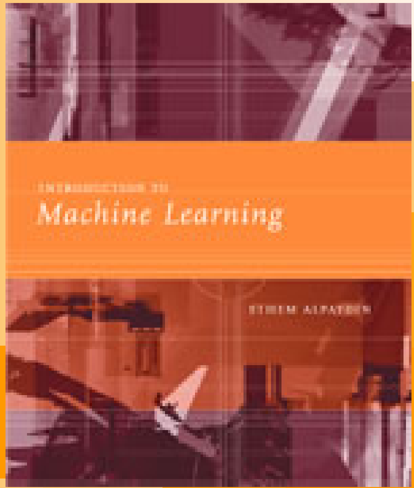


Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml

CHAPTER 6:

Dimensionality Reduction

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), linear discriminant analysis (LDA), factor analysis (FA)

Subset Selection

- There are 2^d 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 $E(F \cup x_j) < E(F)$
- Hill-climbing $O(d^2)$ algorithm
- Backward search: Start with all features and remove one at a time, if possible.
- Floating search (Add k, remove l)

Principal Components Analysis (PCA)

- Find a low-dimensional space such that when $\boldsymbol{x}$ is projected there, information loss is minimized.
- The projection of $\boldsymbol{x}$ on the direction of $\boldsymbol{w}$ is: $z = \boldsymbol{w}^T\boldsymbol{x}$
- Find $\boldsymbol{w}$ such that Var(z) is maximized

$$
\begin{aligned}
\mathrm{Var}(z) = \mathrm{Var}(\boldsymbol{w}^T\boldsymbol{x}) &= E[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})^2] \\
&= E[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})] \\
&= E[\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{w}] \\
&= \boldsymbol{w}^T E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]\boldsymbol{w} = \boldsymbol{w}^T \boldsymbol{\Sigma}\, \boldsymbol{w}
\end{aligned}
$$

where $\mathrm{Var}(\boldsymbol{x}) = E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$

- Maximize Var(z) subject to $||\boldsymbol{w}||=1$

$$\max_{\boldsymbol{w}_1} \boldsymbol{w}_1^T \Sigma \boldsymbol{w}_1 - \alpha\left(\boldsymbol{w}_1^T \boldsymbol{w}_1 - 1\right)$$

$\boldsymbol{\Sigma} \boldsymbol{w}_1 = \alpha \boldsymbol{w}_1$ that is, $\boldsymbol{w}_1$ is an eigenvector of $\boldsymbol{\Sigma}$

Choose the one with the largest eigenvalue for Var(z) to be max

- Second principal component: Max Var(z_2), s.t., $||\boldsymbol{w}_2||=1$ and orthogonal to $\boldsymbol{w}_1$

$$\max_{\boldsymbol{w}_2} \boldsymbol{w}_2^T \Sigma \boldsymbol{w}_2 - \alpha\left(\boldsymbol{w}_2^T \boldsymbol{w}_2 - 1\right) - \beta\left(\boldsymbol{w}_2^T \boldsymbol{w}_1 - 0\right)$$

$\boldsymbol{\Sigma}\ \boldsymbol{w}_2 = \alpha\ \boldsymbol{w}_2$ that is, $\boldsymbol{w}_2$ is another eigenvector of $\boldsymbol{\Sigma}$

