Week 5

Dimensionality Reduction

Emre Ugur, BM 33

emre.ugur@boun.edu.tr

http://www.cmpe.boun.edu.tr/~emre/courses/cmpe462

cmpe462@listeci.cmpe.boun.edu.tr

Why Reduce Dimensionality?

1. Reduces time complexity: Less computation
2. Reduces space complexity: Less parameters
3. Saves the cost of observing the feature
4. Simpler models are more robust on small datasets
5. More interpretable; simpler explanation
6. Data visualization (structure, groups, outliers, etc) if plotted in 2 or 3 dimensions

Feature Selection vs Extraction

- Feature selection: Choosing $k<d$ important features, ignoring the remaining $d - k$

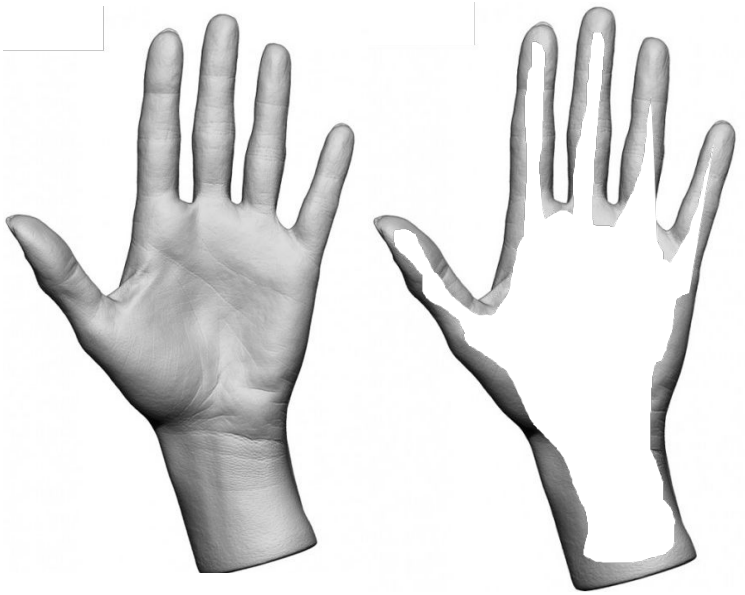
  Subset selection algorithms

- Feature extraction: Project the original x_i , i =1,...,d dimensions to new $k<d$ dimensions, z_j , j =1,...,k

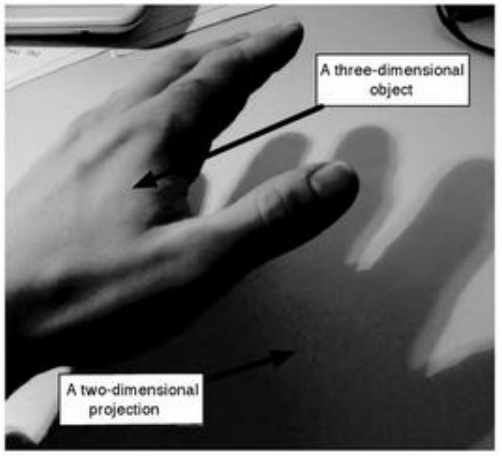
  - Principal components analysis (PCA)
  - Feature embedding
  - Factor analysis (FA)
  - Multidimensional scaling
  - Linear discriminant analysis (LDA),
  - Canonnical correlation analysis (CCA)

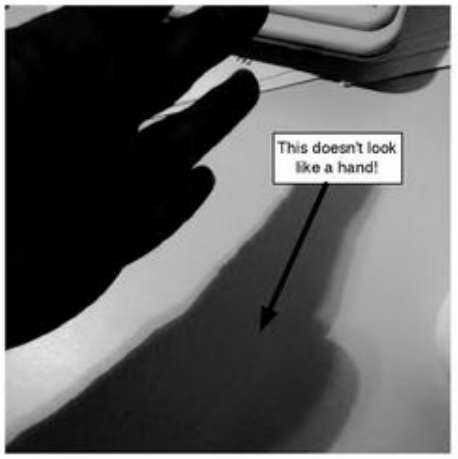
Feature Selection vs Extraction

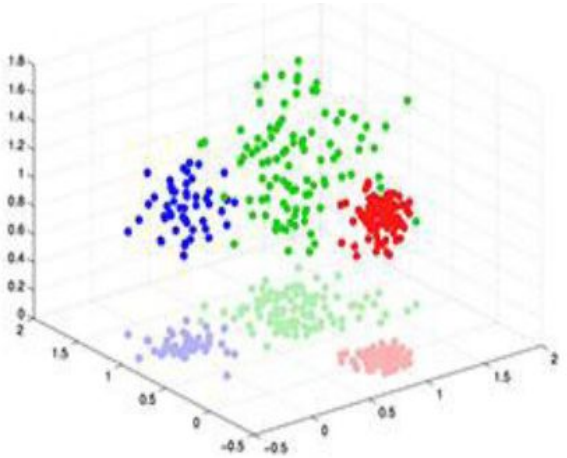
- Feature selection



- Feature extraction (through projection)







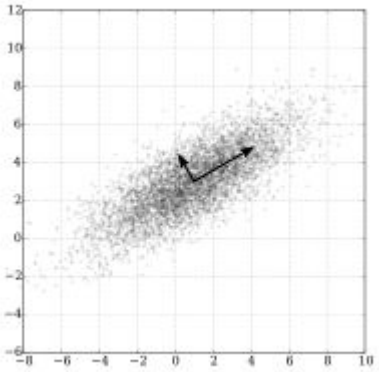
Feature Selection: Subset selection

- How many subsets of *d* features?

- Forward search: Add the best feature at each step
  - Set of features *F* initially Ø.
  - At each iteration, find the best new feature

    $$j = \text{argmin}_i E (F \cup x_i)$$
  - Add x_j to *F* if $E (F \cup x_j) < E (F)$
  - Local search! no guarantees.. why?
- O(?)
- Backward search: Start with all features and remove one at a time, if possible. O(?)
- Floating search (Add *k*, remove *l*)

Feature Extraction: PCA



- Principle Component Analysis
  - Unsupervised method
  - Mapping from original d-dimensional space to k-dimensional space
    - With minimum loss of information
  - Maximize the variance in k-dimensions

