BLG 527E Machine Learning

FALL 2021-2022

Assoc. Prof. Yusuf Yaslan & Assist. Prof. Ayşe Tosun

Dimensionality Reduction

Why Reduce Dimensionality?

- Ideally, we should not need feature selection or extraction as a separate process; the classifier (or regressor) should be able to use whichever features are necessary, discarding the irrelevant.
- There are several reasons why we are interested in reducing dimensionality as a separate preprocessing step:

Why Reduce Dimensionality?

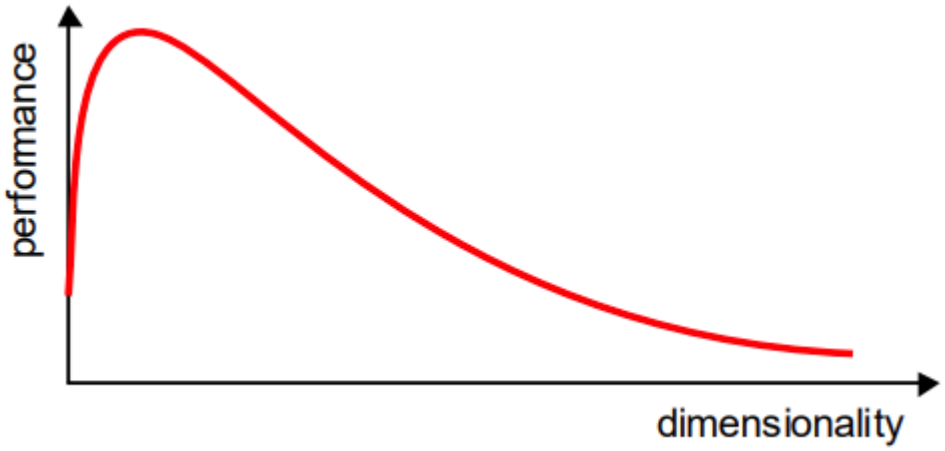
- Reduces time complexity: Less computation
- Reduces space complexity: Less parameters
- Saves the cost of observing the feature
- Simpler models are more robust on small datasets
- More interpretable; simpler explanation
- Data visualization (structure, groups, outliers, etc) if plotted in 2 or 3 dimensions

Curse of Dimensionality

- A term coined by Bellman in 1961
- Refers to the problems associated with multivariate data analysis as the dimensionality increases
- Describes the increasing difficulty in training a model when more predictor variables are added to it
- For a given sample size, there is a maximum number of features above which the performance of our classifier will degrade rather than improve

Curse of Dimensionality

- In most cases, the additional information that is lost by discarding some features is (more than) compensated by a more accurate mapping in the lower-dimensional space



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

- We are interested in finding the best subset of the set of features. The best subset contains the least number of dimensions that most contribute to accuracy. We discard the remaining, unimportant dimensions.
- There are 2^d subsets of d features but we cannot test for all of them unless d is small and we employ heuristics to get a reasonable (but not optimal) solution in reasonable (polynomial) time.

Subset Selection: Search Strategy and Objective Function

- A search strategy is used to select candidate subsets. i.e sequential, randomized and exponential algorithms
- An objective function is used to evaluate these candidates
- Objective functions are divided in two groups
- Filters: evaluate subsets by their information content, e.g., interclass distance, statistical dependence or information-theoretic measures
- Wrappers: use a classifier to evaluate subsets by their predictive accuracy (on validation data) by statistical resampling or cross-validation

This slide has been adopted from Dr. Ricardo Gutierrez-Osuna's Lecture Notes

Subset Selection

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

    $j = \text{argmin}_i E (F \cup x_i)$ % *E():* Error on the validation set
  - Add x_j to F if $E (F \cup x_j) < E (F)$

- Hill-climbing $O(d^2)$ algorithm (train and test the system d ,d-1,d-2…..(d-k) times)

Subset Selection

- This is a local search procedure and does not guarantee finding the optimal subset, namely, the minimal subset causing the smallest error
- Backward search: Start with all features and remove one at a time, if possible.
- Floating search (Add *k*, remove *l*) number of added features and removed features can also change at each step.