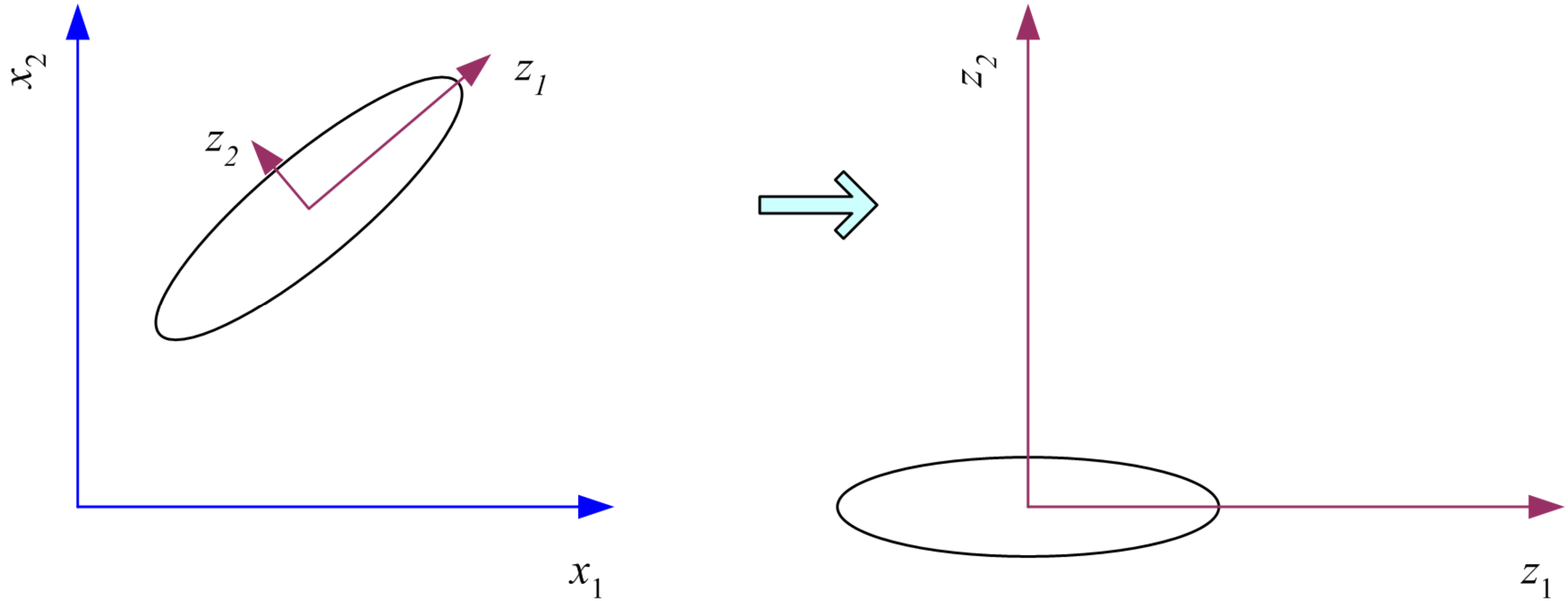
and so on.

What PCA does

$$\boldsymbol{z} = \mathbf{W}^T(\boldsymbol{x} - \boldsymbol{m})$$

where the columns of $\mathbf{W}$ are the eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{m}$ is sample mean