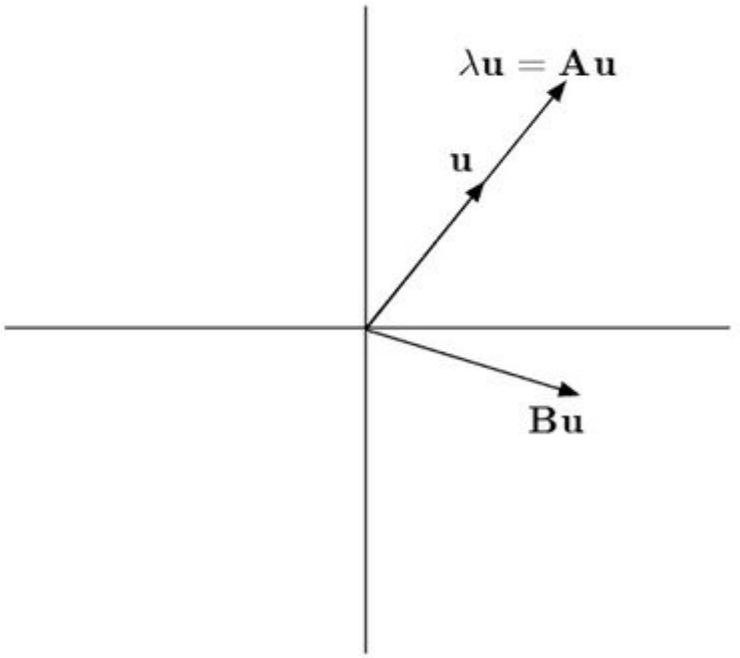
Linear algebra review

Comment 7.1 – Eigenvectors and eigen values: The eigenvector/eigenvalue equation for some square matrix $\mathbf{A}$ is given as:

$$\lambda_i \mathbf{u}_i = \mathbf{A}\mathbf{u}_i. \tag{7.4}$$

The solutions to this equation are pairs of eigenvalues (λ_i) and eigenvectors ($\mathbf{u}_i$).

The figure on the right provides some intuition for this equation. Multiplying an M-dimensional vector $\mathbf{u}$ by an $M \times M$ matrix $\mathbf{B}$ results in another M-dimensional vector. Therefore, we can consider the matrix $\mathbf{B}$ as defining a rotation of the vector $\mathbf{u}$. Different $\mathbf{B}$ matrices will produce different rotations. The solutions to Equation 7.3 for a particular matrix $\mathbf{A}$ are the vectors $\mathbf{u}$ for which applying the rotation $\mathbf{A}$ only results in a change in the length of $\mathbf{u}$. The magnitude of this change is given by the scalar λ.



$$|A| = \prod_{i=1}^{n} \lambda_i$$

Linear algebra review

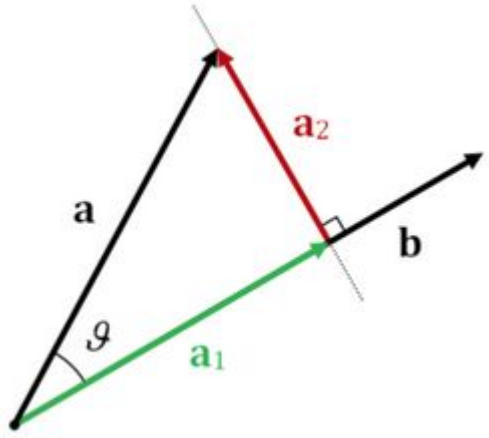
- Derivatives

$f(\mathbf{w})$	$\frac{\partial f}{\partial \mathbf{w}}$
$\mathbf{w}^\mathsf{T}\mathbf{x}$	$\mathbf{x}$
$\mathbf{x}^\mathsf{T}\mathbf{w}$	$\mathbf{x}$
$\mathbf{w}^\mathsf{T}\mathbf{w}$	$2\mathbf{w}$
$\mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w}$	$2\mathbf{C}\mathbf{w}$

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$(\mathbf{X}\mathbf{w})^\mathsf{T} = \mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}$$

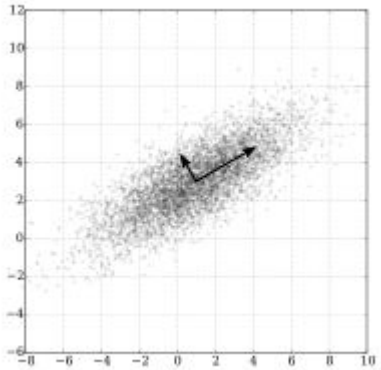
- Projection



$$\mathbf{a}_1 = a_1 \hat{\mathbf{b}}$$

$$a_1 = |\mathbf{a}| \cos\theta = \mathbf{a} \cdot \hat{\mathbf{b}} = \mathbf{a} \cdot \frac{\mathbf{b}}{|\mathbf{b}|}$$

Principle Component Analysis (PCA)



- Principle Component Analysis
  - Unsupervised method
  - Mapping from original d-dimensional space to k-dimensional space
    - With minimum loss of information
  - Maximize the variance in k-dimensions

- Assume original dimensions has zero mean
- Projection of ***x*** on the direction ***w*** is

$$z = \boldsymbol{w}^T \boldsymbol{x}$$

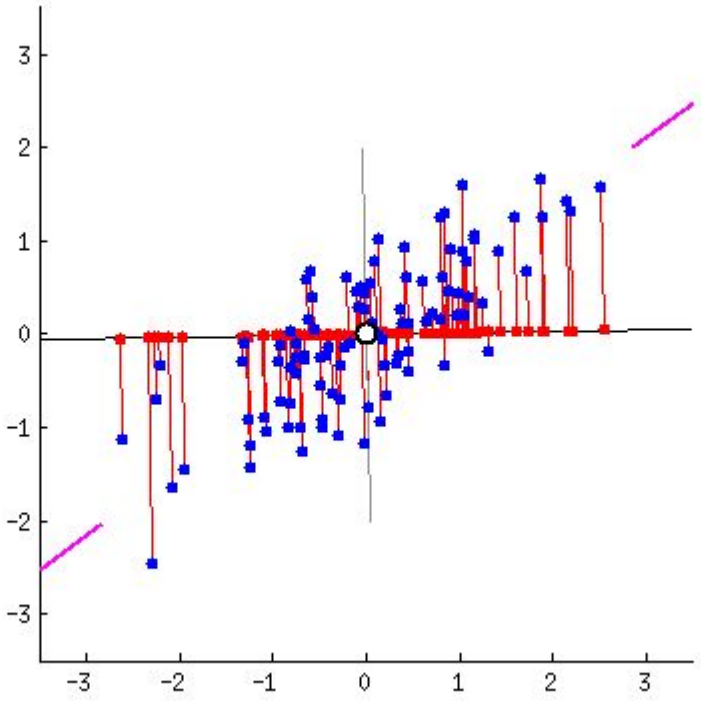
- w is unit length
- Maximize the variance

$$\|\boldsymbol{w}_1\| = 1$$

$$\Sigma \boldsymbol{w}_1 = \alpha \boldsymbol{w}_1$$

$$\Sigma \boldsymbol{w}_2 = \alpha \boldsymbol{w}_2$$

Var(z)

