Principal Component Analysis

Vojtěch Franc
Center for Machine Perception
Department of Cybernetics, FEE CTU Prage

RPZ ZS 2015

Principal Component Analysis

PCA is a dimension reduction method transforming input n-dimensional data set

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathbb{R}^{n\times m}$$

by a applying a linear orthogonal transform on centered version of $\mathbf{X}$, i.e.

$$\mathbf{z} = \mathbf{W}^T(\mathbf{x}-\mu) = \mathbf{W}^T\bar{\mathbf{x}} = \begin{bmatrix} \mathbf{w}_1^T(\mathbf{x}-\mu) \\ \mathbf{w}_2^T(\mathbf{x}-\mu) \\ \vdots \\ \mathbf{w}_d^T(\mathbf{x}-\mu) \end{bmatrix} \quad \text{where} \quad \mathbf{W} \in \mathbb{R}^{n\times d}, \mathbf{W}^T\mathbf{W} = \mathbf{I}, \mu = \frac{1}{m}\sum_{j=1}^{m} \mathbf{x}_j \,,$$

which yields a lower $(d < n)$-dimensional representation, so called principal scores,

$$\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m) \in \mathbb{R}^{d\times m}$$

with the following properties:

- $\mathbf{Z}$ retains as much as possible of the variation presented in the data set $\mathbf{X}$
- $\mathbf{Z}$ is centered and uncorrelated
- $\mathbf{Z}$ are coordinates of points $\mathbf{X}'$ obtained by projecting $\mathbf{X}$ onto an d-dimensional affine subspace such that the approximation error $\|\mathbf{X} - \mathbf{X}'\|_F$ is minimal possible.

Orthogonal transformation retaining the maximum of variance

- The projection vectors $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_d) \in \mathbb{R}^{n \times d}$ are called the principal components.
- The first principal component is a unit vector $\mathbf{w}_1 \in \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{w}\| = 1\}$ maximizing the variance

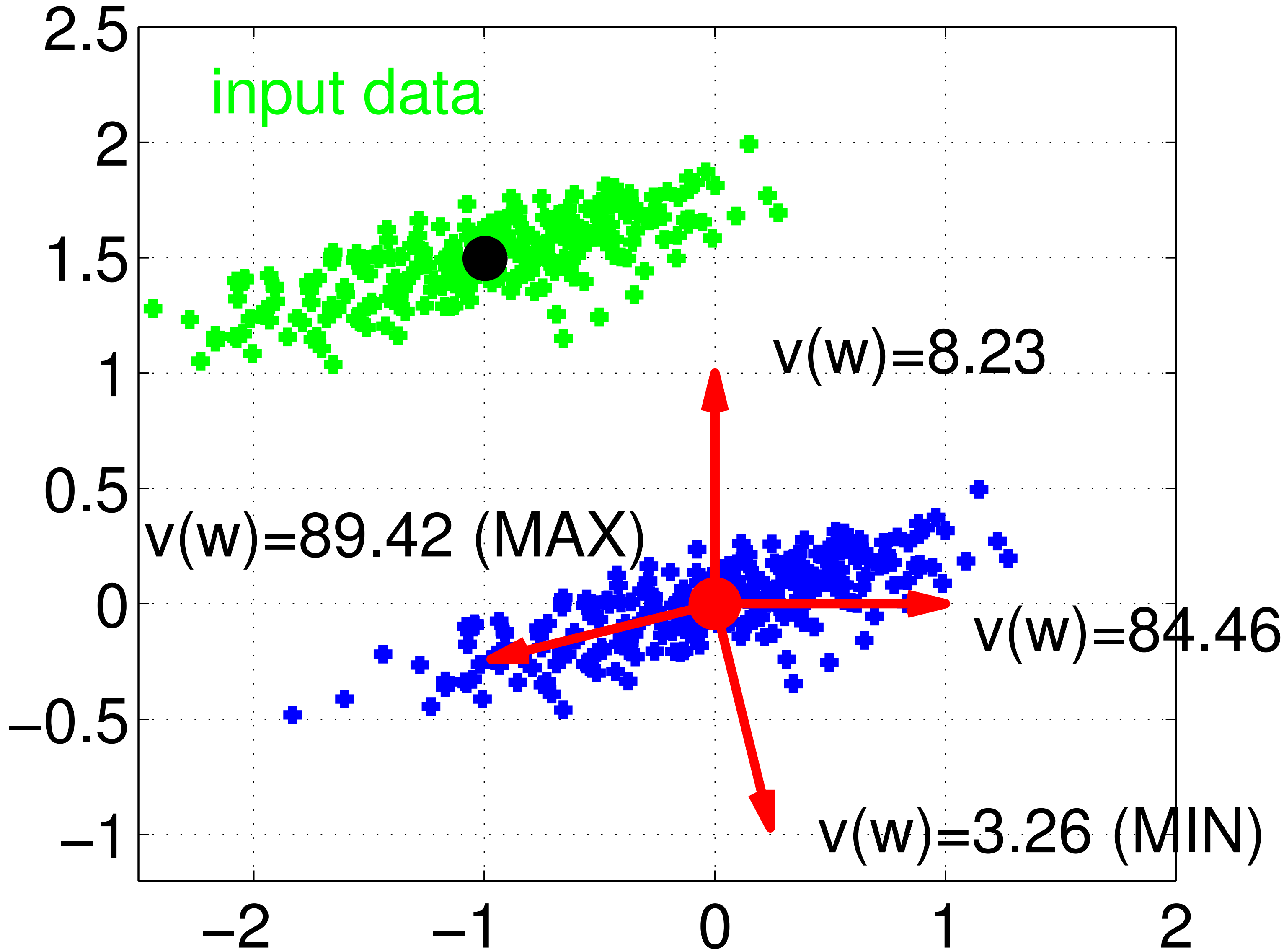
$$v(\mathbf{w}) = \mathrm{var}(\mathbf{w}^T(\mathbf{x} - \mu)) = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{w}^T(\mathbf{x}_j - \mu))^2 = \mathbf{w}^T\mathbf{C}\mathbf{w}$$