Centers the data at the origin and rotates the axes

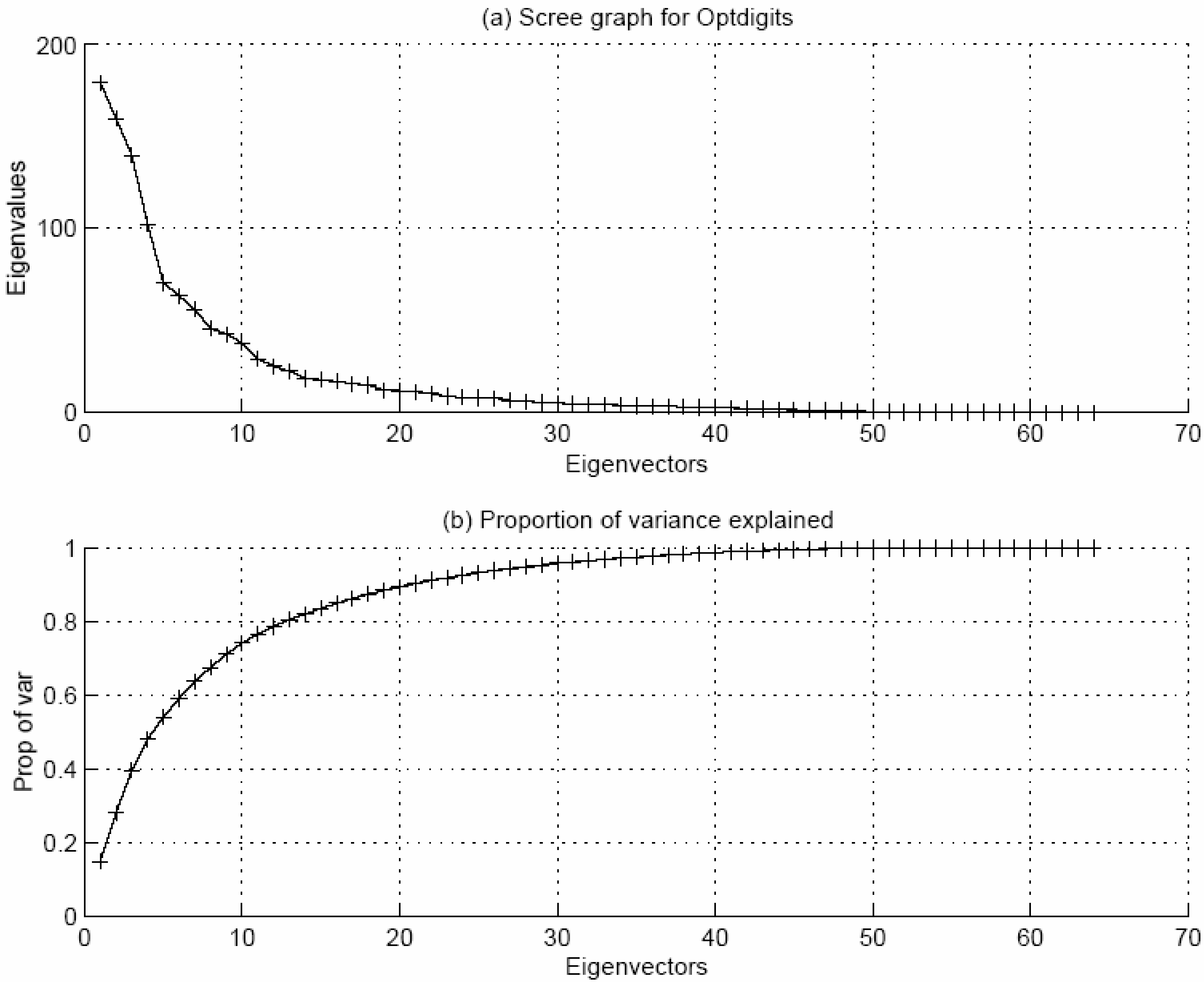