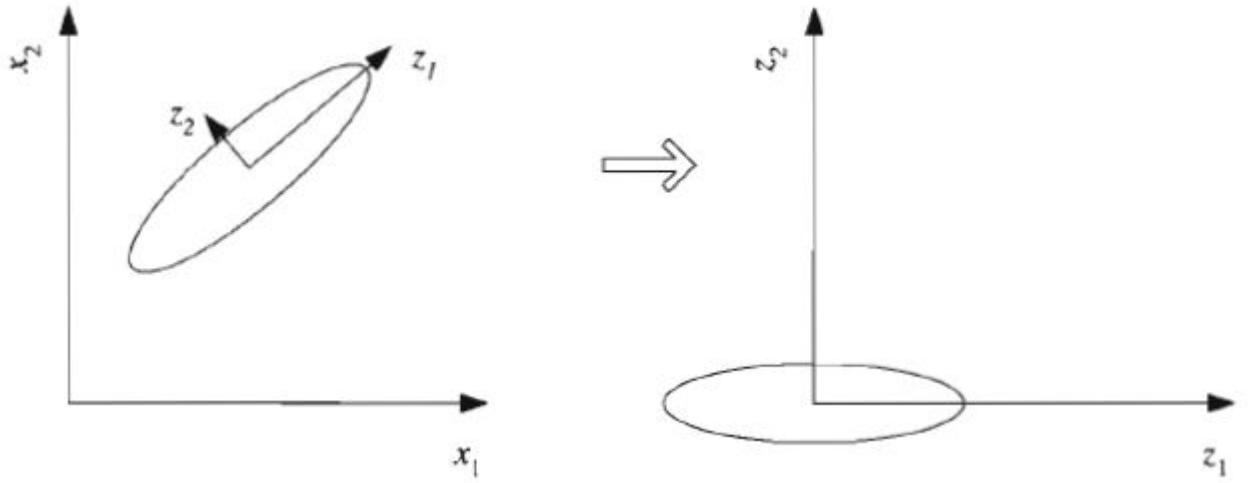

3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

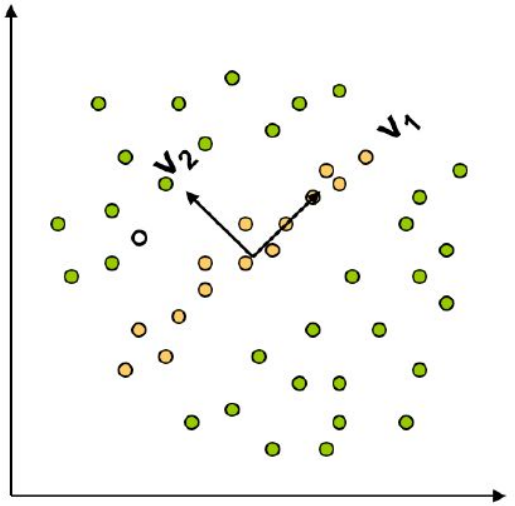
Principle Component Analysis (PCA)

- Maximize the variance

$$\Sigma w_1 = \alpha w_1$$
$$\Sigma w_2 = \alpha w_2$$
$$z = \mathbf{W}^T (x - m)$$

where the k columns of **W** are the leading eigenvectors.
k-dimensional space: dims are the eigenvectors, and the variances over these new dimensions are equal to the eigenvalues.





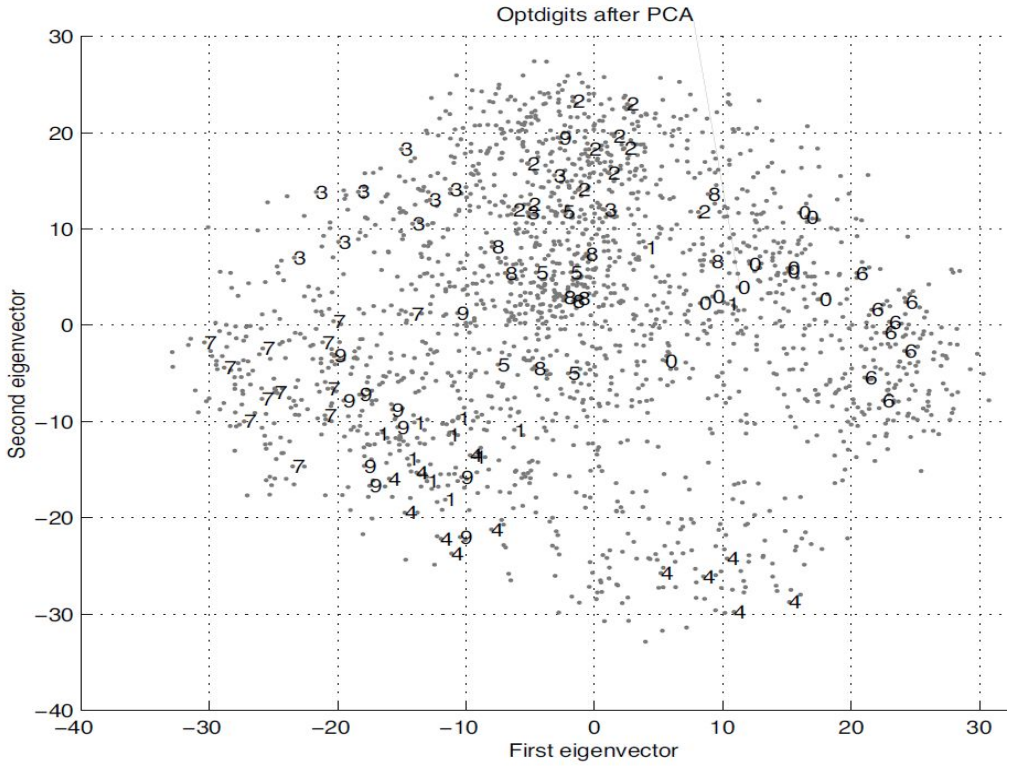
Principle Component Analysis (PCA)

- How many leading eigenvectors? Proportion of variance

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_d}$$

- Reconstruction (and error) $\hat{\mathbf{x}}^t = \mathbf{W}\mathbf{z}^t + \boldsymbol{\mu}$