where $\mathbf{C} \in \mathbb{R}^{n \times n}$ is the sample covariance matrix defined as

$$\mathbf{C} = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{x}_j - \mu)(\mathbf{x}_j - \mu)^T$$

- $(k \geq 2)$-th principal component is a unit vector $\mathbf{w}_k \in \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{w}\| = 1\}$ maximizing the variance $v(\mathbf{w})$ and being orthogonal to all previous principal components, i.e. $\mathbf{w}_k^T\mathbf{w}_i = 0,\ i = 1, \ldots, k-1$.

Directions with maximal variance



$$v(\mathbf{w}) = \mathrm{var}(\mathbf{w}^T(\mathbf{x}-\mu)) = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{w}^T(\mathbf{x}_j-\mu))^2 = \mathbf{w}^T\mathbf{C}\mathbf{w}$$

Finding the first principal component

- Finding the first principal component amounts to solving

$$\mathbf{w}_1 = \operatorname*{argmax}_{\mathbf{w}\in\mathbb{R}^n} \mathbf{w}^T\mathbf{C}\mathbf{w} \qquad \text{s.t.} \qquad \|\mathbf{w}\|^2 = 1 \tag{1}$$

- The first order conditions for $\mathbf{w}_1$ to solve (1): there exist $\lambda_1 \in \mathbb{R}$ such that $\partial_{\mathbf{w}} L(\mathbf{w}_1, \lambda_1) = 0$ and $\partial_{\lambda} L(\mathbf{w}_1, \lambda_1) = 0$ where

$$L(\mathbf{w}_1, \lambda_1) = \mathbf{w}_1^T\mathbf{C}\mathbf{w}_1 - \lambda_1(\|\mathbf{w}_1\|^2 - 1)$$

  is the Lagrange function of (1).
- The condition $\partial_{\mathbf{w}} L(\mathbf{w}_1, \lambda_1) = 0$ is a set of n non-linear equations with variables $\mathbf{w}_1 \in \mathbb{R}^n$ and $\lambda_1 \in \mathbb{R}$ known as the Eigenvalue problem

$$\mathbf{C}\mathbf{w}_1 = \lambda_1\mathbf{w}_1$$

  which for symmetric PSD matrix $\mathbf{C}$ has n solutions: eigen-values $(\lambda'_1, \ldots, \lambda'_n) \in \mathbb{R}^n$ and associated orthogonal eigen-vectors $\mathbf{W}' = (\mathbf{w}'_1, \ldots, \mathbf{w}'_n) \in \mathbb{R}^{n\times n}$, $\mathbf{W}'^T\mathbf{W}' = \mathbf{I}$.
- The first principal component is the eigen-vector with the highest eigenvalue because

$$v(\mathbf{w}'_i) = \mathbf{w}'^T_i\mathbf{C}\mathbf{w}'_i = \lambda'_i$$

Finding the second and other principal components

- Finding the second principal component amounts to solving

$$\mathbf{w}_2 = \operatorname*{argmax}_{\mathbf{w}\in\mathbb{R}^n} \mathbf{w}^T\mathbf{C}\mathbf{w} \qquad \text{s.t.} \qquad \|\mathbf{w}\|^2 = 1 \quad \text{and} \quad \mathbf{w}^T\mathbf{w}_1 = 0 \tag{2}$$