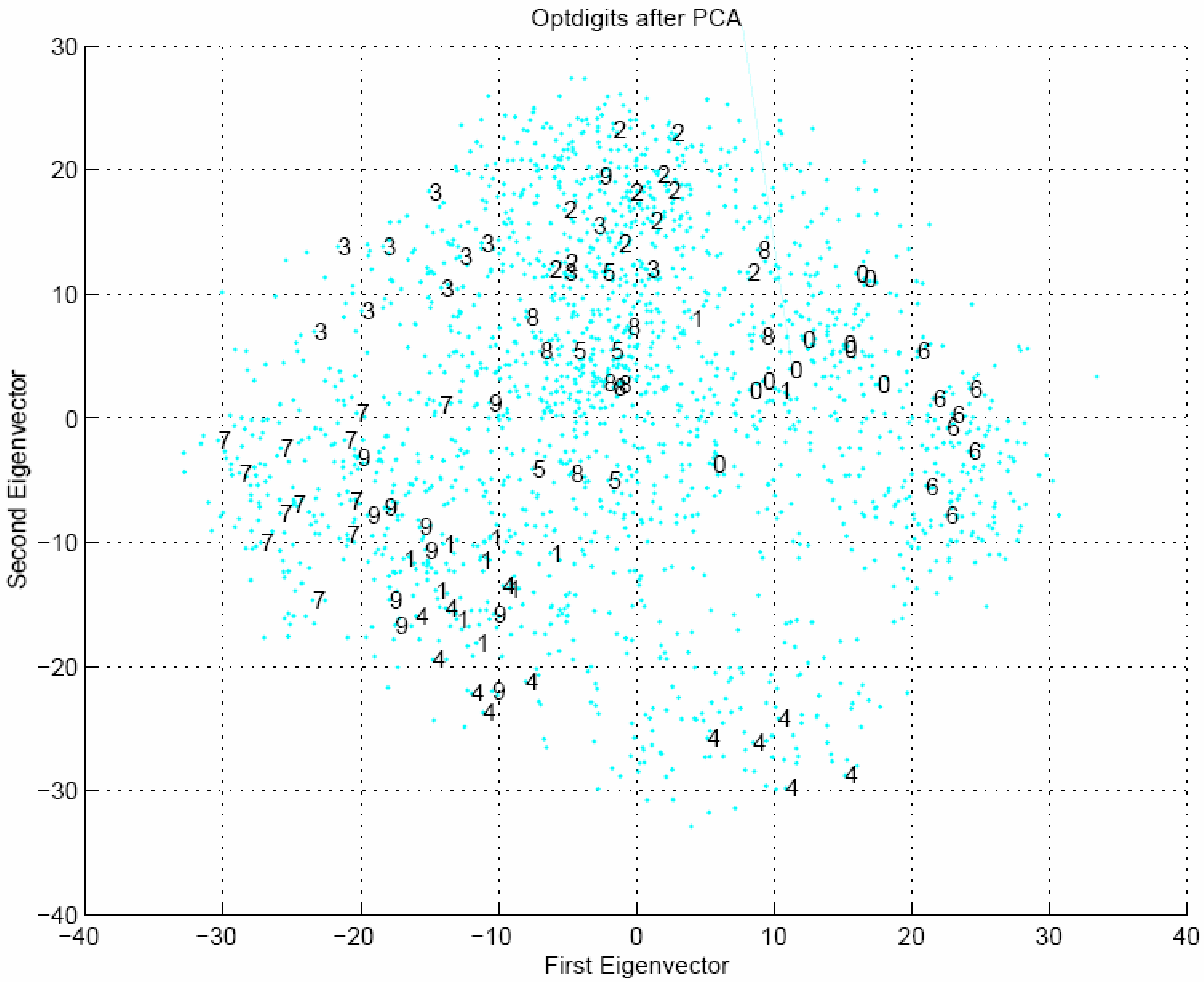
How to choose k ?