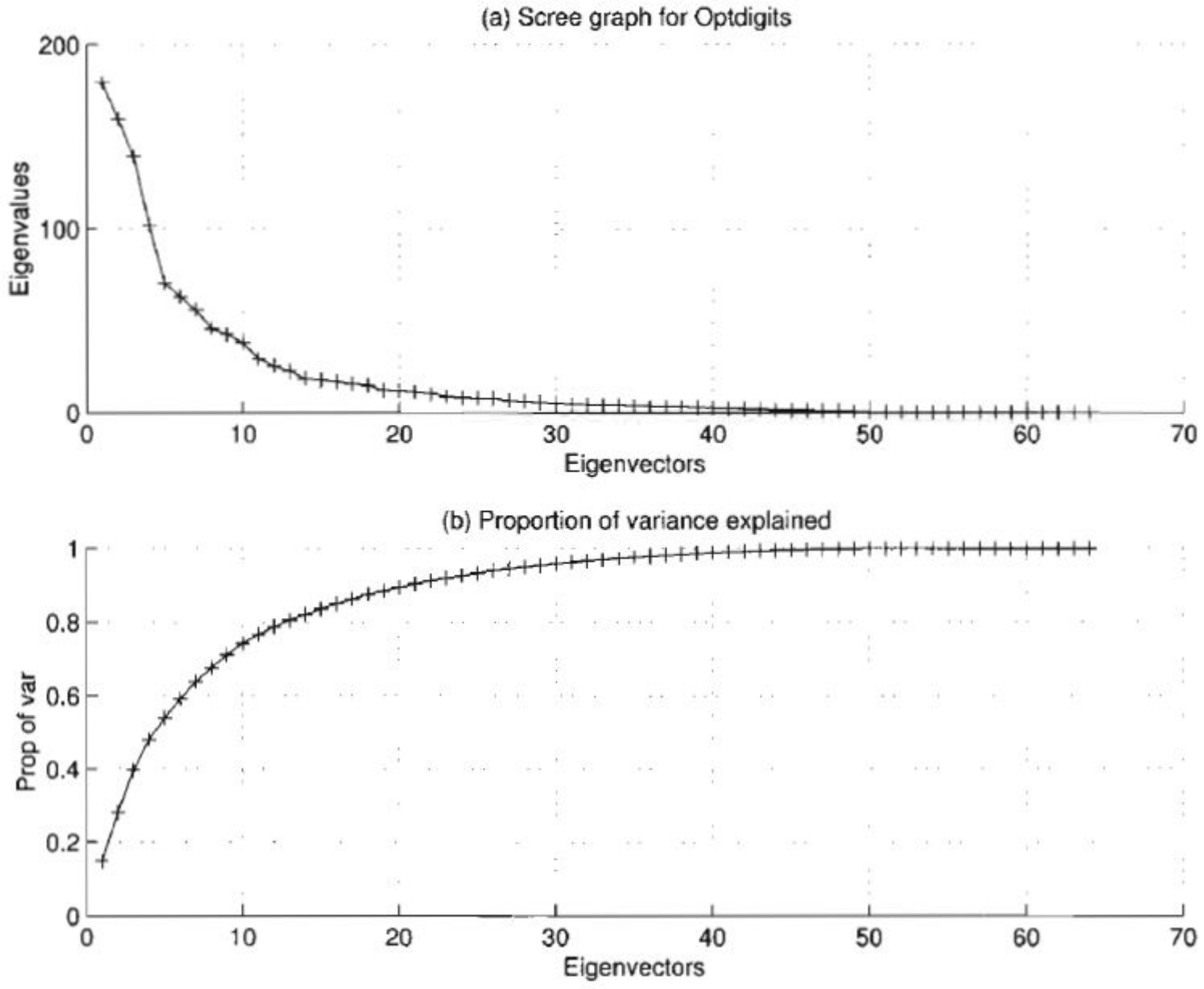
- Visual analysis



Principle Component Analysis (PCA)

- How many leading eigenvectors? Proportion of variance
- Visual analysis $\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_d}$

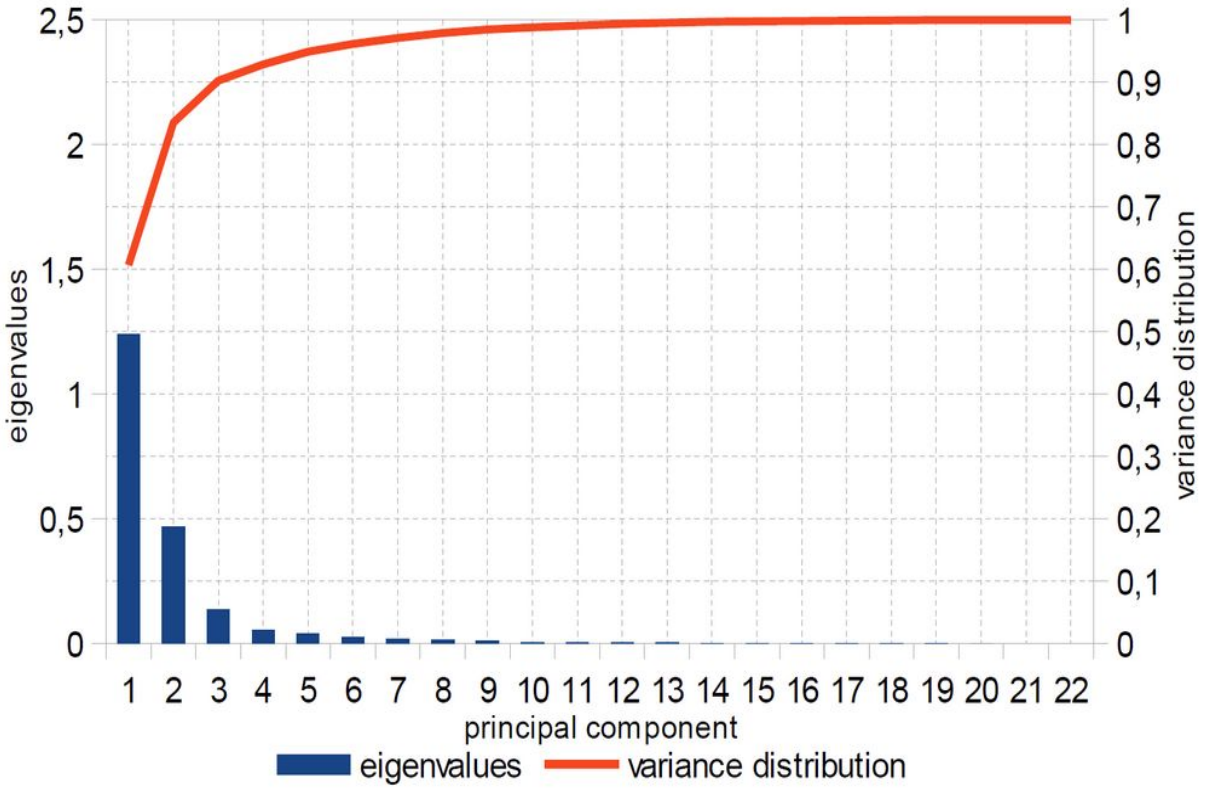
PCA - Scree graph

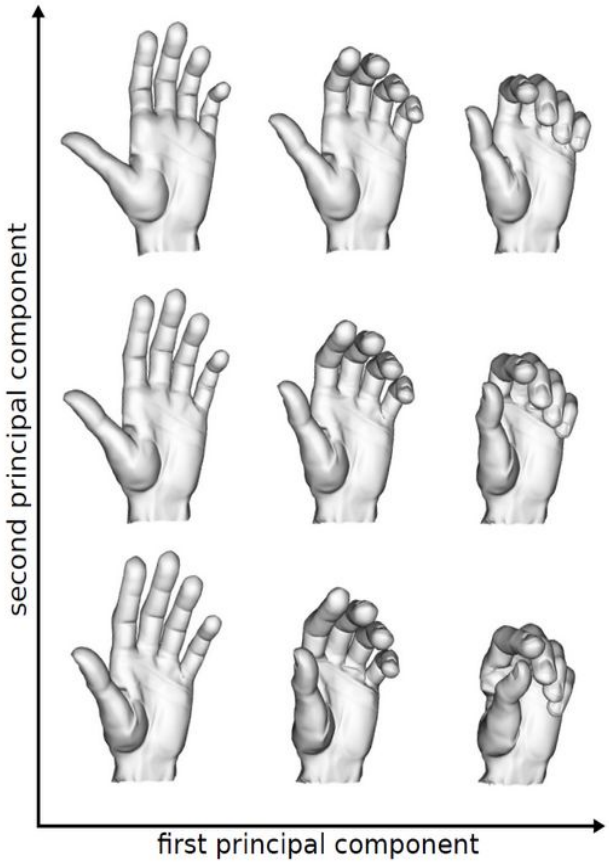


Principle Component Analysis (PCA)

