Intro to ML

October 20th, 2021

CHAPTER 6:

Dimensionality Reduction

Why Reduce Dimensionality?

- Reduces time complexity: Less computation
- Reduces space complexity: Fewer parameters
- Saves the cost of observing the feature
- Simpler models are more robust on small datasets
- More interpretable; simpler explanation
- 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

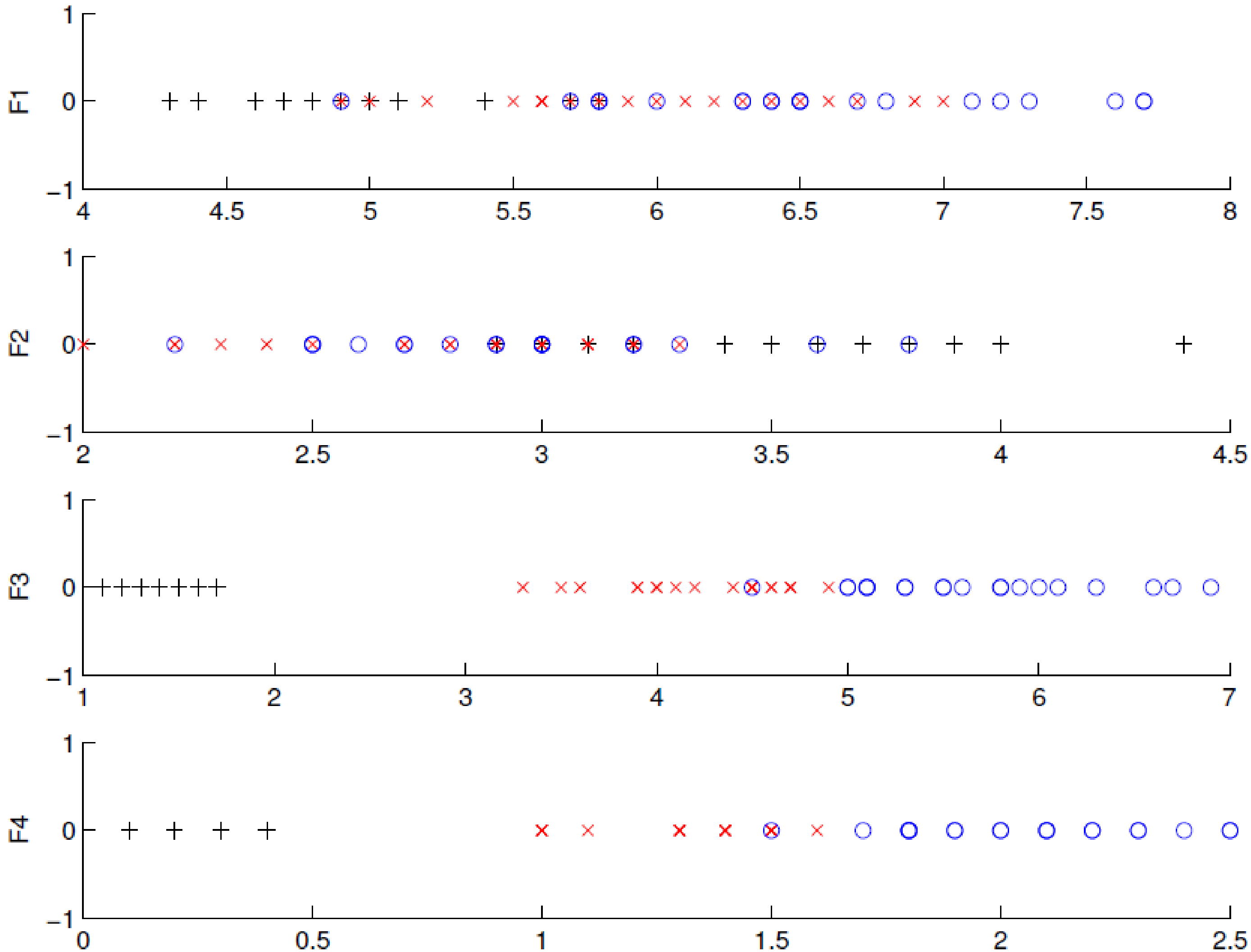
Subset Selection

- There are 2^d subsets of d features
  - Exhaustive search is not possible
- **Forward search**: Add the best feature at each step
  - Set of features F initially Ø.
  - At each iteration, find the best new feature

