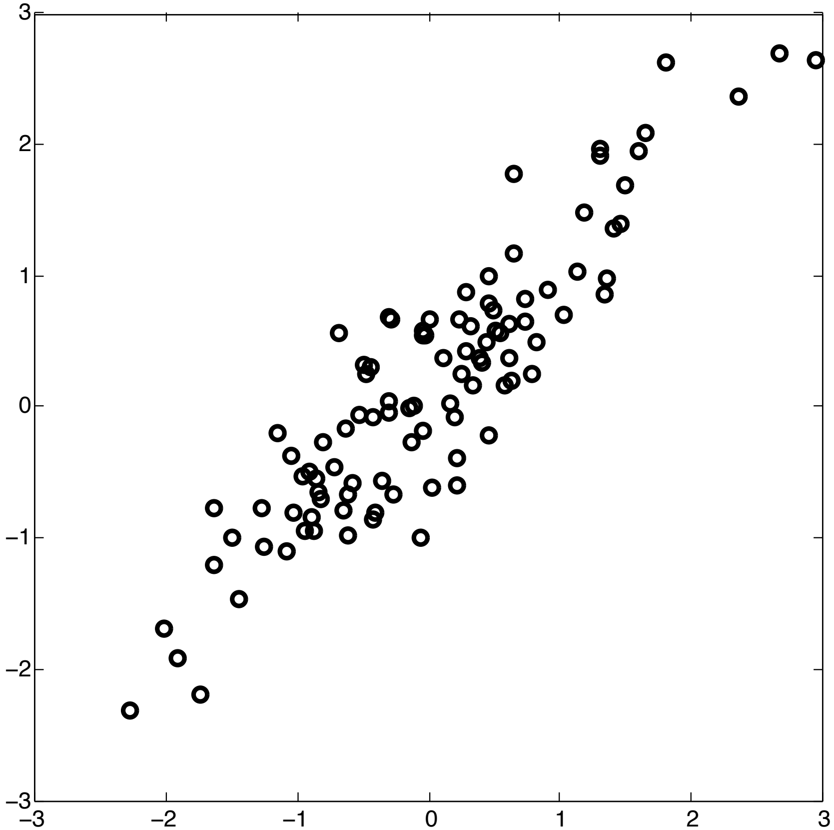
[m x n] projection matrix

[n x d] data matrix (zero-meaned)

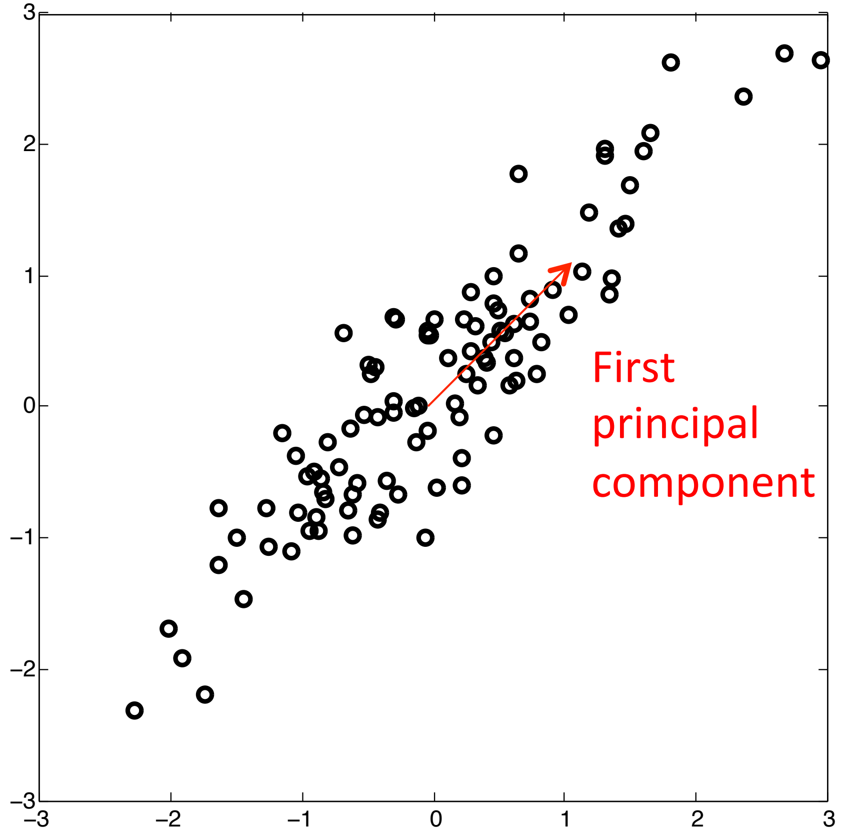
Selects a *subspace* to best represent the data