- How many leading eigenvectors? Proportion of variance
- Visual analysis

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_d}$$





Feature embedding

- In PCA: Eigenvectors of $\mathbf{X}^T\mathbf{X}$. dxd matrix
- Feature embedding: Eigenvectors of $\mathbf{X}\mathbf{X}^T$. NxN matrix
  - If d >> N
  - Turk and Pentland 1991, 256x256 images, 40 face images

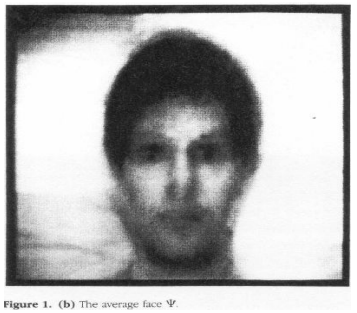


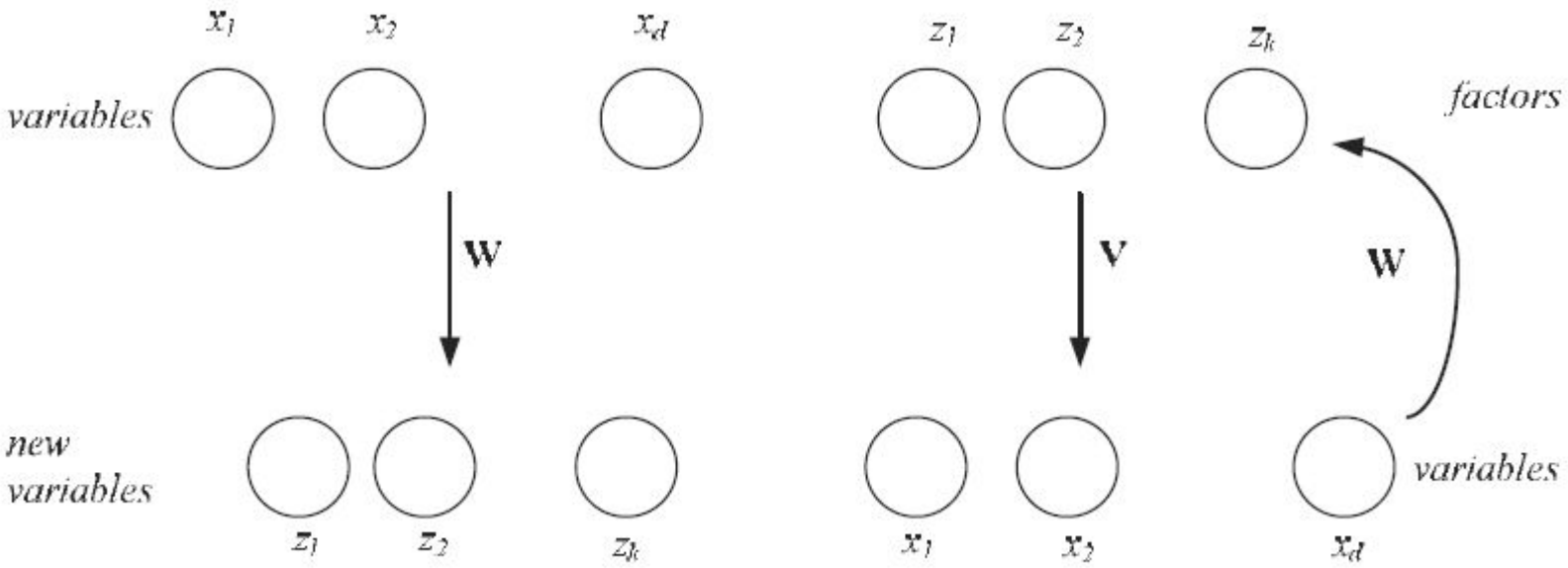
Figure 1. (b) The average face Ψ.

$$\mathbf{u}_l = \sum_{k=1}^{M} \mathbf{v}_{lk}\Phi_k,$$

New image: $\omega_k = \mathbf{u}_k^T(\Gamma - \Psi)$ $\Omega^T = [\omega_1, \omega_2 \ldots \omega_{M'}]$

Factor analysis

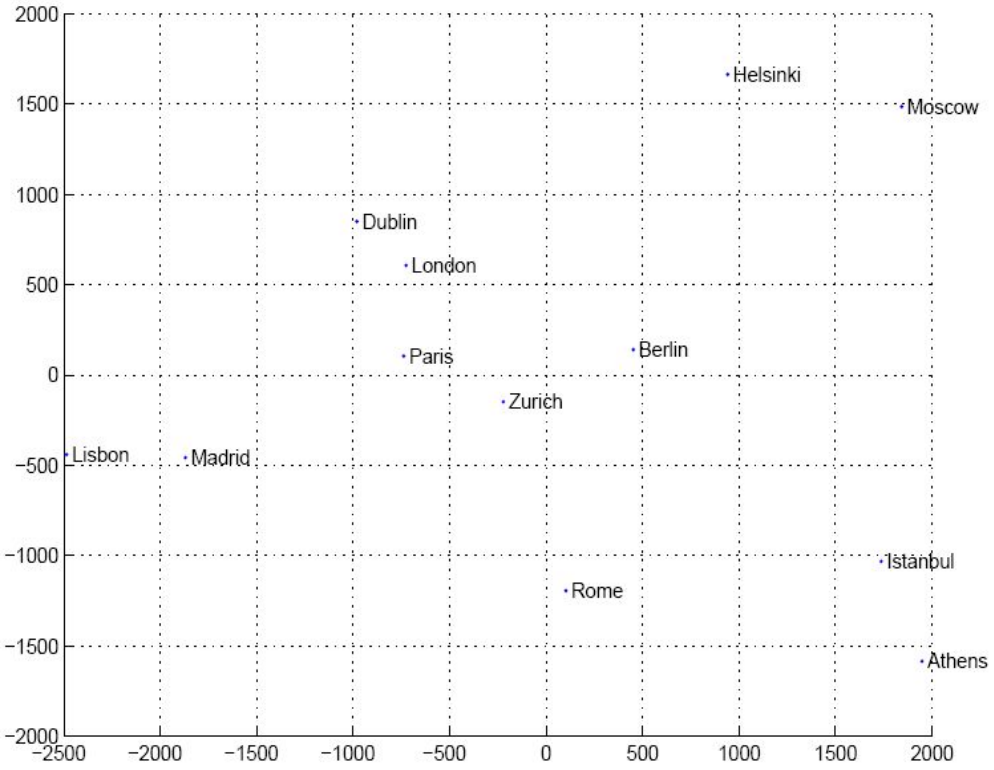
- Assume that there is a set of unobservable variables, *latent factors,* which when acting in combination *generate* ***x****.*
- It assumes that each input dimension, x_i, can be written as a weighted sum of the k < d factors, plus a residual term.