- The first order condition for $\mathbf{w}_2$: there exists $\lambda_1 \in \mathbb{R}$, $\lambda_2 \in \mathbb{R}$ such that $\partial_{\mathbf{w}} L(\mathbf{w}_2, \lambda_1, \lambda_2) = 0$, $\partial_{\lambda_1} L(\mathbf{w}_2, \lambda_1, \lambda_2) = 0$ and $\partial_{\lambda_2} L(\mathbf{w}_2, \lambda_1, \lambda_2) = 0$ where

$$L(\mathbf{w}_2, \lambda_1, \lambda_2) = \mathbf{w}_2^T\mathbf{C}\mathbf{w}_2 - \lambda_1\mathbf{w}_1^T\mathbf{w}_2 - \lambda_2(\|\mathbf{w}_2\|^2 - 1)$$

- The condition $\partial_{\mathbf{w}} L(\mathbf{w}_2, \lambda_1, \lambda_2) = 0$ implies that

$$\begin{aligned}
2\mathbf{C}\mathbf{w}_2 - \lambda_1\mathbf{w}_1 - 2\lambda_2\mathbf{w}_2 &= 0 \\
2\underbrace{\mathbf{w}_1^T\mathbf{C}\mathbf{w}_2}_{=\lambda_1\mathbf{w}_1^T\mathbf{w}_2=0} - \lambda_1\mathbf{w}_1^T\mathbf{w}_1 - 2\lambda_2\underbrace{\mathbf{w}_1^T\mathbf{w}_2}_{=0} &= 0 \qquad \Rightarrow \qquad \lambda_1 = 0 \\
\mathbf{C}\mathbf{w}_2 &= \lambda_2\mathbf{w}_2
\end{aligned}$$

- The last line is again the Eigenvalue problem and thus the second principal component is the eigen-vector with the second largest eigen-value.
- The 3rd, 4th, $\cdots$, d-th principal components are found analogically.

PCA: The Algorithm

1: Input: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathbb{R}^{n\times m}$ and $d \in [0, n]$.
2: Output: $\mathbf{W} \in \mathbb{R}^{n\times d}$, $\mu \in \mathbb{R}^n$
3: Compute mean and covariance

$$\mu = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i \qquad \mathbf{C} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

4: Find d eigen-vectors $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_d)$ with highest eigen-values, i.e. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, of the eigen-value problem

$$\mathbf{C}\mathbf{w} = \lambda\mathbf{w}$$

Transform the data $\mathbf{X}$ onto a PCA scores $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m) \in \mathbb{R}^{d\times m}$ by

$$\mathbf{z}_i = \mathbf{W}^T(\mathbf{x}_i - \mu)\,, \qquad i \in \{1, \ldots, m\}$$

PCA: Matlab code


```
function [W,mu] = pca( X, d)
% PCA Principal Component Analysis
% Synopsis:
%    [W,mu] = pca( X, d)
%
% Input:
%   X   [n x m] Input data
%   d   [1 x 1] Output dimension
% Output:
%   W   [n x d] Principal components
%   mu  [m x 1] Data mean
%
   mu       = mean(X,2);
   C        = cov( X', 1);
   [V,D]    = eig( C );
   [~,idx]  = sort( diag(D), 'descend');
   W        = V(:,idx(1:d));
end

% Lower dimensional data representation of X
Z = W'*( X - repmat( mu, 1, size(X,2) ));
```

The PCA scores are centered and uncorrelated

- The PCA scores $\mathbf{z}_i = \mathbf{W}^T(\mathbf{x}_i - \mu)$, $i \in \{1, \ldots, m\}$, are centered

$$\tilde{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{z}_i = \frac{1}{m}\sum_{i=1}^{m}\mathbf{W}^T(\mathbf{x}_i - \mu) = \mathbf{W}^T\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i\right) - \mathbf{W}^T\mu = \mathbf{0}$$

- The PCA scores are uncorrelated

$$\tilde{\mathbf{C}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{z}_i^T\mathbf{z}_i = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{W}^T(\mathbf{x}_i - \mu))(\mathbf{W}^T(\mathbf{x}_i - \mu))^T = \mathbf{W}^T\mathbf{C}\mathbf{W}$$

so that

$$\tilde{C}_{i,j} = \mathbf{w}_i^T\mathbf{C}\mathbf{w}_j = \lambda_j\mathbf{w}_i^T\mathbf{w}_j = \lambda_i\mathbf{w}_i^T\mathbf{w}_j = \begin{cases} 0 & \text{for} \quad i \neq j \\ \lambda_i & \text{for} \quad i = j \end{cases}$$