The most common method is **Principal Component Analysis (PCA)**

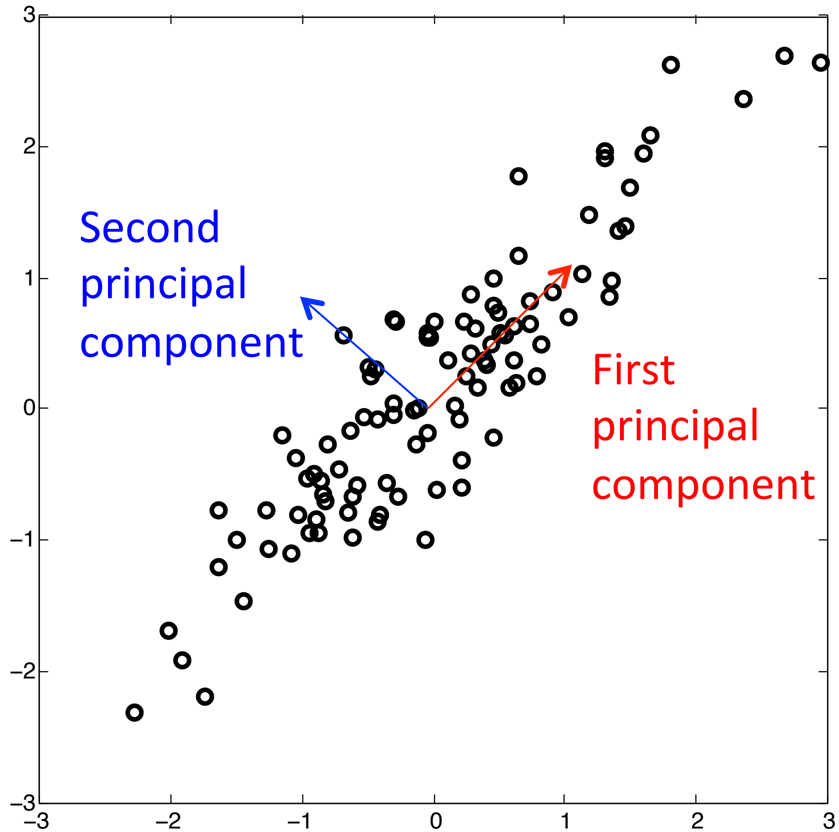
PCA selects orthogonal directions maximizing variance



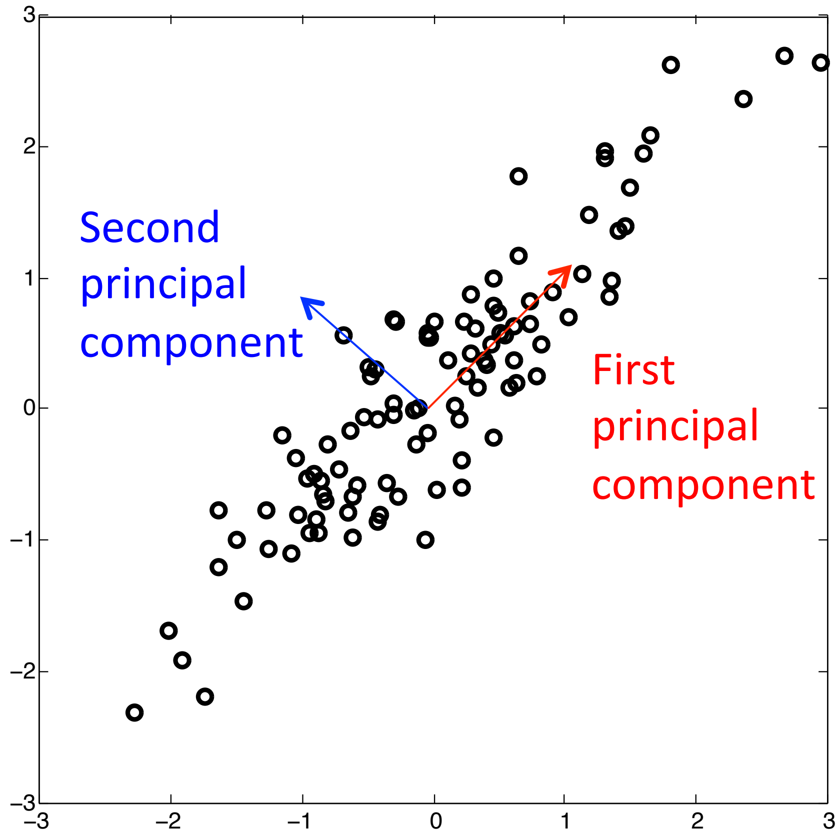
PCA selects orthogonal directions maximizing variance



