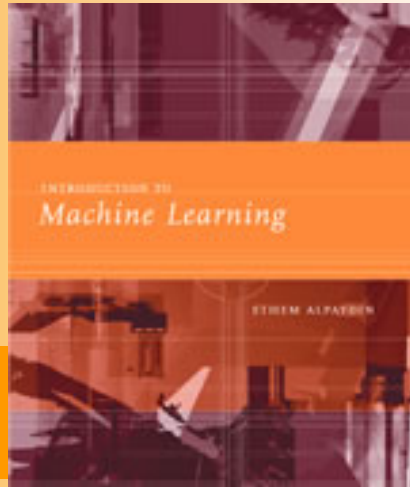




Lecture Slides for

INTRODUCTION TO

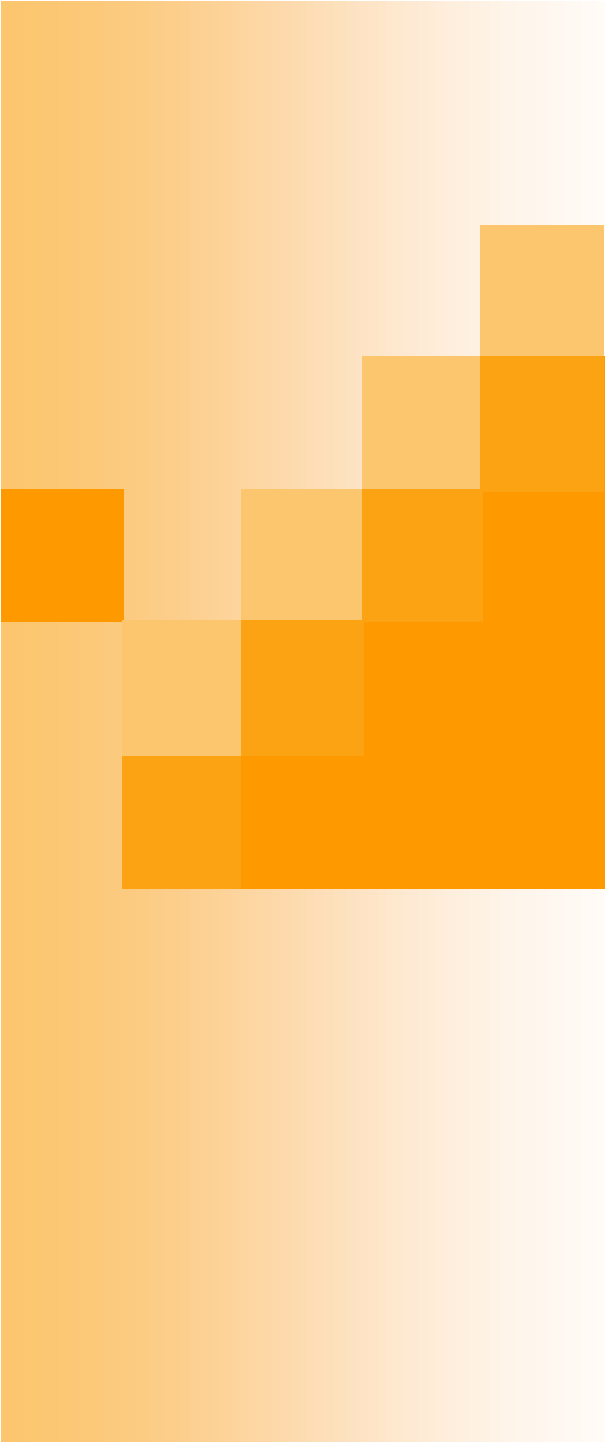
Machine Learning

ETHEM ALPAYDIN

© The MIT Press, 2004

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 , $i = 1, \dots, d$ dimensions to new $k < d$ dimensions, z_j , $j = 1, \dots, k$

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 $\emptyset$.
 - At each iteration, find the best new feature
$$j = \operatorname{argmin}_i E (F \cup x_i)$$
 - 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 $\mathbf{x}$ is projected there, information loss is minimized.
- The projection of $\mathbf{x}$ on the direction of $\mathbf{w}$ is: $z = \mathbf{w}^T \mathbf{x}$
- Find $\mathbf{w}$ such that $\text{var}(z)$ is maximized