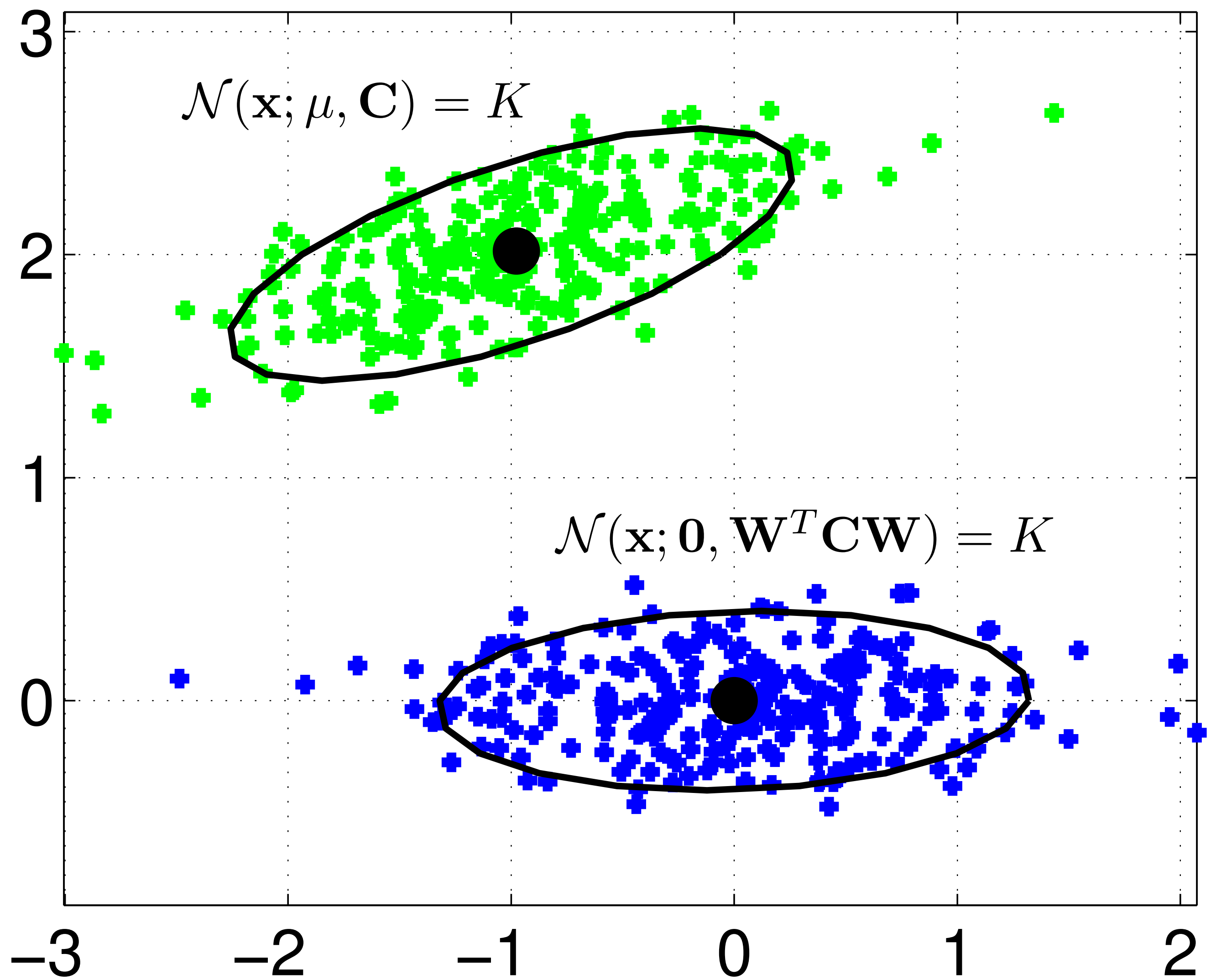
and the retained variance is

$$\operatorname{tr}(\mathbf{W}^T\tilde{\mathbf{C}}\mathbf{W}) = \sum_{i=1}^{d}\lambda_i\mathbf{w}_i^T\mathbf{w}_i = \sum_{i=1}^{d}\lambda_i$$

The PCA scores are centered and uncorrelated

The PCA can be interpreted as follows:

1. Center and rotate the data such that they become uncorrelated.
2. Forget the dimensions with lowest variance.



PCA presents data in affine sub-space

- The PCA scores $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_m) \in \mathbb{R}^{d \times m}$ are coordinates of the original data $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m) \in \mathbb{R}^{n \times m}$ projected onto a d-dimensional affine sub-space

$$\mathcal{P} = \left\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \mathbf{W}\mathbf{z} + \mu, \mathbf{z} \in \mathbb{R}^d \right\} \subseteq \mathbb{R}^n$$

- The orthogonal projection of $\mathbf{x}$ onto an affine sub-space $\mathcal{P}$ is