PCA selects orthogonal directions maximizing variance



PCA selects orthogonal directions maximizing variance



Vectors have unit length - PCA provides an *orthornormal basis*

Covariance matrix

- An m x m matrix, with:

$$\Sigma_{\mathbf{X}}(i,j) = E[(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)]$$

Mean of jth attribute

- Estimate using:

$$\Sigma_{\mathbf{X}} \equiv \frac{1}{d-1}\mathbf{X}\mathbf{X}^T$$

- Properties:
  - Diagonal of $\Sigma_{\mathbf{X}}$ has the *variance* of **x**
  - Off-diagonal terms of $\Sigma_{\mathbf{X}}$ represent *covariances of **x***
  - It's square, symmetric, positive semi-definite

Maximizing variance of a projection (one dimension)

$$\begin{aligned} E[\mathbf{v}^T\mathbf{x} - E[\mathbf{v}^T\mathbf{x}]]^2 &= E[(\mathbf{v}^T[\mathbf{x} - E\mathbf{x}])^2] \\ &= \mathbf{v}^T E[(\mathbf{x} - E\mathbf{x})(\mathbf{x} - E\mathbf{x})^T]\mathbf{v} \\ &= \mathbf{v}^T \Sigma_{\mathbf{X}} \mathbf{v} \end{aligned}$$

For a zero-mean covariance matrix, the unit vector minimizing this quantity is the top eigenvector of Σ_X

$$(\mathbf{X}\mathbf{X}^T)\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

Summary: PCA Recipe

1. Convert the data to have zero mean
2. Form ***A*** using the top n eigenvectors of the sample covariance matrix $\mathbf{XX}^\mathbf{T}$
   – Equivalently, use the left singular vectors of **X** associated with the largest singular values
3. Project the data: $\mathbf{y} = \mathbf{A}(\mathbf{x} - \mu)$
4. To reconstruct: $\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{y} + \mu$

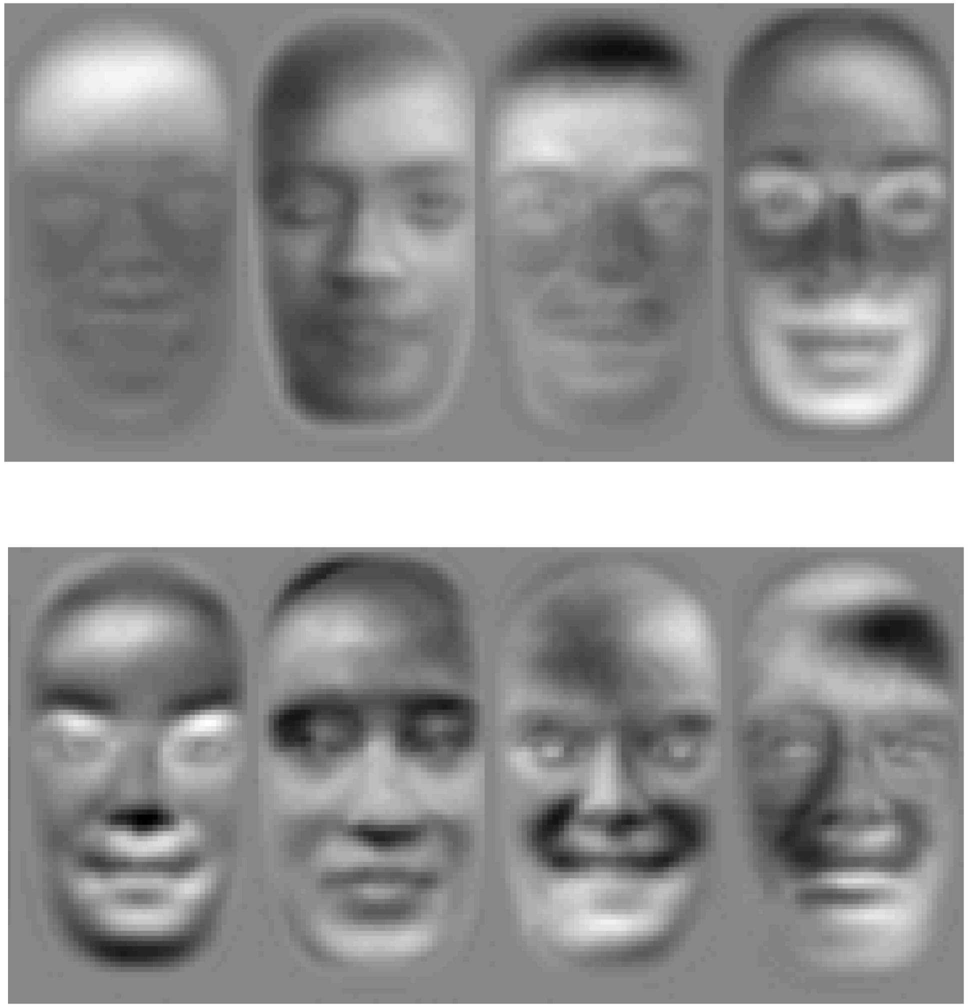
david.r.thompson@jpl.nasa.gov

Example: Eigenfaces

FERET frontal-view image database, from [Moghaddam, Wahid, and Pentland 1998]