    $j = \text{argmin}_i\, E\,(\,F \cup x_i\,)$

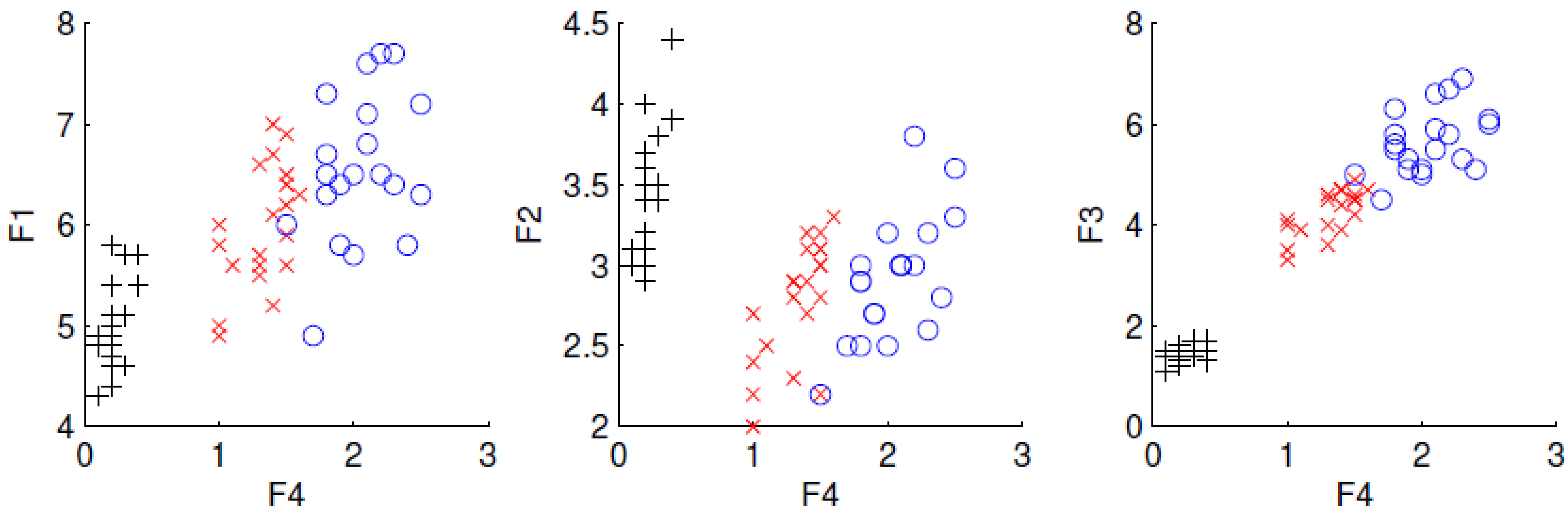
    Use train, validation, test setup
  - Add x_j to F if $E\,(\,F \cup x_j\,) < E\,(\,F\,)$
- **Backward search**: Start with all features and remove one at a time, if possible.
- Hill-climbing $O(d^2)$ algorithm
- Floating search (Add k, remove l, $k>l$)

Iris data: Single feature



Chosen

Iris data: Add one more feature to F4



Chosen

Principal Components Analysis

- 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 $\boxed{\mathrm{Var}(z)}$ is maximized

information

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

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

- First projection: Maximize Var(z_1) subject to $|| \boldsymbol{w_1} ||=1$

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

A Lagrange problem

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

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 $\boldsymbol{w}_1$

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

$\sum \boldsymbol{w}_2 = \alpha\, \boldsymbol{w}_2$ that is, $\boldsymbol{w}_2$ is another eigenvector of $\sum$ with the second largest eigenvalues and so on