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

where $\text{var}(\mathbf{x}) = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$

- 
- Maximize $\text{var}(z)$ subject to $\|\mathbf{w}\|=1$

$$\max_{\mathbf{w}_1} \mathbf{w}_1^T \Sigma \mathbf{w}_1 - \alpha (\mathbf{w}_1^T \mathbf{w}_1 - 1)$$

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

Choose the one with the largest eigenvalue for $\text{var}(z)$ to be max

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

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

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

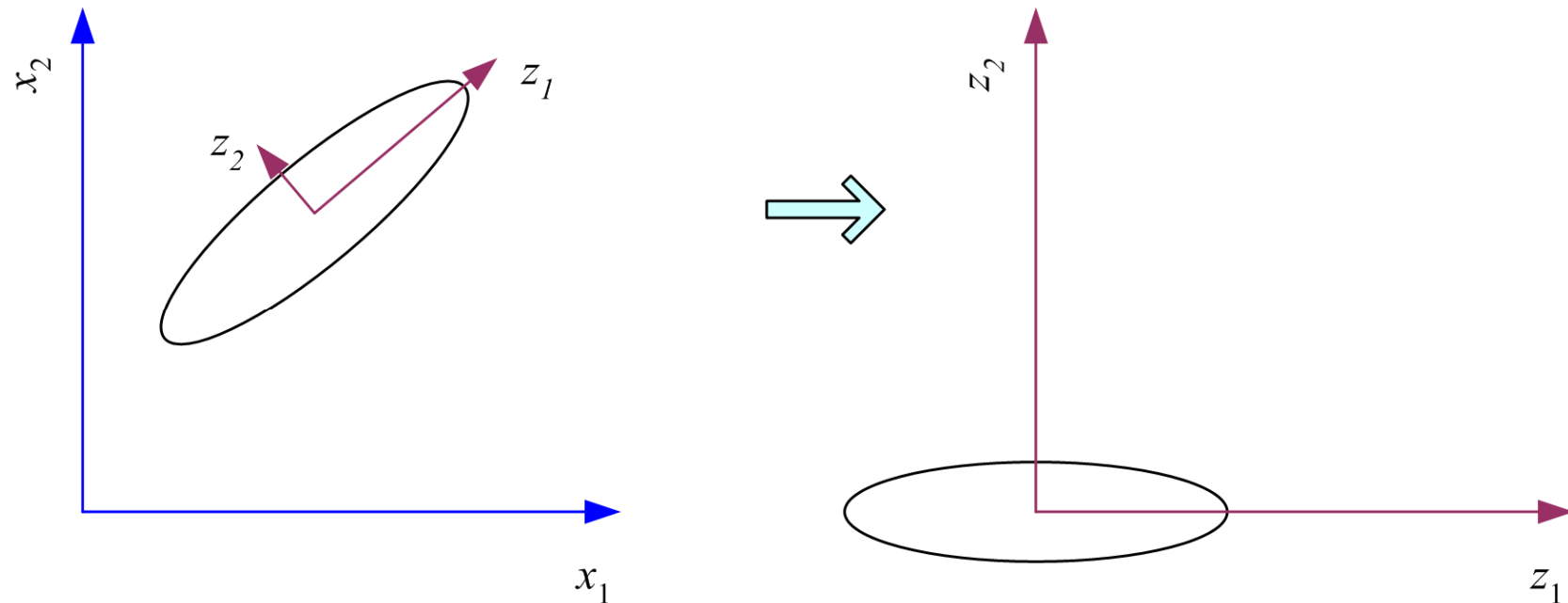
- and so on.