$$x_i - \mu_i = \sum_{j=1}^{k} v_{ij} z_j + \epsilon_i$$

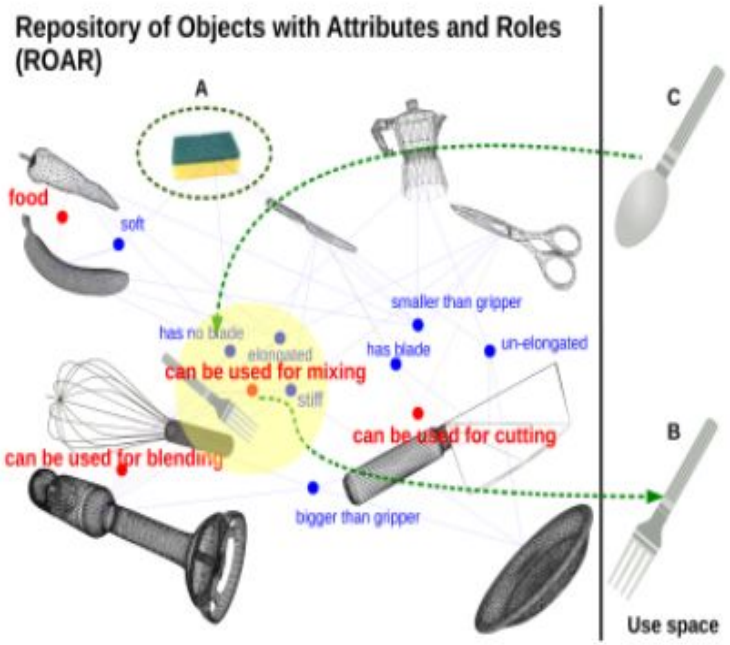
Multidimensional scaling

- Visualizing the level of similarity of individual cases of a dataset
- Preserve d_{ij}, the given distance in the original space.
- PCA on correlation matrix (not covariance) = MDS with Euclidean distances with each variable unit variance



Multidimensional scaling

- Visualizing the level of similarity of individual cases of a dataset
- Preserve d_{ij}, the given distance in the original space.
- PCA on correlation matrix (not covariance) = MDS with Euclidean distances with each variable unit variance



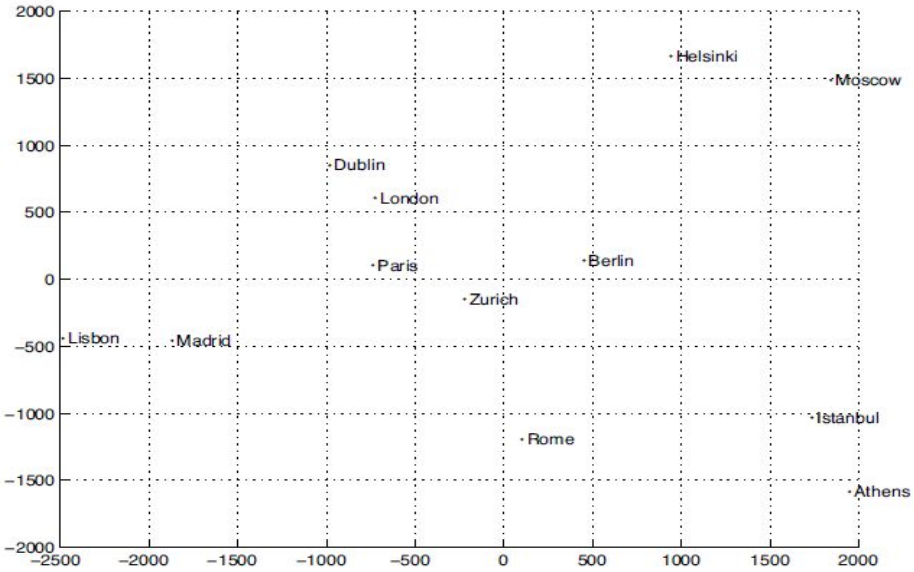
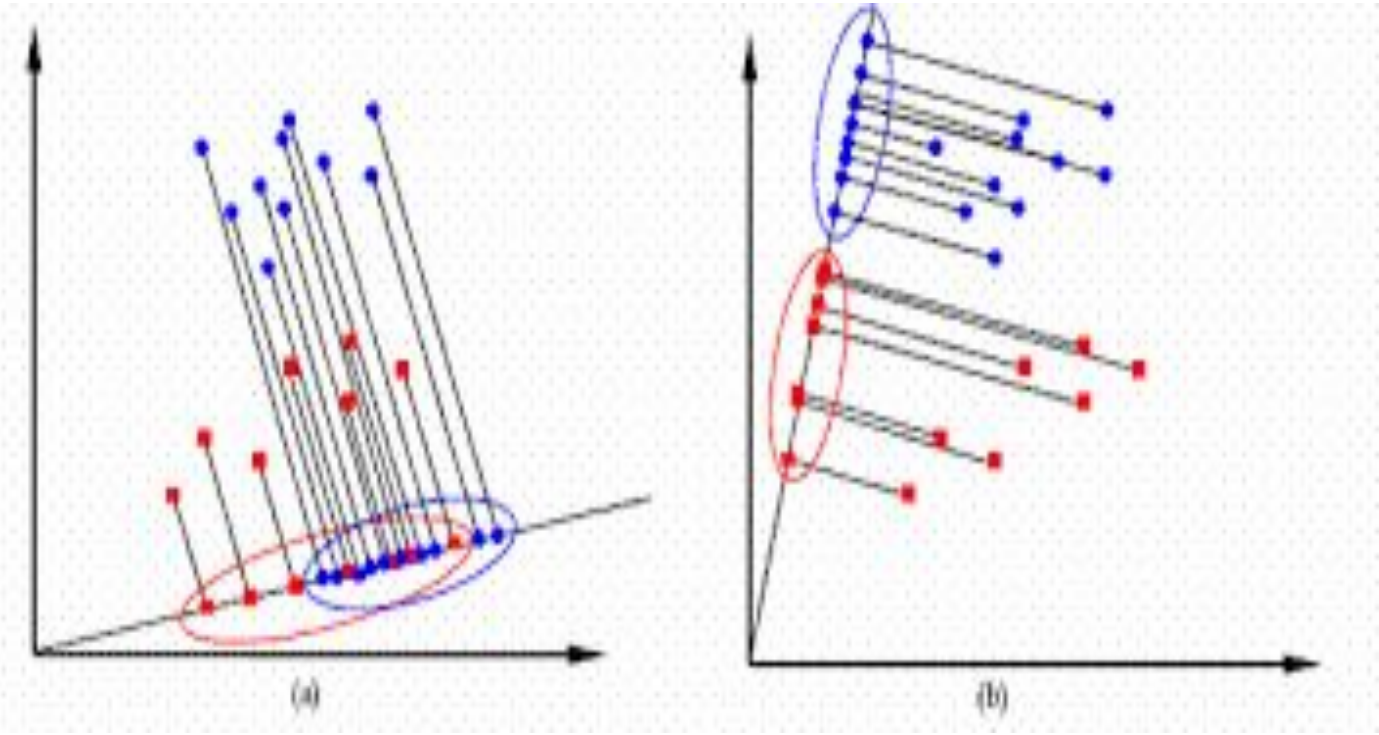


Figure 6.6 Map of Europe drawn by MDS. Pairwise road travel distances between these cities are given as input, and MDS places them in two dimensions such that these distances are preserved as well as possible.