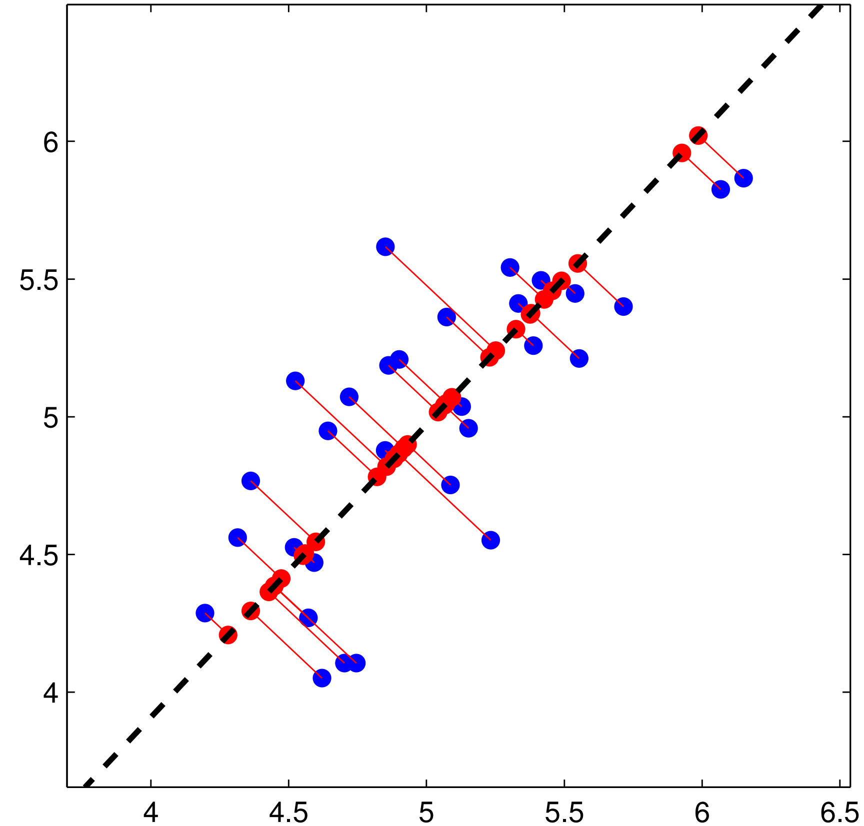
$$\tilde{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}' \in \mathcal{P}} \|\mathbf{x} - \mathbf{x}'\|^2 = \mathbf{W}\mathbf{z} + \mu = \mathbf{w}_1 z_1 + \mathbf{w}_2 z_2 + \cdots \mathbf{w}_d z_d + \mu$$

where $\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \mu)$.

This follows from

$$\begin{aligned} \mathbf{z} &= \operatorname*{argmin}_{\mathbf{z}' \in \mathbb{R}^d} \|\mathbf{W}\mathbf{z}' + \mu - \mathbf{x}\|^2 \\ &= (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T(\mathbf{x} - \mu) \\ &= \mathbf{W}^T(\mathbf{x} - \mu) \end{aligned}$$

which is so called least squares problem.



PCA minimizes the reconstruction error

- Let $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m) \in \mathbb{R}^{n \times m}$ be the points reconstructed from the PCA scores $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m) \in \mathbb{R}^{d \times m}$ by

$$\tilde{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i + \mu \qquad \text{where} \qquad \mathbf{z}_i = \mathbf{W}^T(\mathbf{x}_i - \mu)$$

- The reconstruction error is

$$\mathrm{Err}(\mathbf{W}, \mu, \mathbf{Z}) = \|\tilde{\mathbf{X}} - \mathbf{X}\|_F^2 = \sum_{i=1}^{m} \|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|^2 = \sum_{i=1}^{m} \|\mathbf{W}\mathbf{z}_i + \mu - \mathbf{x}_i\|^2$$

- The PCA is the optimal solution of the problem

$$\min_{\mathbf{W}' \in \mathbb{R}^{n \times d}, \mu' \in \mathbb{R}^n, \mathbf{Z}' \in \mathbb{R}^{d \times m}} \mathrm{Err}(\mathbf{W}', \mu', \mathbf{Z}')$$

  i.e. it minimizes the reconstruction error which equals to

$$\mathrm{Err}(\mathbf{W}, \mu, \mathbf{Z}) = \sum_{i=d+1}^{n} \lambda_i$$

- The cumulative sum of the sorted eigen-vectors can be used to select the output dimension d.