What PCA does

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

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

Centers the data at the origin and rotates the axes





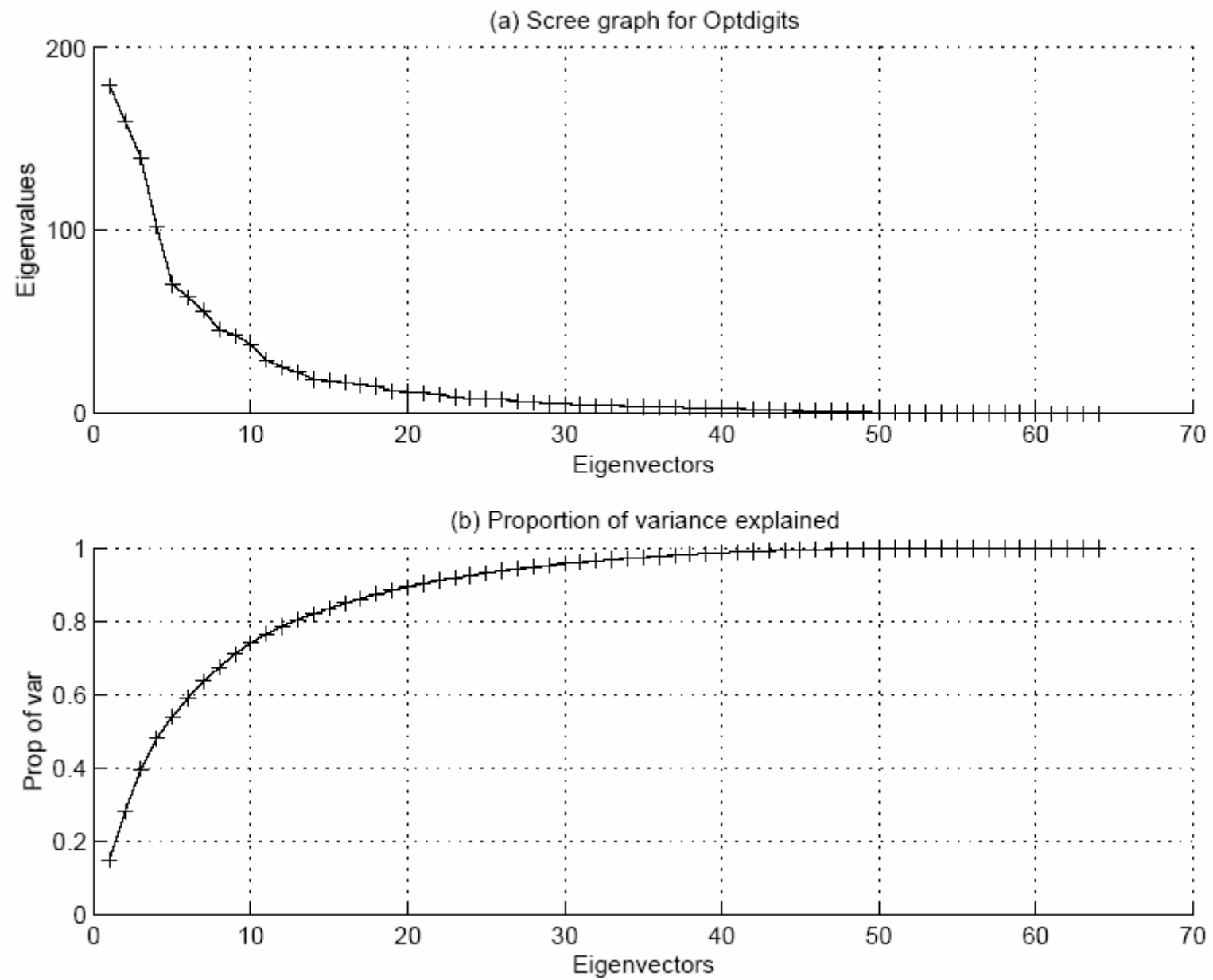
How to choose k ?