- Proportion of Variance (PoV) explained

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

  when λ_i are sorted in descending order
- Typically, stop at PoV>0.9
- Scree graph plots of PoV vs *k*, stop at "elbow"


(a) Scree graph for Optdigits
Eigenvalues
200
100
0
0
10
20
30
40
50
60
70
Eigenvectors
(b) Proportion of variance explained
Prop of var
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
Eigenvectors


Optdigits after PCA
Second Eigenvector
First Eigenvector
30
20
10
0
−10
−20
−30
−40
40

Factor Analysis

- Find a small number of factors $\boldsymbol{z}$, which when combined generate $\boldsymbol{x}$:

  $$x_i - \mu_i = v_{i1}z_1 + v_{i2}z_2 + \ldots + v_{ik}z_k + \varepsilon_i$$

  where z_j, $j = 1,\ldots,k$ are the latent factors with

  $E[z_j] = 0$, $\mathrm{Var}(z_j) = 1$, $\mathrm{Cov}(z_i, z_j) = 0$, $i \neq j$,

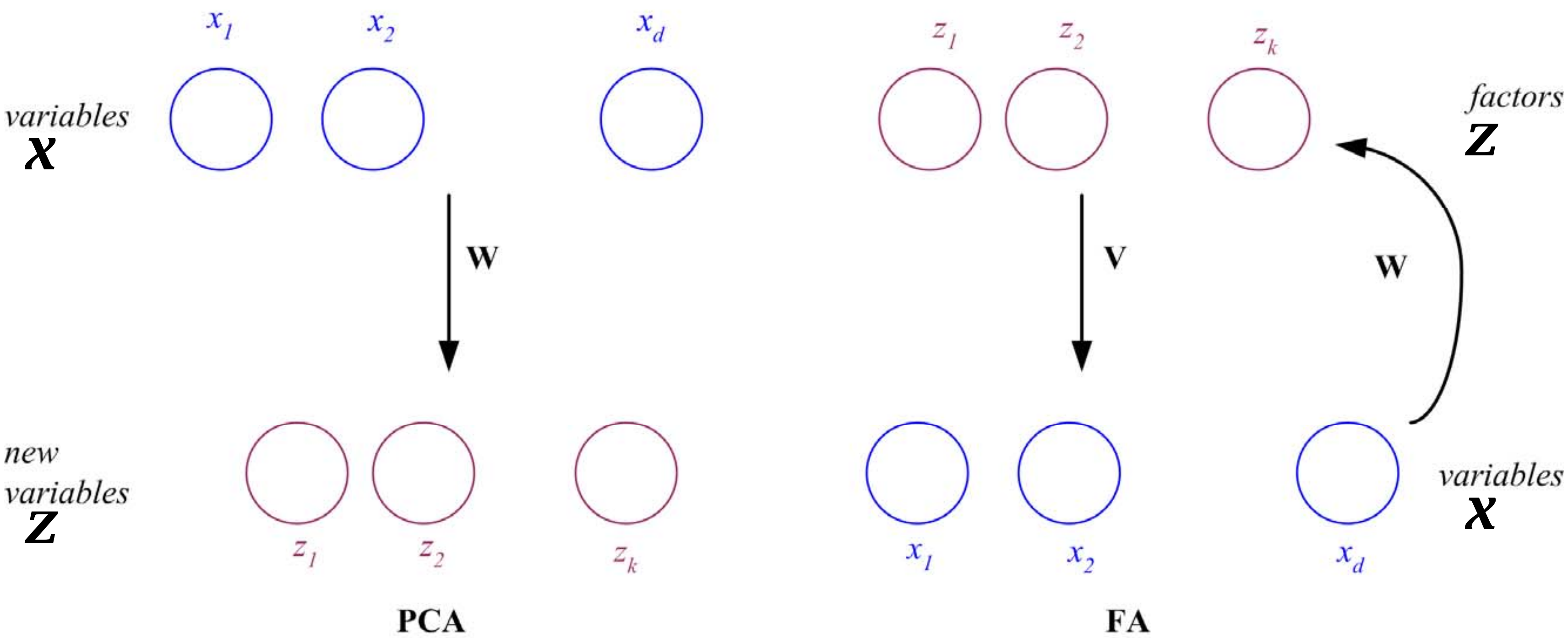
  ε_i are the noise sources

  $E[\varepsilon_i] = \psi_i$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, $\mathrm{Cov}(\varepsilon_i, z_j) = 0$,