Example: Eigenfaces

- The face image represented by column vector $\mathbf{x} \in \mathbb{R}^n$ containing the intensity values is compressed to PCA scores

$$\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \mu) = (\underbrace{\mathbf{w}_1^T(\mathbf{x} - \mu)}_{z_1}, \ldots, \underbrace{\mathbf{w}_d^T(\mathbf{x} - \mu)}_{z_d})^T$$

- Face $\mathbf{x}$ is approximated by a linear combination of d-principal components, so called "eigenfaces":

$$\mathbf{x} \qquad \tilde{\mathbf{x}} = \mu + z_1\mathbf{w}_1 + z_2\mathbf{w}_2 + z_3\mathbf{w}_3 + z_4\mathbf{w}_4 + z_5\mathbf{w}_5 + z_6\mathbf{w}_6 + z_7\mathbf{w}_7 + z_8\mathbf{w}_8$$

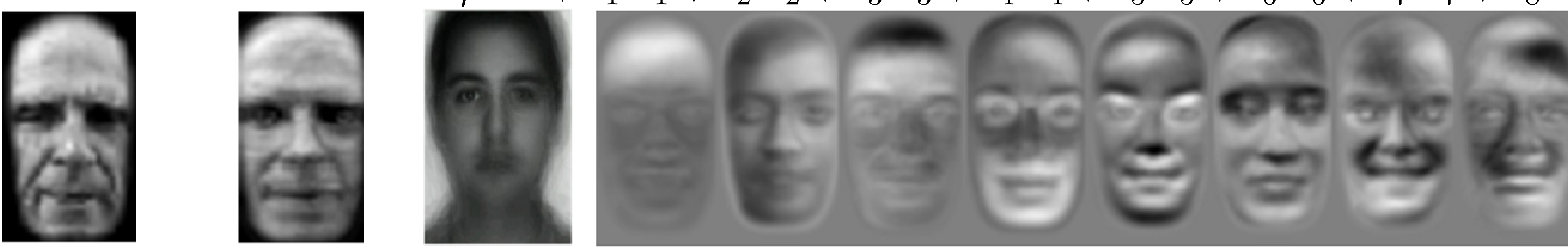


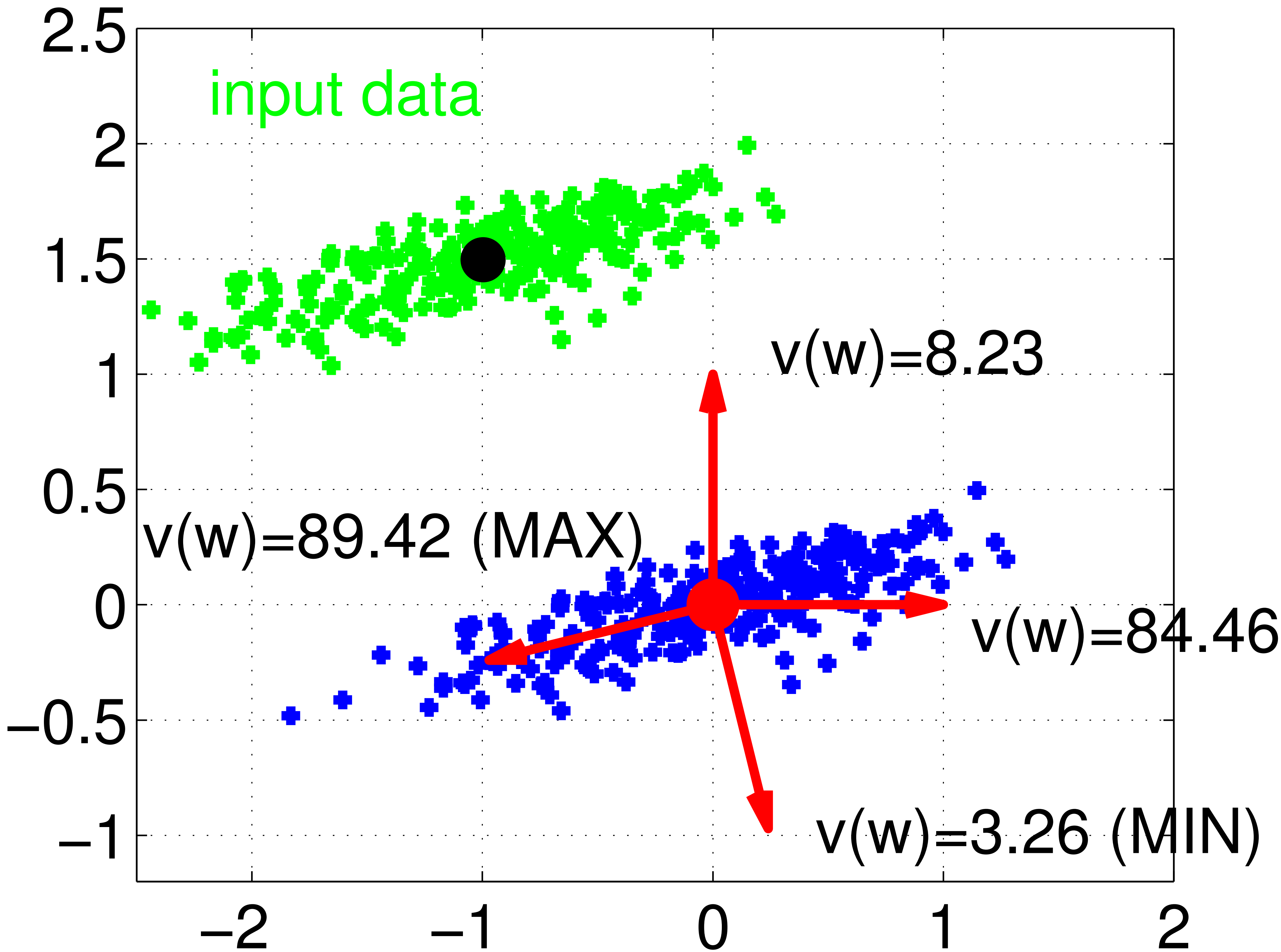
Image originates from http://vision.stanford.edu/teaching/cs231a/lecture/lecture2_face_r%ecognition_cs231a_marked.pdf