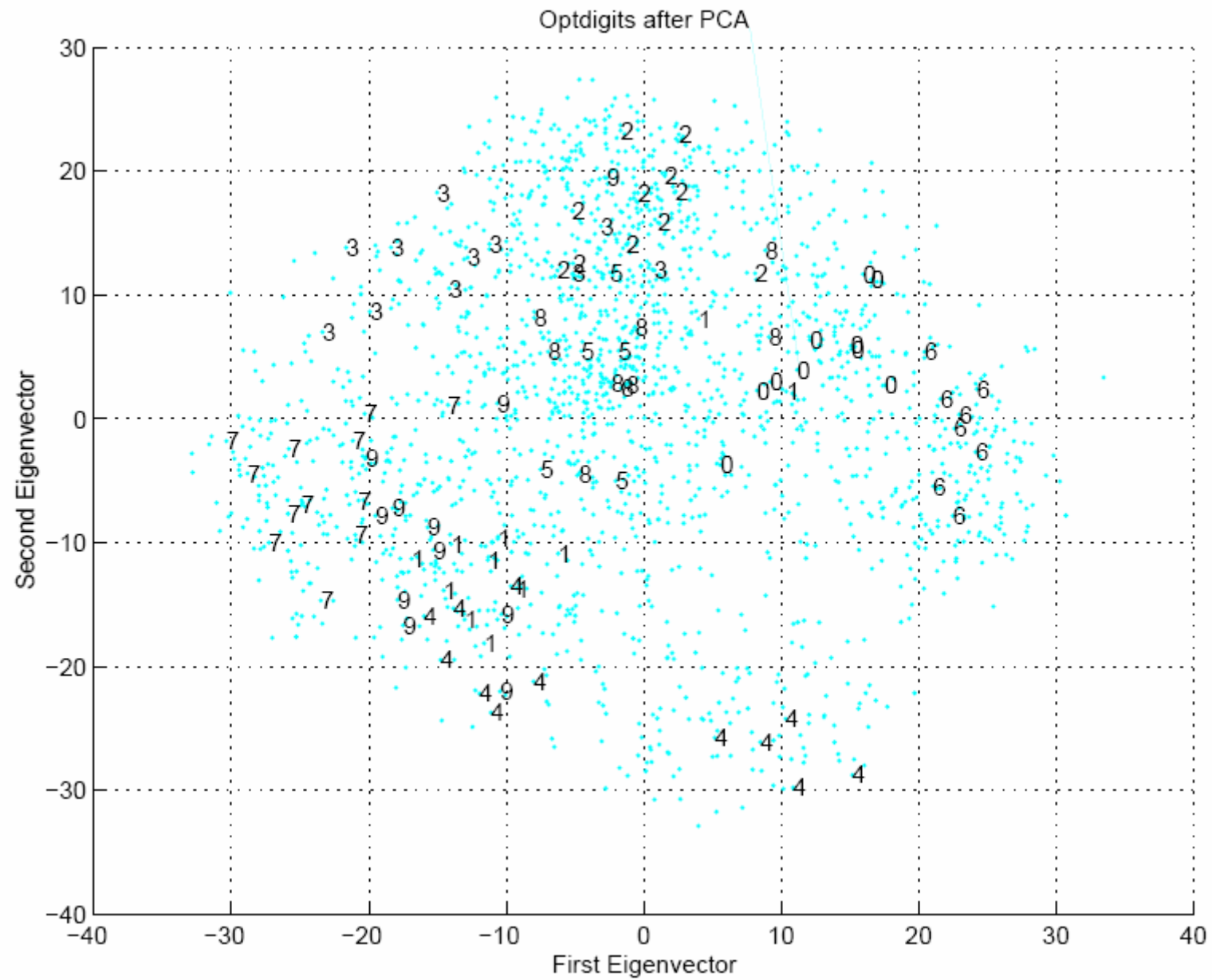
- Proportion of variance (PoV) explained

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

when λ_i are sorted in descending order

- Typically, stop at PoV>0.9
- Scree graph plots of PoV vs k , stop at “elbow”