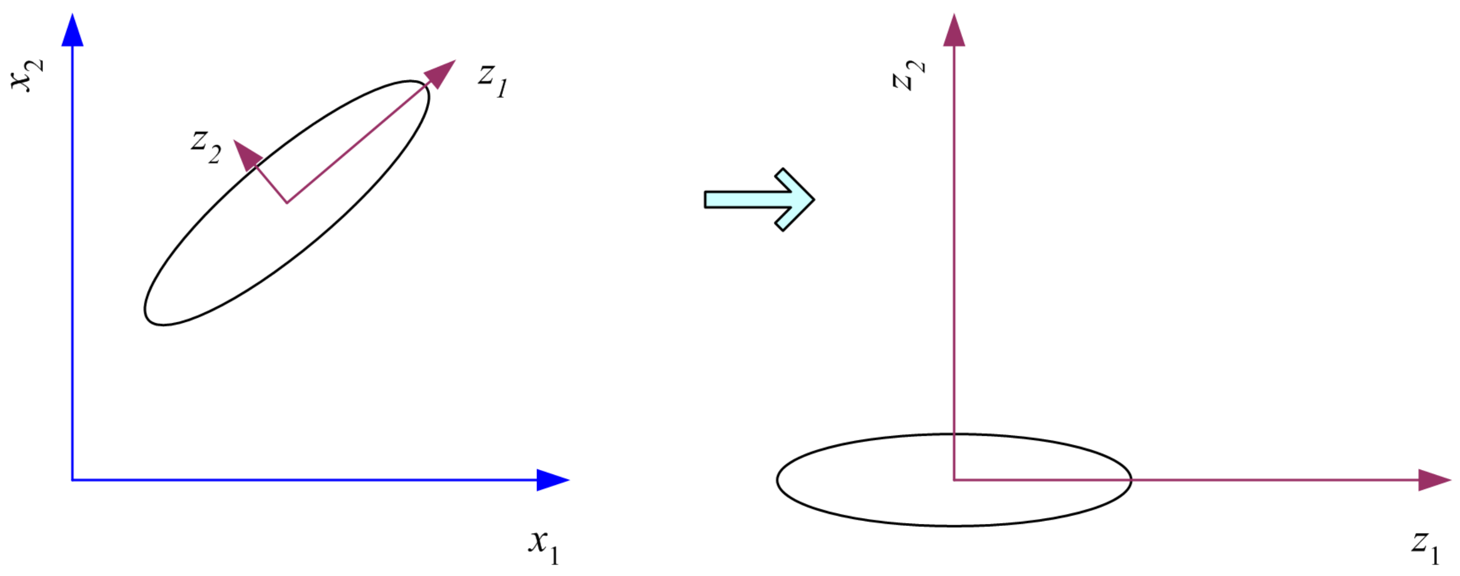
- $\sum$ is symmetric -> two eigenvectors orthogonal
- $\sum$ is positive definite -> all eigenvalues are positive
- $\sum$ is singular -> it has a lower rank (dimension require k < d)