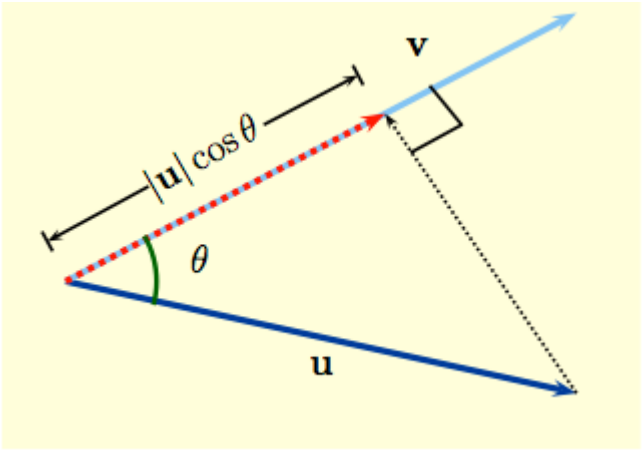
Principal Components Analysis (PCA)

- Find a low-dimensional space such that when $\boldsymbol{x}$ is projected there, information loss is minimized.

Vector Projections

- The red vector has length $\text{length} = \|\mathbf{u}\| \cos\theta = \dfrac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{v}\|}$,
- $\mathbf{v}/\|\mathbf{v}\|$ is a unit vector in the direction of $\mathbf{v}$.
- The dashed red vector

$$\text{dashed red vector} = \left(\frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}}\right)\mathbf{v}\,.$$



This lide is adopted from the following lecture notes:
https://web.ma.utexas.edu/users/m408m/Display12-3-4.shtml

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

$$\text{Var}(z) = \text{Var}(\boldsymbol{w}^T\boldsymbol{x}) = E[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})^2]$$
$$= E[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})]$$
$$= E[\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{w}]$$
$$= \boldsymbol{w}^T E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]\boldsymbol{w} = \boldsymbol{w}^T \sum \boldsymbol{w}$$

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

Principal Components Analysis (PCA)

- Maximize Var(z) subject to $||\boldsymbol{w}||=1$
- Use Lagrange Multipliers:
- max f(x,y) subject to g(x,y)=c we need both f and g continuous with first partial derivative.
- We introduce a new variable (λ) Lagrange Multiplier and optimize:
- $L(x,y, \lambda) = f(x,y) - \lambda(g(x,y)-c)$

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

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

Take the derivative w.r.t. $\boldsymbol{w}_1$ and set equal to 0.

$$2\Sigma \mathbf{w}_1 - 2\alpha \mathbf{w}_1 = 0$$

$\sum \boldsymbol{w}_1 = \alpha \boldsymbol{w}_1$ 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 $\boldsymbol{w}_1$
- We should add the second constraint with β Lagrange multiplier

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

- Take the derivative w.r.t. $\boldsymbol{w}_2$ and set equal to 0.

$$2\Sigma \mathbf{w}_2 - 2\alpha \mathbf{w}_2 - \beta \mathbf{w}_1 = 0$$

- Premultiply with $\boldsymbol{w}_1^T$

$$2\mathbf{w}_1^T\Sigma\mathbf{w}_2 - 2\alpha\mathbf{w}_1^T\mathbf{w}_2 - \beta\mathbf{w}_1^T\mathbf{w}_1 = 0 \qquad 2\alpha\mathbf{w}_1^T\mathbf{w}_2 = 0$$

$$\mathbf{w}_1^T\Sigma\mathbf{w}_2 = \mathbf{w}_2^T\Sigma\mathbf{w}_1 \qquad \Sigma\mathbf{w}_1 = \lambda_1\mathbf{w}_1 \qquad \lambda_1\mathbf{w}_2^T\mathbf{w}_1 = 0$$