Factor Analysis

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

$$x_i - \mu_i = v_{i1}Z_1 + v_{i2}Z_2 + \dots + v_{ik}Z_k + \varepsilon_i$$

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

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

ε_i are the **noise sources**

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

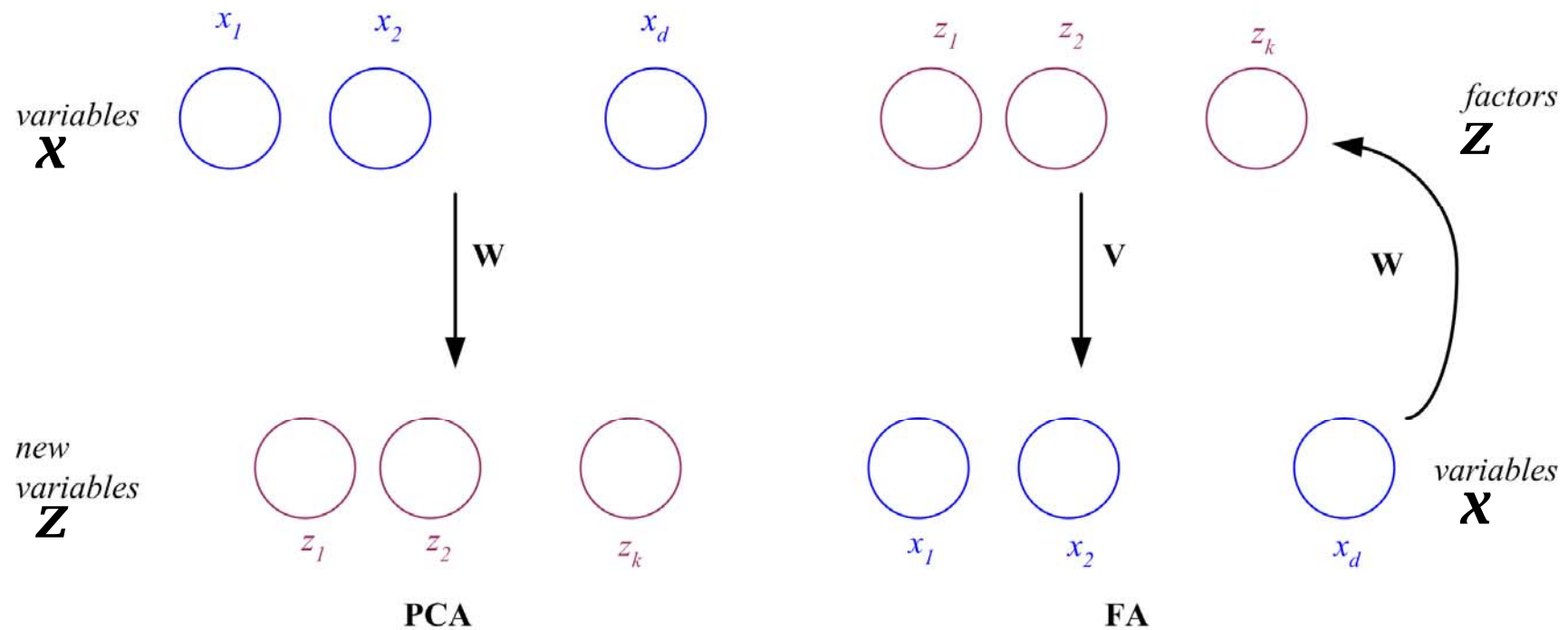
and v_{ij} are the **factor loadings**

PCA vs FA

- PCA From $\mathbf{x}$ to $\mathbf{z}$
- FA From $\mathbf{z}$ to $\mathbf{x}$

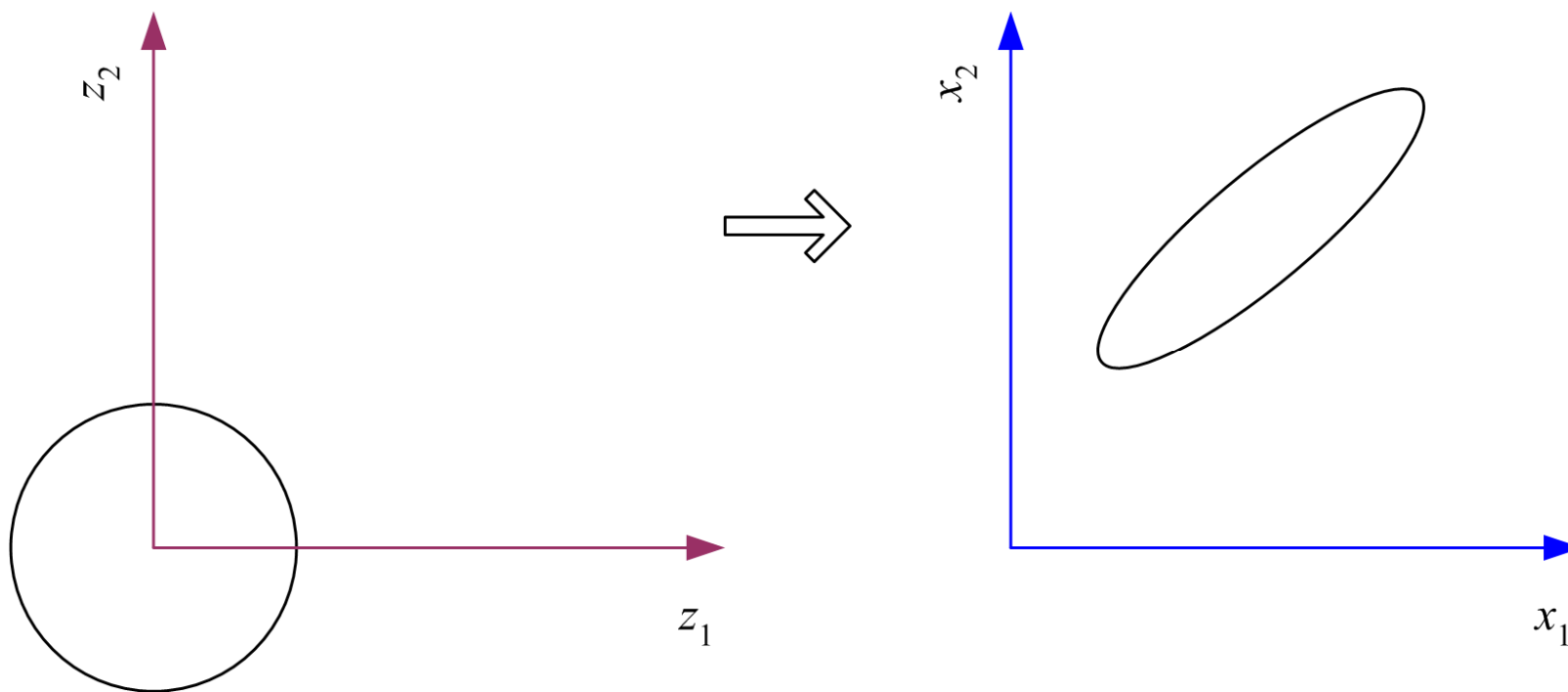