What PCA does

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

where the columns of W are the eigenvectors of Σ and m is sample mean

Centers the data at the origin and rotates the axes

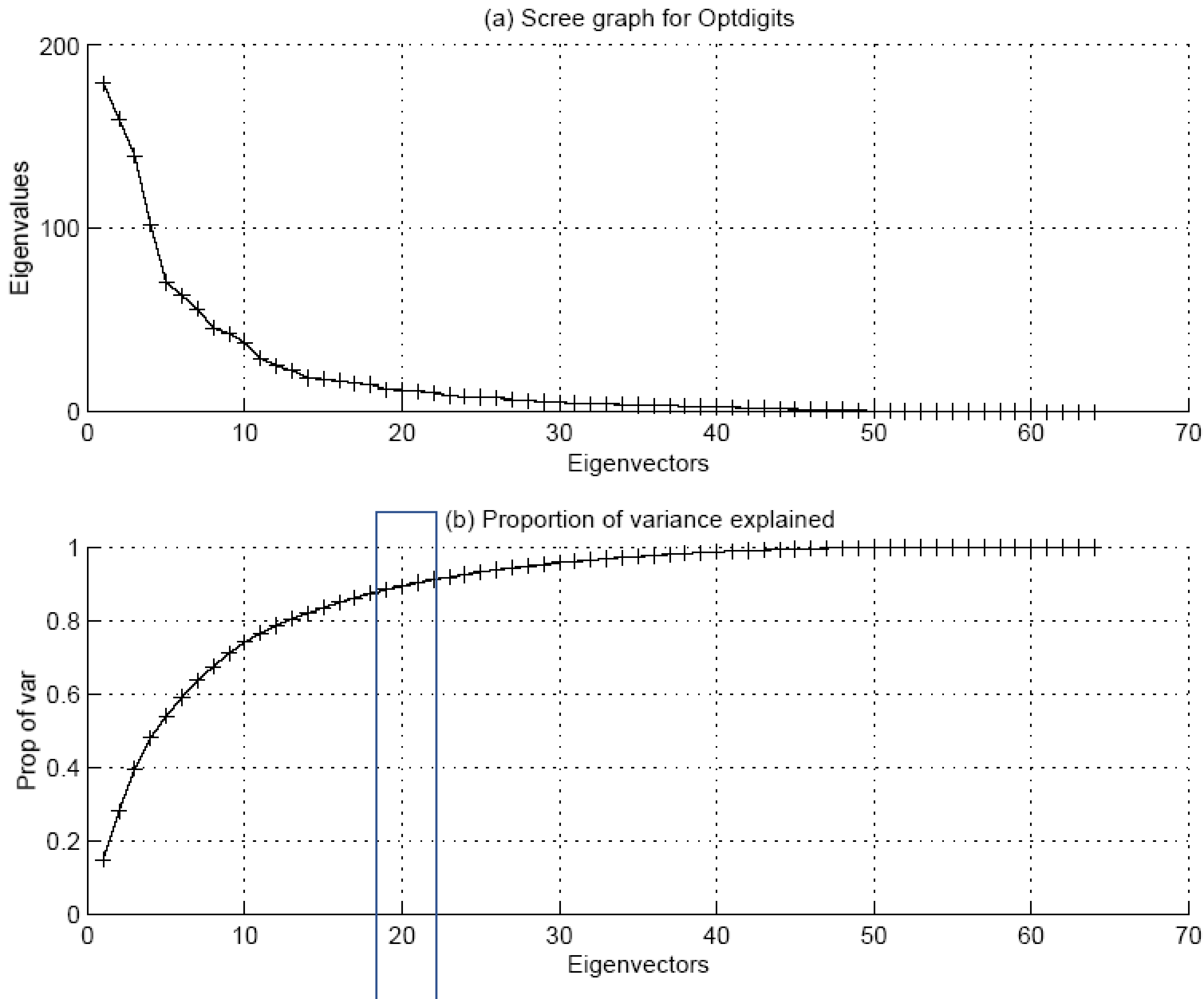


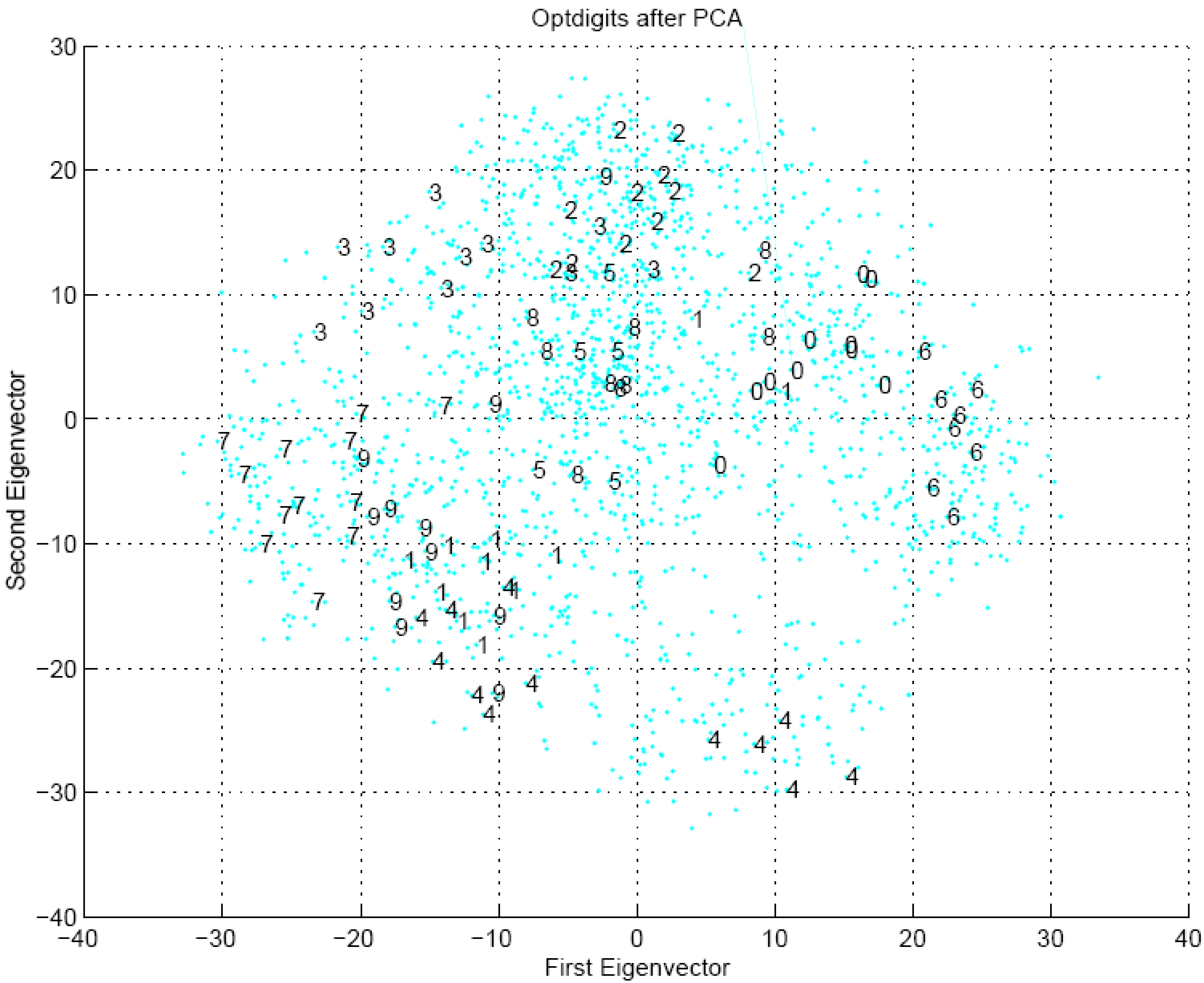
How to choose k ?

- 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 (eigenvalues) are sorted in descending order
- Typically, stop at PoV>0.9
- Scree graph plots of PoV vs *k*, stop at "elbow"
  - *adding another eigenfactor does not add much variances*


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



Visual display

After PCA

- From d-dimension to k-dimension
    - K<d, parametric discriminant function rely on fewer parameters (due to lesser features)
    - Since these features are projected to be uncorrelated (orthogonal is you assume Normal distribution)
        - The multivariate Gaussian’s covariance matrix can now be assumed to be diagonal

- PCA is an unsupervised method of dimension reduction -> does not require the knowledge of label, just the data

Another formulation of PCA

- Find a matrix W such that when we have $z=W^Tx$, we will get cov(z) = D, where D is any diagonal matrix (uncorrelated zi)
- Form a dxd matrix C, whose ith column is the normalized eigenvectors ci of S, $C^TC = I$

$$S=SCC^T= S(c1, c2, c3\ldots cd)C^T= (\lambda_1 c_1, \lambda_2 c_2 \ldots \lambda_d c_d)C^T=CDC^T$$

Where D = diagonal matrix with eigenvalues on the diagonal

- We also call this spectral decomposition of S
- $C^TSC= D$
- $z=W^Tx$, $Cov(z) = W^TSW$, want this to be a diagonal matrix, just set W=C

Factor Analysis

- Find a small number of unobservable 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