- Then $\beta = 0$

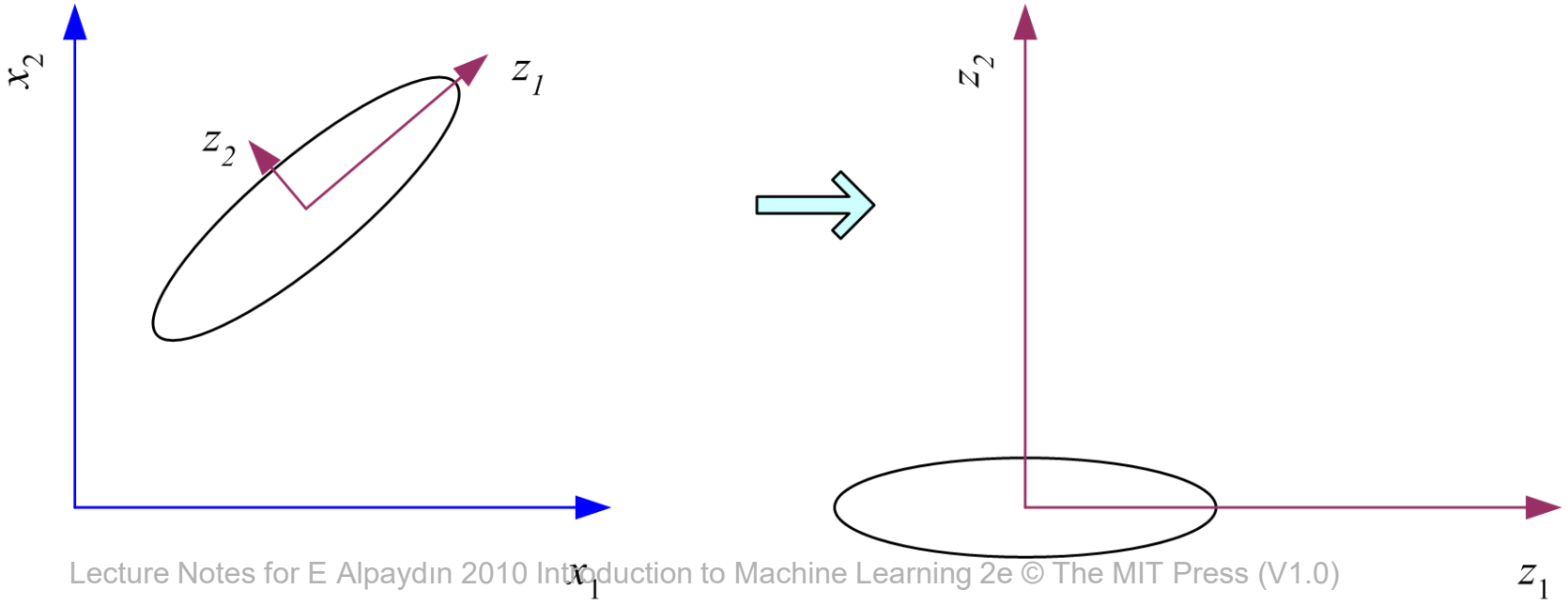
$\sum \boldsymbol{w}_2 = \alpha \boldsymbol{w}_2$ that is, $\boldsymbol{w}_2$ is another eigenvector of $\sum$ and so on.