Interpreting the principal components

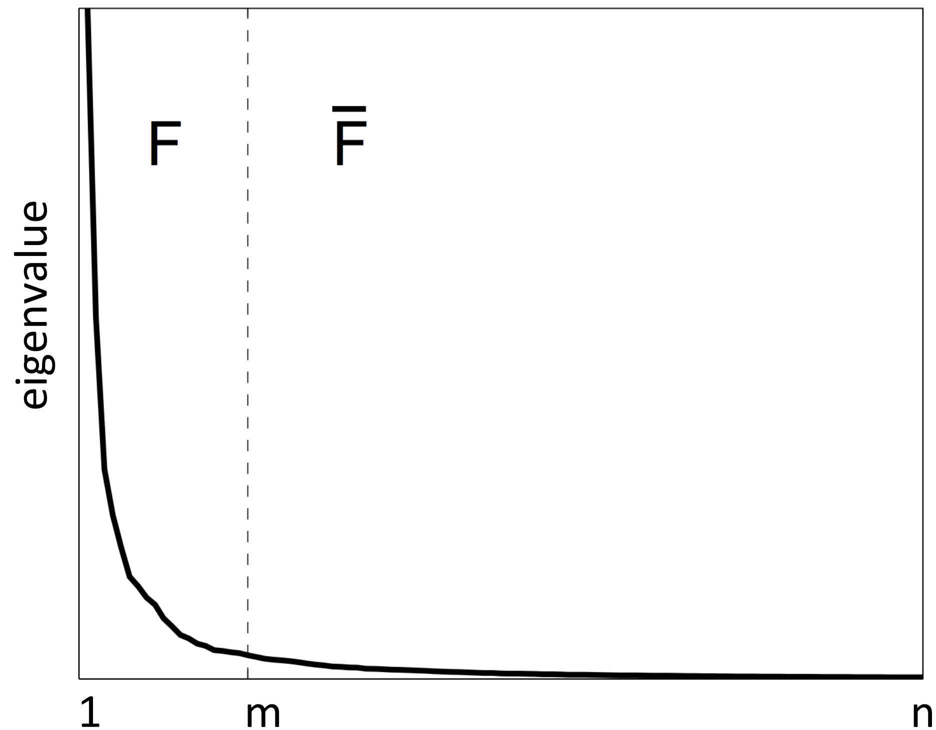
The vectors describe main axes of variation



[Moghaddam, Wahid, and Pentland 1998]

Interpreting eigenvalues

Eigenvalue "fall off" suggests the number of non-noise components



[Moghaddam, Wahid, and Pentland 1998]

Efficient implementations for large *d* or large *n*

- Use Singular Value Decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

U is a valid orthonormal basis

Diagonal matrix of singular values – use instead of covariance matrix eigenvalues

- Calculate one component at a time [e.g. Roweis, *NIPS 1998*]
- Use sequential estimation [e.g. Warmuth & Kuzmin, *JMLR* 2008]

david.r.thompson@jpl.nasa.gov

Summary

- Principal Component Analysis (PCA) is a reliable, standard method for dimensionality reduction and visualization
- It finds an orthonormal basis to maximize the variance of the projected data
- The basis vectors and eigenvalues can provide insight about principal axes of variation in your data

Formal definitions

Data points $\mathbf{x} \in \mathbb{R}^n, \quad \mathbf{y} \in \mathbb{R}^m \quad m << n$

$$\mathbf{x}_i = \underbrace{\begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{in} \end{bmatrix}}_{\text{Data point having n features}} \qquad \mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \cdots \\ y_{im} \end{bmatrix}$$

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_d}] \qquad \mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \ldots \mathbf{y_d}]$$

Data set as an [n x d] matrix

Dimensionality reduction is a mapping $\mathbf{x} \mapsto \mathbf{y}$

Formal definitions

Data point as an n-dimensional row vector

$$\mathbf{x} = [x_1, x_2, \ldots x_n]$$

Projected datapoint

$$\mathbf{y} = [y_1, y_2, \ldots y_m]$$

Data set as a [d x n] matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdots \\ \mathbf{x}_d \end{bmatrix}$$

New data set as a [d x m] matrix

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \cdots \\ \mathbf{y}_d \end{bmatrix}$$

Dimensionality reduction is a mapping $\mathbf{x} \mapsto \mathbf{y}$