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

Factors are:
Unit normal and
uncorrelated

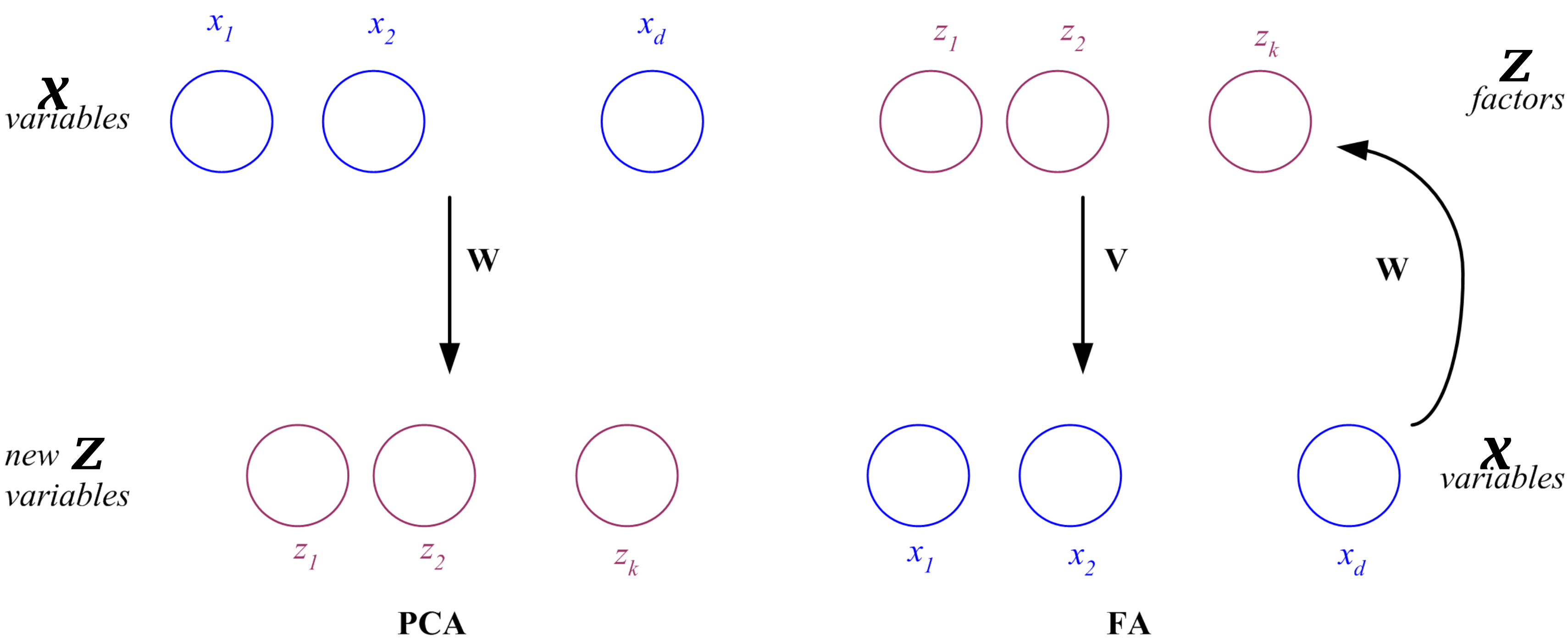
ε_i are the noise sources

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

and v_{ij} are the factor loadings

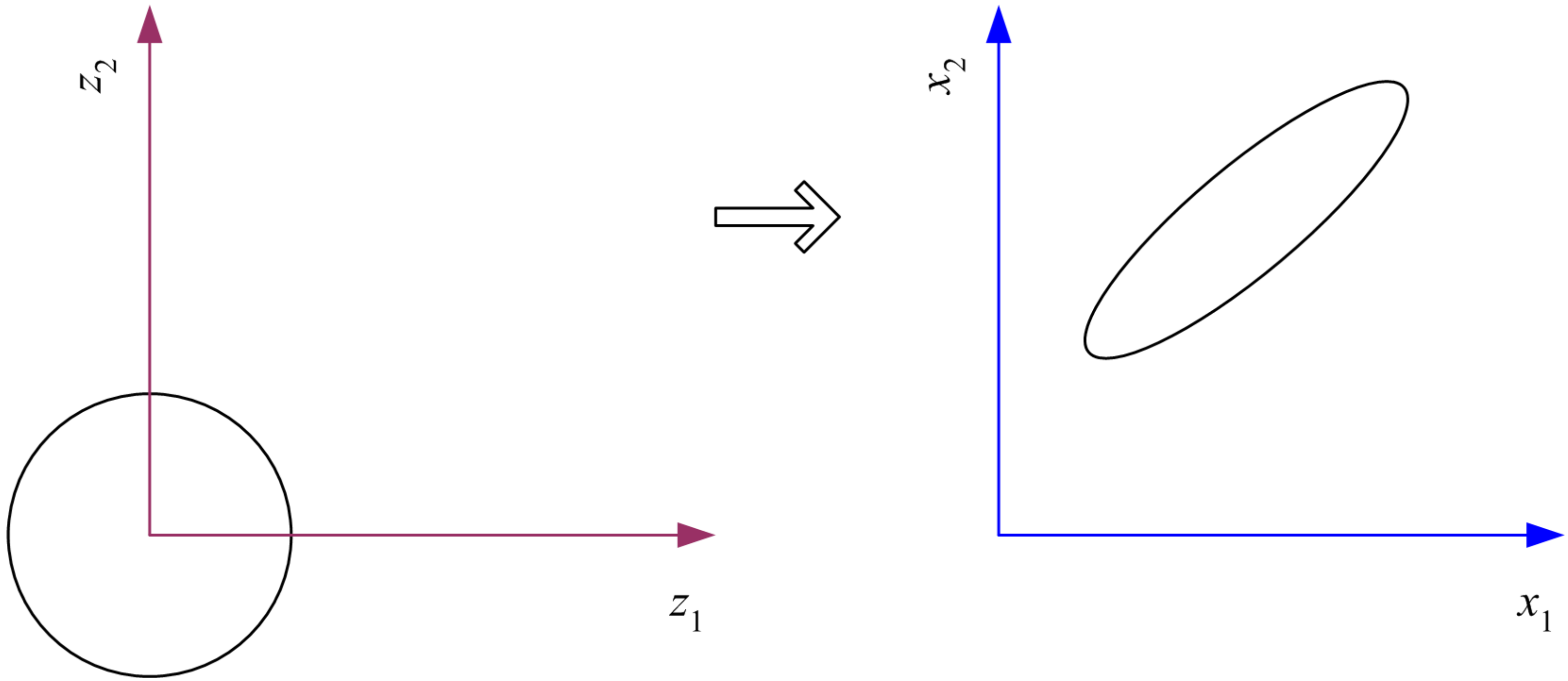
PCA vs FA

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



Factor Analysis

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



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

$Var(Xi) = V_{i1}\hat{}2 + V_{i2}\ \hat{}2 + \ldots.\ \Psi_i$

- Variance of Xi attributed to common factor 1-k and noise

$$\sum = \mathrm{cov}\,(X) = \mathrm{cov}(Vz + \varepsilon)$$

$$= \mathrm{cov}(Vz) + \mathrm{cov}\,(\varepsilon)$$

$$= V\,\boxed{\mathrm{cov}(z)}\,V^T + \Psi$$

Identity matrix

$$= VV^T + \boxed{\Psi}$$

Diagonal matrix

Assume there are two 'hidden factors', and two observable features

Cov(x1,x2)= $v_{11}v_{21} + v_{12}v_{22}$

- $Cov(x1,x2) = v_{11}v_{21} + v_{12}v_{22}$
- If x1 x2 have high covariance due to the first 'factor'
  - The $v_{11}v_{12}$ will be high or $v_{12}v_{22}$ will be high
  - Either way the above terms will be high
- If they depend on different factors, one term high one term low, this summation will be small

$Cov(x_1,z_2) = Cov(v_{12}z_{2,}\ z_2) = v_{12}$

- $Cov(x,z) = V$
- Loading represent correlation between variables and the factors

PCA: FIND THE EIGENVECTOR OF S

- S estimate $\sum$
- Factor analysis basically solves the following equation

$S=VV^T+\Psi$

- $z=W^Tx$, $Cov(z) = W^TSW$, want this to be a diagonal matrix, just set W=C