Linear Discriminant Analysis (LDA)



Linear Discriminant Analysis (LDA)

- Supervised method for dimensionality reduction
- Given samples from C_1 and C_2, find the vector **w**, when the data is projected onto **w**, examples of two classes are maximally separated.

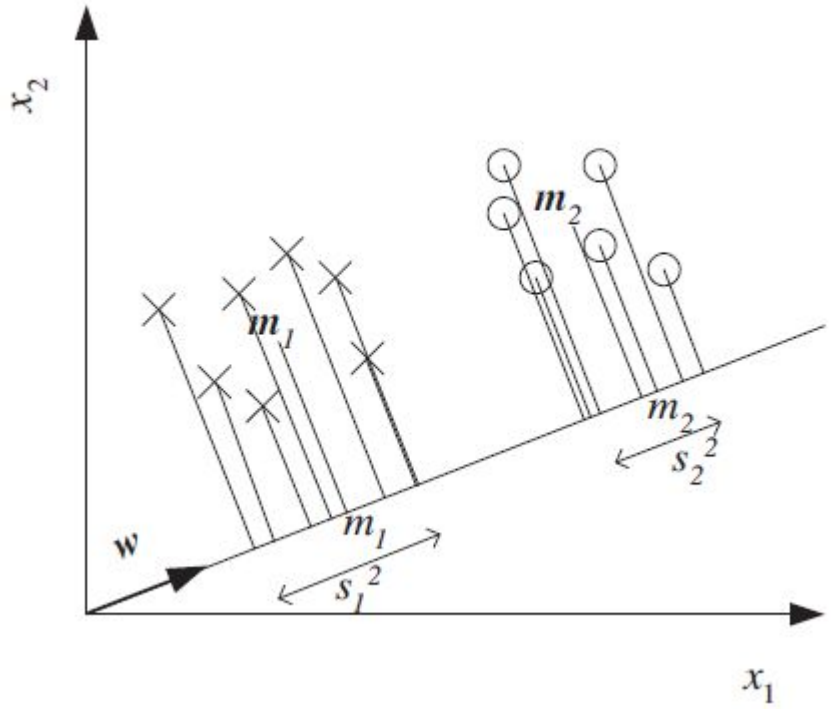
$$z = \mathbf{w}^T \mathbf{x}$$

- Find **w** that maximizes

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

$$m_1 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t r^t}{\sum_t r^t} = \mathbf{w}^T \mathbf{m}_1$$

$$s_1^2 = \sum_t (\mathbf{w}^T \mathbf{x}^t - m_1)^2 r^t$$



Linear Discriminant Analysis (LDA)

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

Take derivative of *J* wrt. **w**:

$$\mathbf{w} = c\mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$$\begin{aligned}(m_1 - m_2)^2 &= (\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2 \\ &= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} \\ &= \mathbf{w}^T\mathbf{S}_B\mathbf{w}\end{aligned}$$

between-class scatter matrix

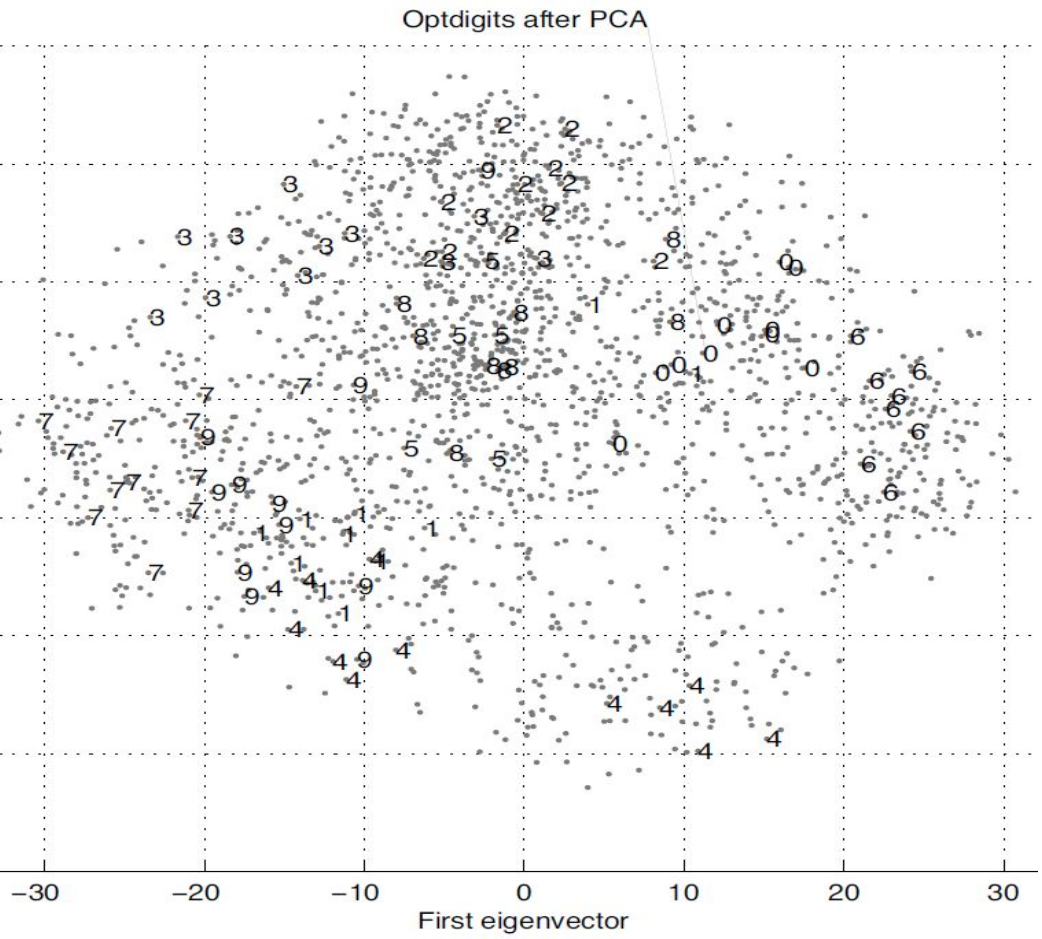
$$\begin{aligned}s_1^2 &= \sum_t (\mathbf{w}^T\mathbf{x}^t - m_1)^2 r^t \\ &= \sum_t \mathbf{w}^T(\mathbf{x}^t - \mathbf{m}_1)(\mathbf{x}^t - \mathbf{m}_1)^T\mathbf{w}r^t \\ &= \mathbf{w}^T\mathbf{S}_1\mathbf{w}\end{aligned}$$