  and v_{ij} are the factor loadings

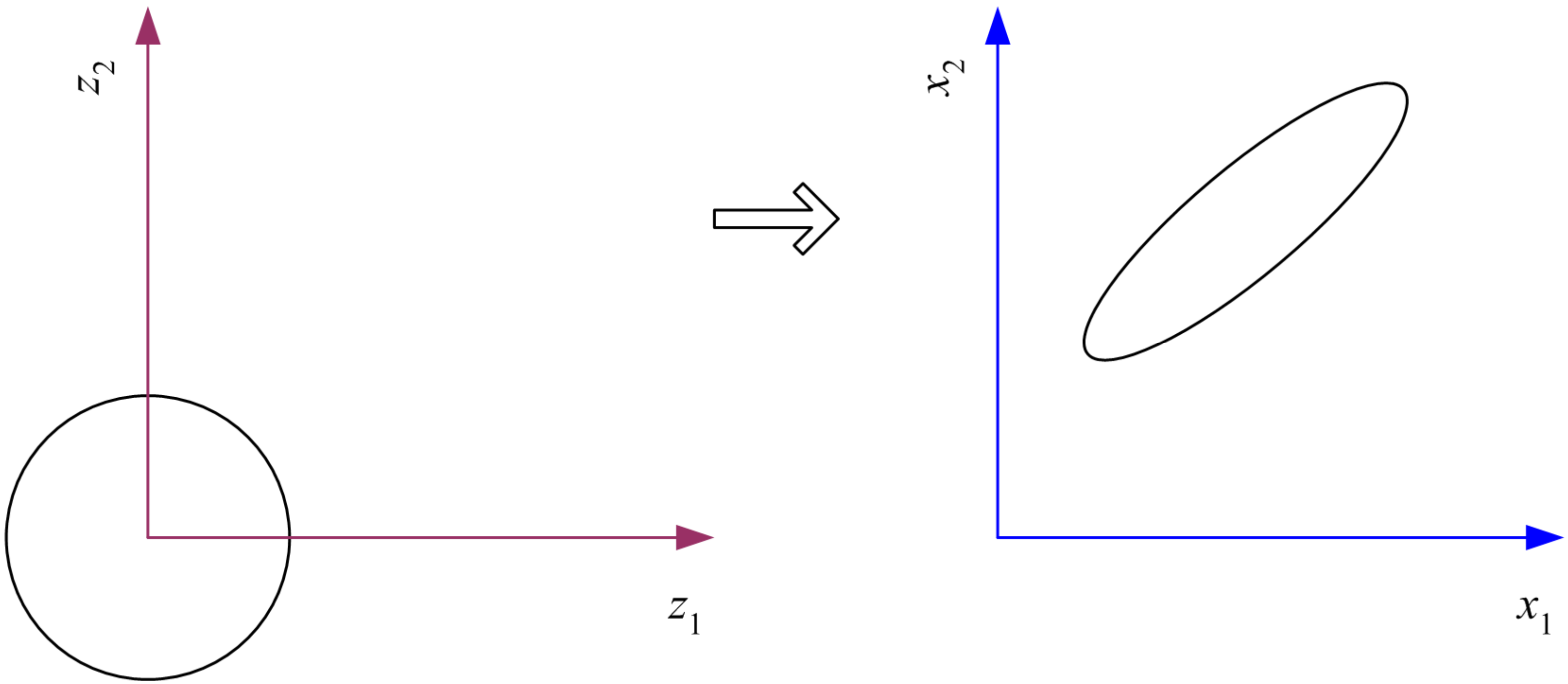
PCA vs FA

- PCA From $\boldsymbol{x}$ to $\boldsymbol{z}$ $\boldsymbol{z} = \mathbf{W}^T(\boldsymbol{x} - \boldsymbol{\mu})$
- FA From $\boldsymbol{z}$ to $\boldsymbol{x}$ $\boldsymbol{x} - \boldsymbol{\mu} = \mathbf{V}\boldsymbol{z} + \varepsilon$



Factor Analysis

- In FA, factors z_j are stretched, rotated and translated to generate $\boldsymbol{x}$



Multidimensional Scaling

- Given pairwise distances between N points,

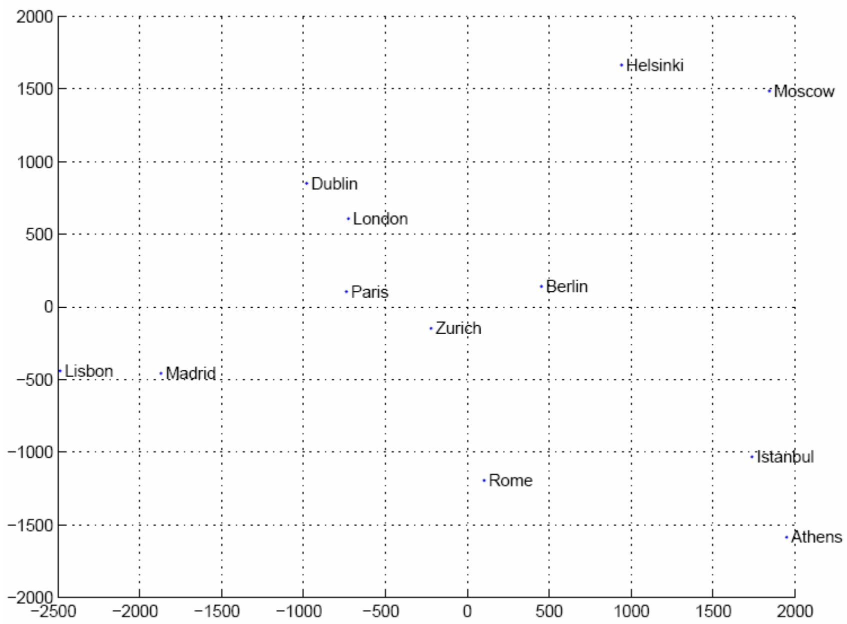
  $d_{ij},\ i,j = 1,\ldots,N$

  place on a low-dim map s.t. distances are preserved.
- $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{x} \mid \theta)$ Find θ that min Sammon stress

$$E(\theta \mid \mathcal{X}) = \sum_{r,s} \frac{\left(\left\| \boldsymbol{z}^r - \boldsymbol{z}^s \right\| - \left\| \boldsymbol{x}^r - \boldsymbol{x}^s \right\| \right)^2}{\left\| \boldsymbol{x}^r - \boldsymbol{x}^s \right\|^2}$$

$$= \sum_{r,s} \frac{\left(\left\| \boldsymbol{g}(\boldsymbol{x}^r \mid \theta) - \boldsymbol{g}(\boldsymbol{x}^s \mid \theta) \right\| - \left\| \boldsymbol{x}^r - \boldsymbol{x}^s \right\| \right)^2}{\left\| \boldsymbol{x}^r - \boldsymbol{x}^s \right\|^2}$$