PCA: Summary

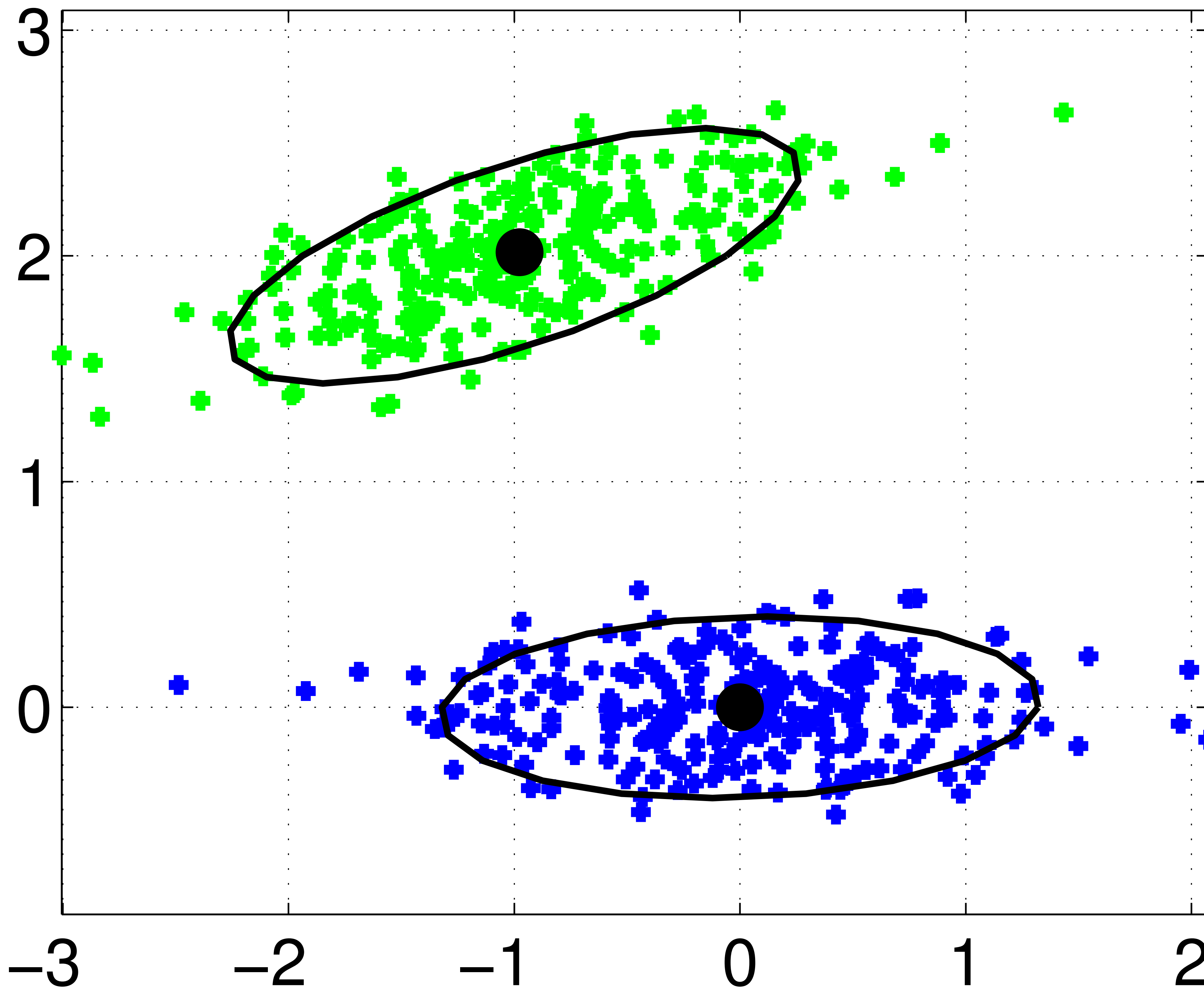
- The PCA represents $\mathbf{X} \in \mathbb{R}^{n \times m}$ as coordinates $\mathbf{Z} \in \mathbb{R}^{d \times m}$ in an affine sub-space with the following properties:
  - $\mathbf{Z}$ retains as much as possible of the variation presented in the data $\mathbf{X}$
  - $\mathbf{Z}$ is centered and uncorrelated
  - $\mathbf{Z}$ is the optimal compression minimizing the reconstruction error $\|\tilde{\mathbf{X}} - \mathbf{X}\|_F$.
- Typical usage of PCA:
  - Feature extraction
  - Compression
  - Visualization
  - Denoising
- PCA is an unsupervised method (no labels are required).