$$\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu})$$

$$\mathbf{x} - \boldsymbol{\mu} = \mathbf{V}\mathbf{z} + \boldsymbol{\varepsilon}$$



Factor Analysis

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



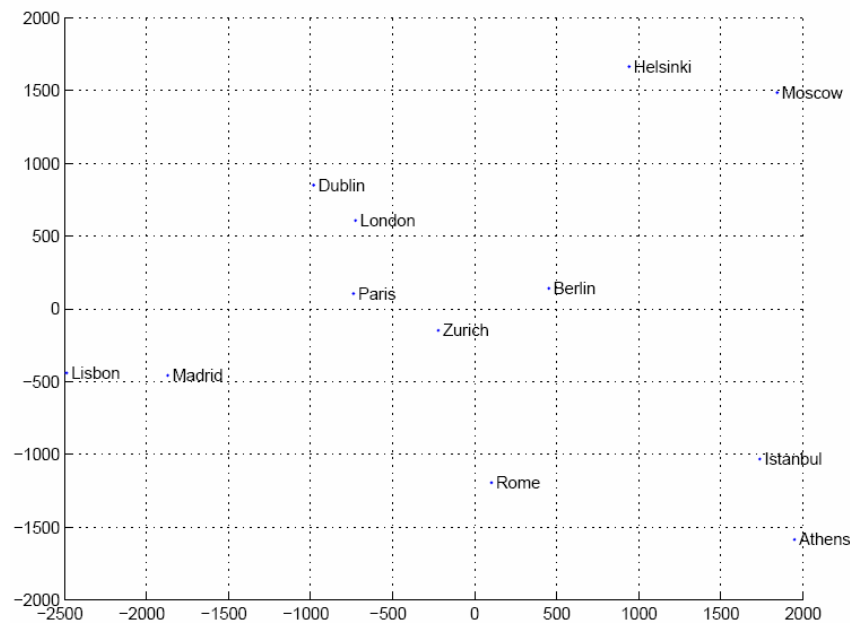


Multidimensional Scaling

- Given pairwise distances between N points,
 $d_{ij}, i, j = 1, \dots, N$
place them on a low-dimensional map such that the distances are preserved.
- $\mathbf{z} = \mathbf{g}(\mathbf{x} | \theta)$ Find θ that min **Sammon stress**