Map of Europe by MDS

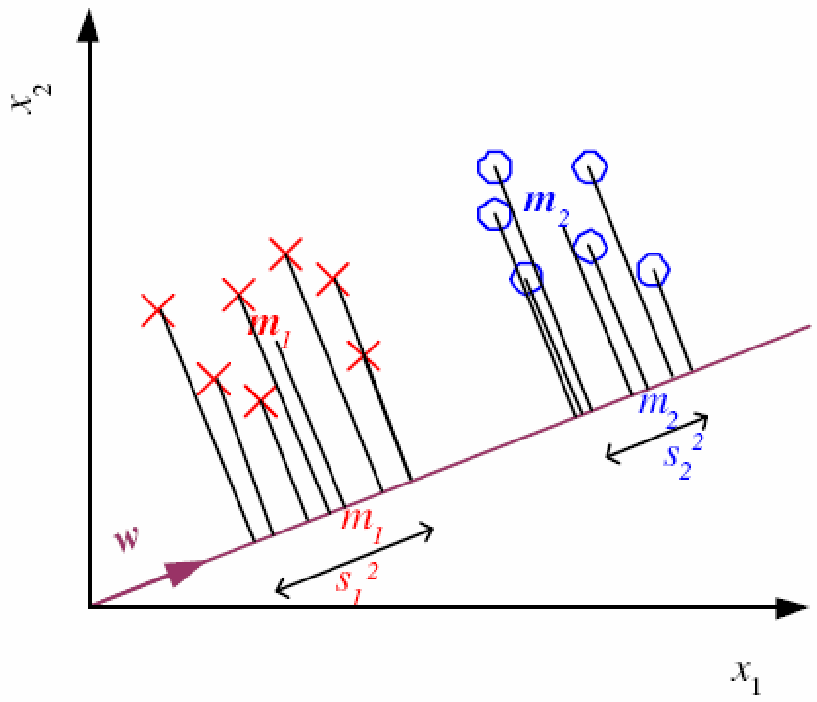


Map from CIA – The World Factbook: http://www.cia.gov/

Linear Discriminant Analysis

- Find a low-dimensional space such that when $\boldsymbol{x}$ is projected, classes are well-separated.
- Find $\boldsymbol{w}$ that maximizes

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



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

- Between-class scatter:

$$(m_1 - m_2)^2 = (\boldsymbol{w}^T\boldsymbol{m}_1 - \boldsymbol{w}^T\boldsymbol{m}_2)^2$$
$$= \boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T\boldsymbol{w}$$
$$= \boldsymbol{w}^T\mathbf{S}_B\boldsymbol{w} \text{ where } \mathbf{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

- Within-class scatter:

$$s_1^2 = \sum_t (\boldsymbol{w}^T\boldsymbol{x}^t - m_1)^2 r^t$$
$$= \sum_t \boldsymbol{w}^T(\boldsymbol{x}^t - \boldsymbol{m}_1)(\boldsymbol{x}^t - \boldsymbol{m}_1)^T\boldsymbol{w}r^t = \boldsymbol{w}^T\mathbf{S}_1\boldsymbol{w}$$
$$\text{where } \mathbf{S}_1 = \sum_t (\boldsymbol{x}^t - \boldsymbol{m}_1)(\boldsymbol{x}^t - \boldsymbol{m}_1)^T r^t$$
$$s_1^2 + s_1^2 = \boldsymbol{w}^T\mathbf{S}_W\boldsymbol{w} \text{ where } \mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

Fisher's Linear Discriminant

- Find $\boldsymbol{w}$ that max

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}} = \frac{\left|\boldsymbol{w}^T (\boldsymbol{m}_1 - \boldsymbol{m}_2)\right|^2}{\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}}$$

- LDA soln:

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

- Parametric soln:

$$\boldsymbol{w} = \Sigma^{-1} (\mu_1 - \mu_2)$$

$$\text{when } p(\boldsymbol{x} \mid C_i) \sim \mathcal{N}(\mu_i, \Sigma)$$

K>2 Classes

- Within-class scatter:

$$\mathbf{S}_W = \sum_{i=1}^{K} \mathbf{S}_i \qquad \mathbf{S}_i = \sum_t r_i^t \left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)\left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)^T$$