END

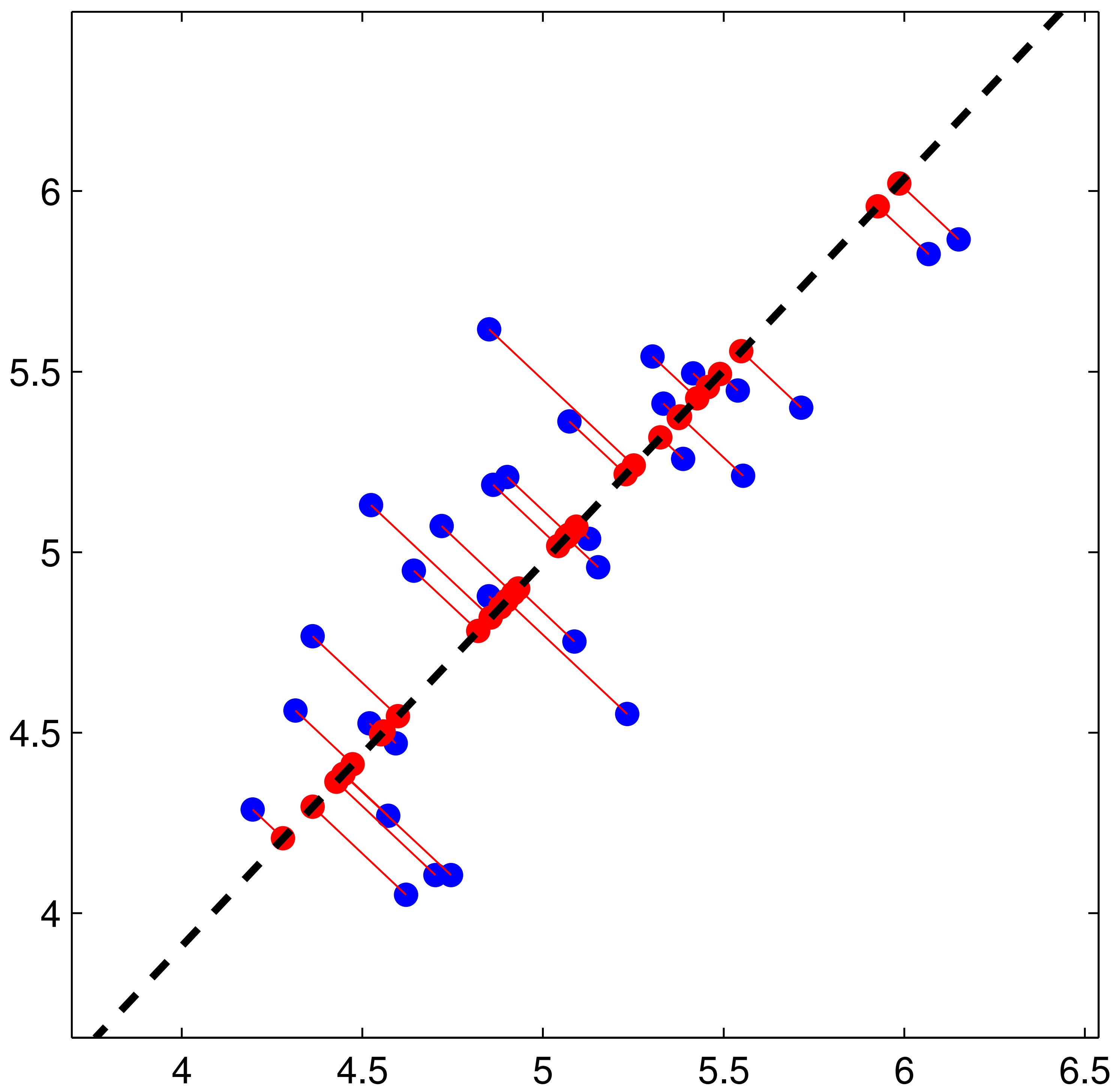
m p



input data
v(w)=8.23
v(w)=89.42 (MAX)
v(w)=84.46
v(w)=3.26 (MIN)
2.5
2
1.5
1
0.5
0
−0.5
−1
−2
−1
0
1
2



3
2
1
0
−3
−2
−1
0
1
2


6
5.5
5
4.5
4
4
4.5
5
5.5
6
6.5