$$\begin{aligned} E(\theta | \mathcal{X}) &= \sum_{r,s} \frac{(\|\mathbf{z}^r - \mathbf{z}^s\| - \|\mathbf{x}^r - \mathbf{x}^s\|)^2}{\|\mathbf{x}^r - \mathbf{x}^s\|^2} \\ &= \sum_{r,s} \frac{(\|\mathbf{g}(\mathbf{x}^r | \theta) - \mathbf{g}(\mathbf{x}^s | \theta)\| - \|\mathbf{x}^r - \mathbf{x}^s\|)^2}{\|\mathbf{x}^r - \mathbf{x}^s\|^2} \end{aligned}$$

Map of Europe by MDS



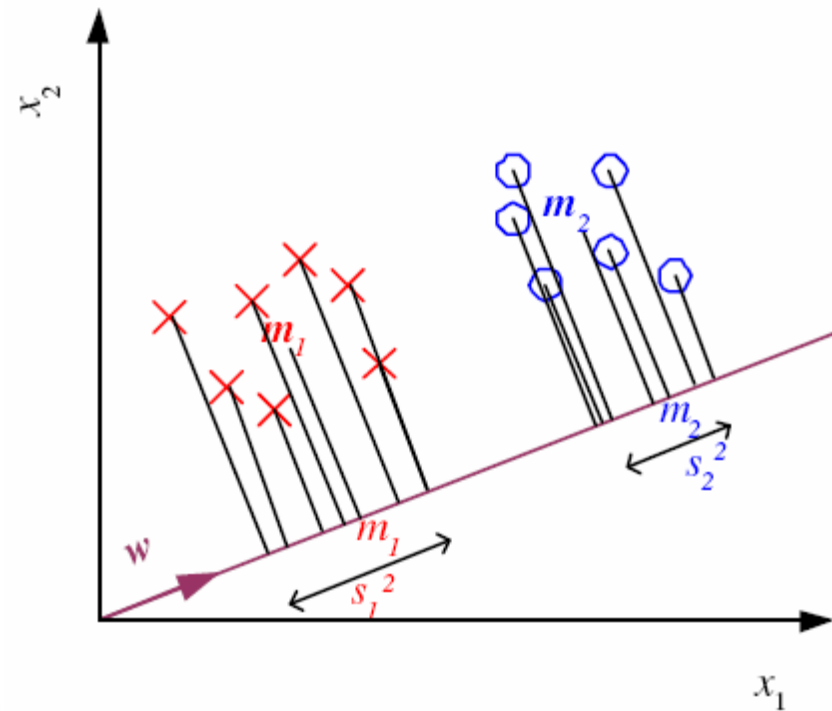
Map from CIA – The World Factbook: <http://www.cia.gov/>

Linear Discriminant Analysis

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

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

$$m_1 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t r^t}{\sum_t r^t} = \mathbf{w}^T \mathbf{m}_1 \quad s_1^2 = \sum_t (\mathbf{w}^T \mathbf{x}^t - m_1)^2 r^t$$



- Between-class scatter

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

- Within-class scatter

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

$$\text{where } \mathbf{S}_1 = \sum_t r^t (\mathbf{x}^t - \mathbf{m}_1) (\mathbf{x}^t - \mathbf{m}_1)^T$$

$$s_1^2 + s_2^2 = \mathbf{w}^T \mathbf{S}_W \mathbf{w}$$

$$\text{where } \mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$



Fisher's Linear Discriminant

- Find $\mathbf{w}$ that max

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

- LDA soln:

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

- Parametric soln:

$$\mathbf{w} = \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$\text{when } p(\mathbf{x}|C_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma)$$

K > 2 Classes

- Within-class scatter:

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

- Between-class scatter:

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

- Find $\mathbf{W}$ that max

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|} \quad \text{The largest eigenvectors of } \mathbf{S}_W^{-1} \mathbf{S}_B$$

maximum rank of $K - 1$