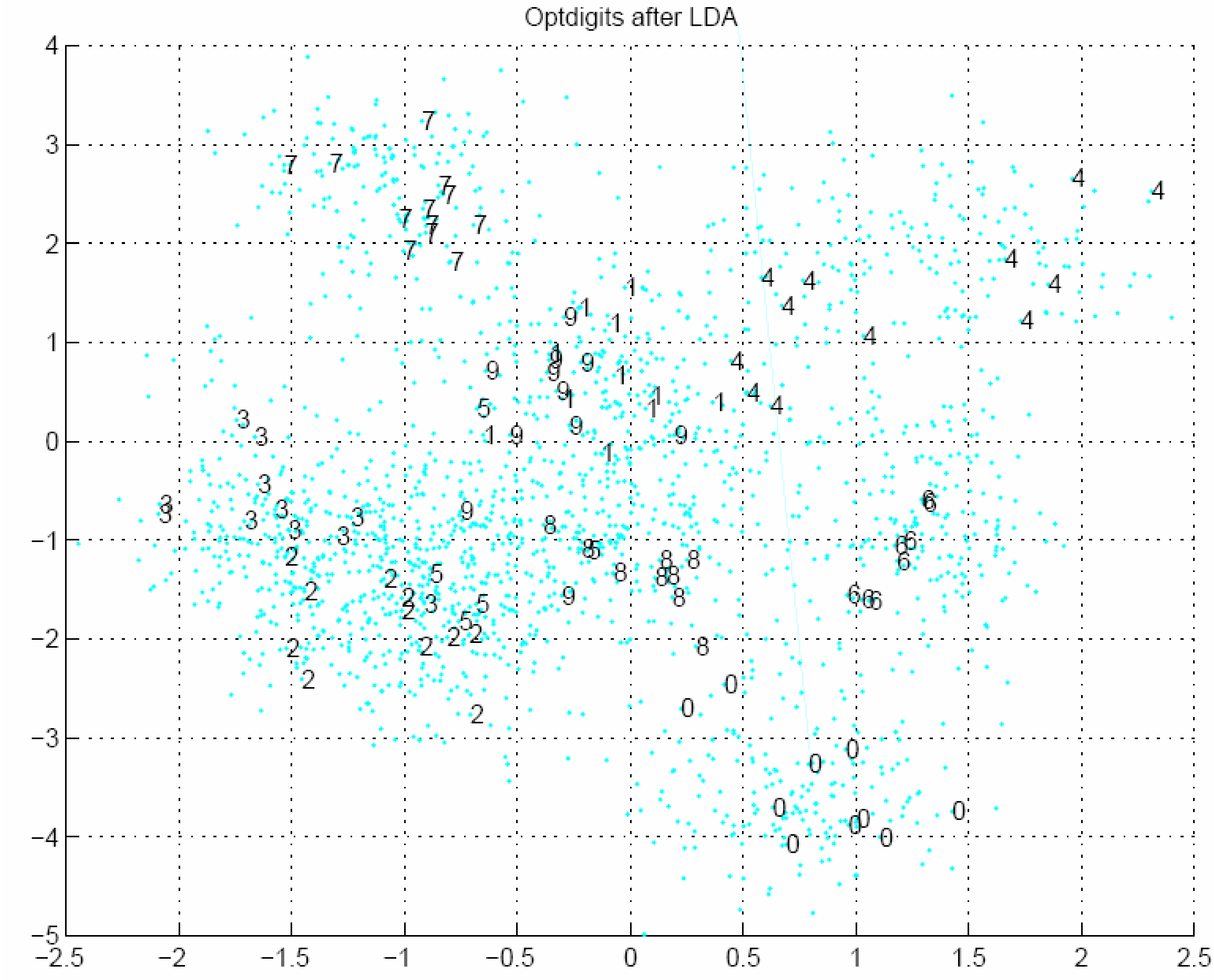
- Between-class scatter:

$$\mathbf{S}_B = \sum_{i=1}^{K} N_i \left(\boldsymbol{m}_i - \boldsymbol{m}\right)\left(\boldsymbol{m}_i - \boldsymbol{m}\right)^T \qquad \boldsymbol{m} = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{m}_i$$

- Find **W** that max

$$J(\mathbf{W}) = \frac{\left|\mathbf{W}^T \mathbf{S}_B \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_W \mathbf{W}\right|}$$

The largest eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$
Maximum rank of K-1


Optdigits after LDA