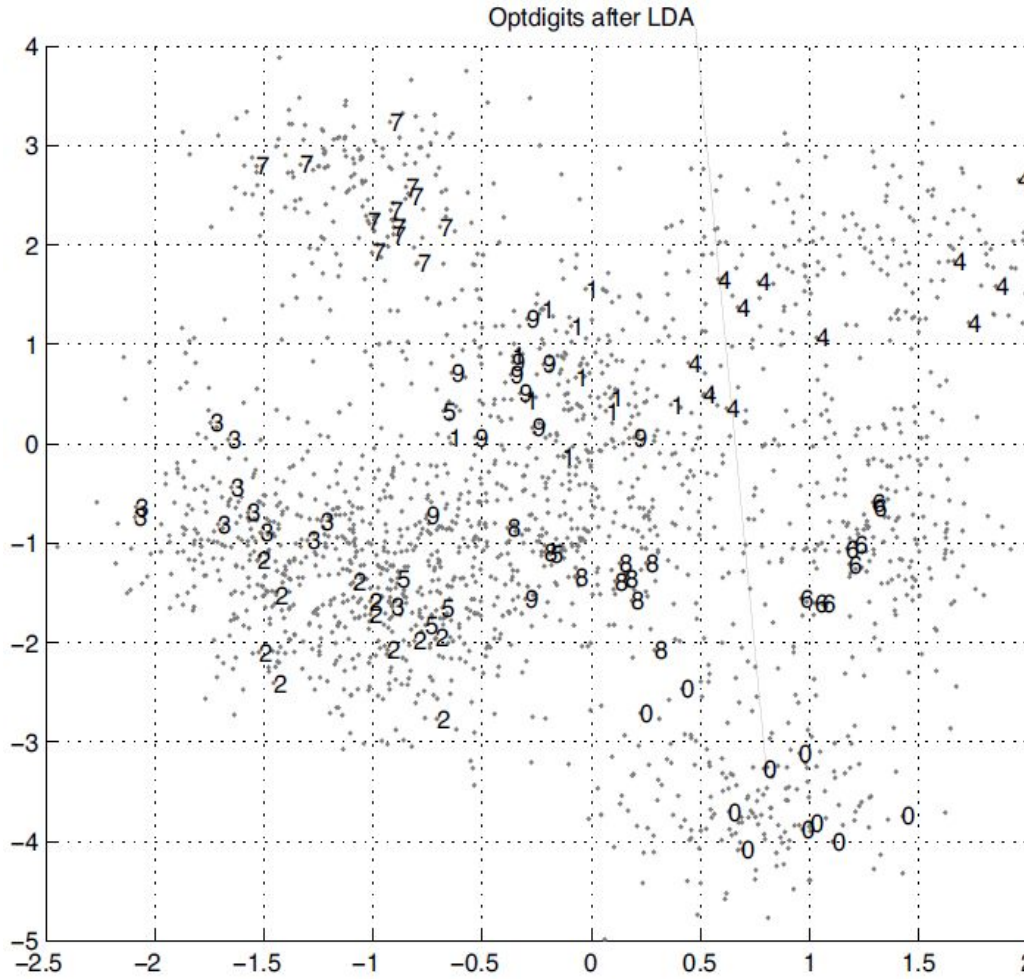
within-class scatter matrix

$$\sum_{i=1}^{C}\sum_{t=1}^{N}(x_t^i - \mu_i)(x_t^i - \mu_i)^T$$

$$\sum_{i=1}^{C} N(\mu_i - \mu)(\mu_i - \mu)^T$$

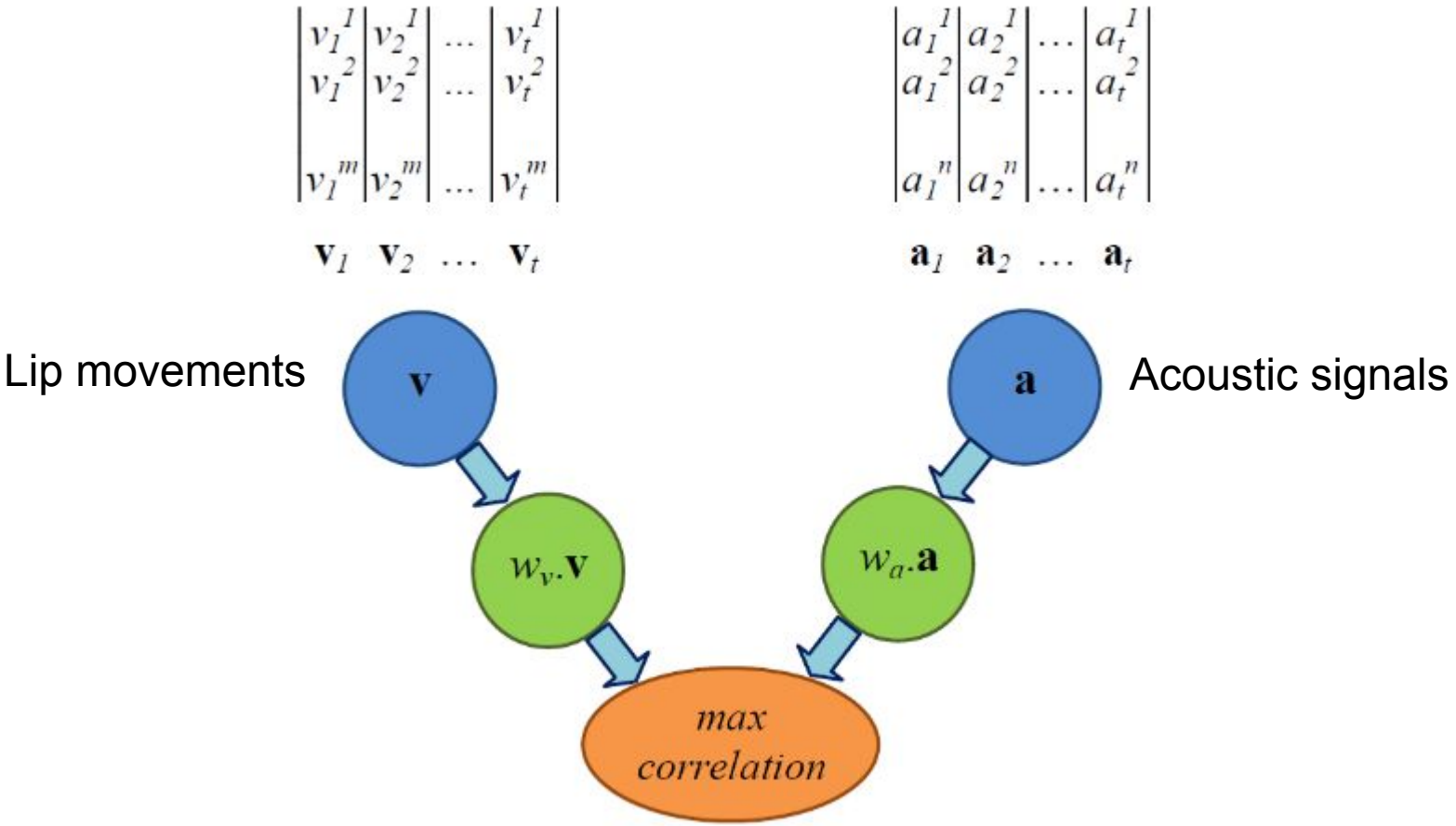
Linear Discriminant Analysis (LDA)





Canonical Correlation Analysis (CCA)

- 2 sets of variables are correlated.
- Reduce dimensionality to a joint space.
- Measure the amount of correlation between **x** and **y** dimensions



Isometric feature mapping (Isomap)

- Similarity between samples may not be simply written in terms of the sum of feature differences
- Estimate geodesic distance