What PCA does

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

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

Centers the data at the origin and rotates the axes

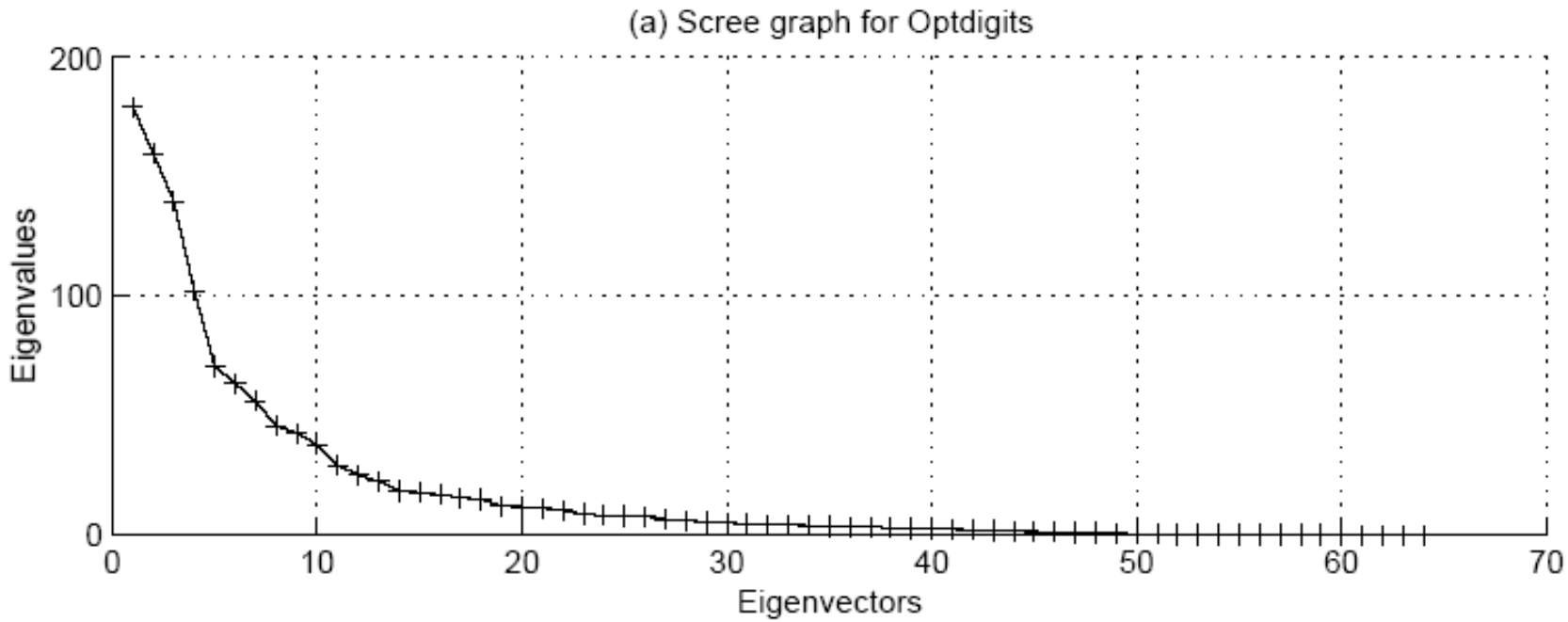


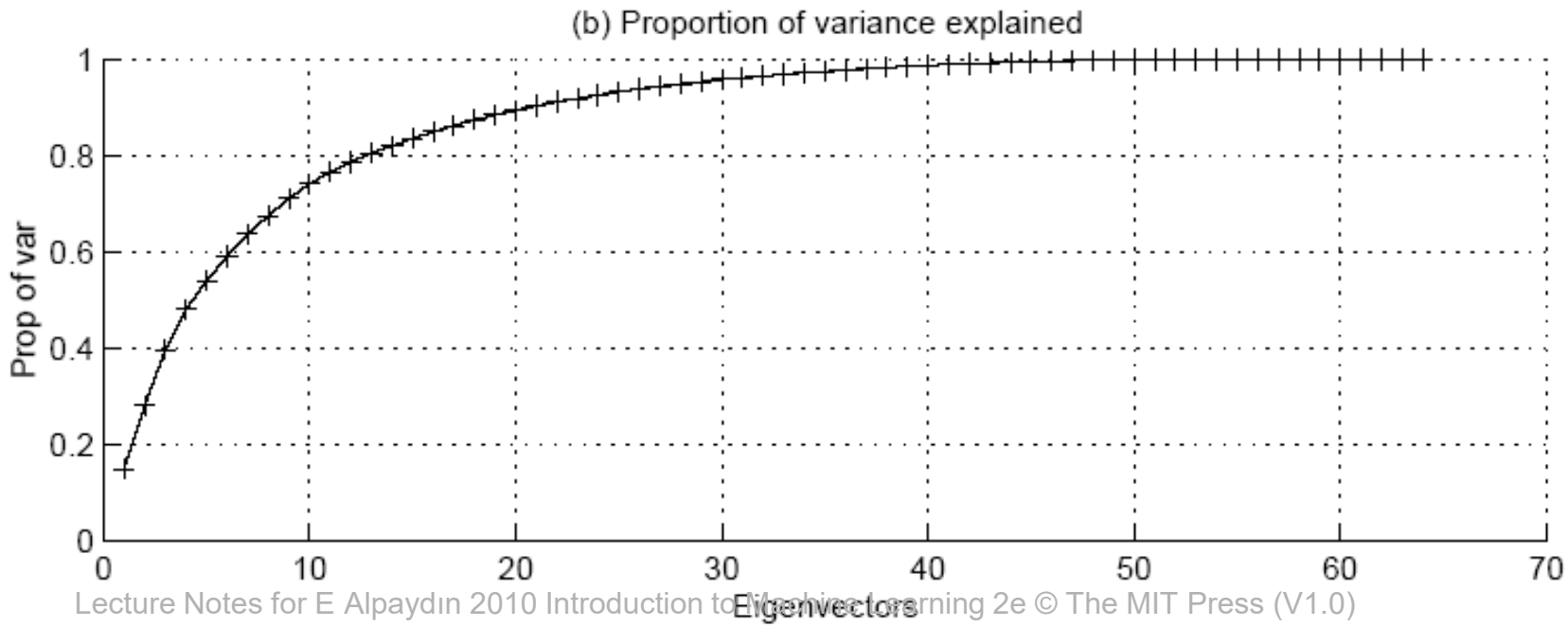
How to choose k ?

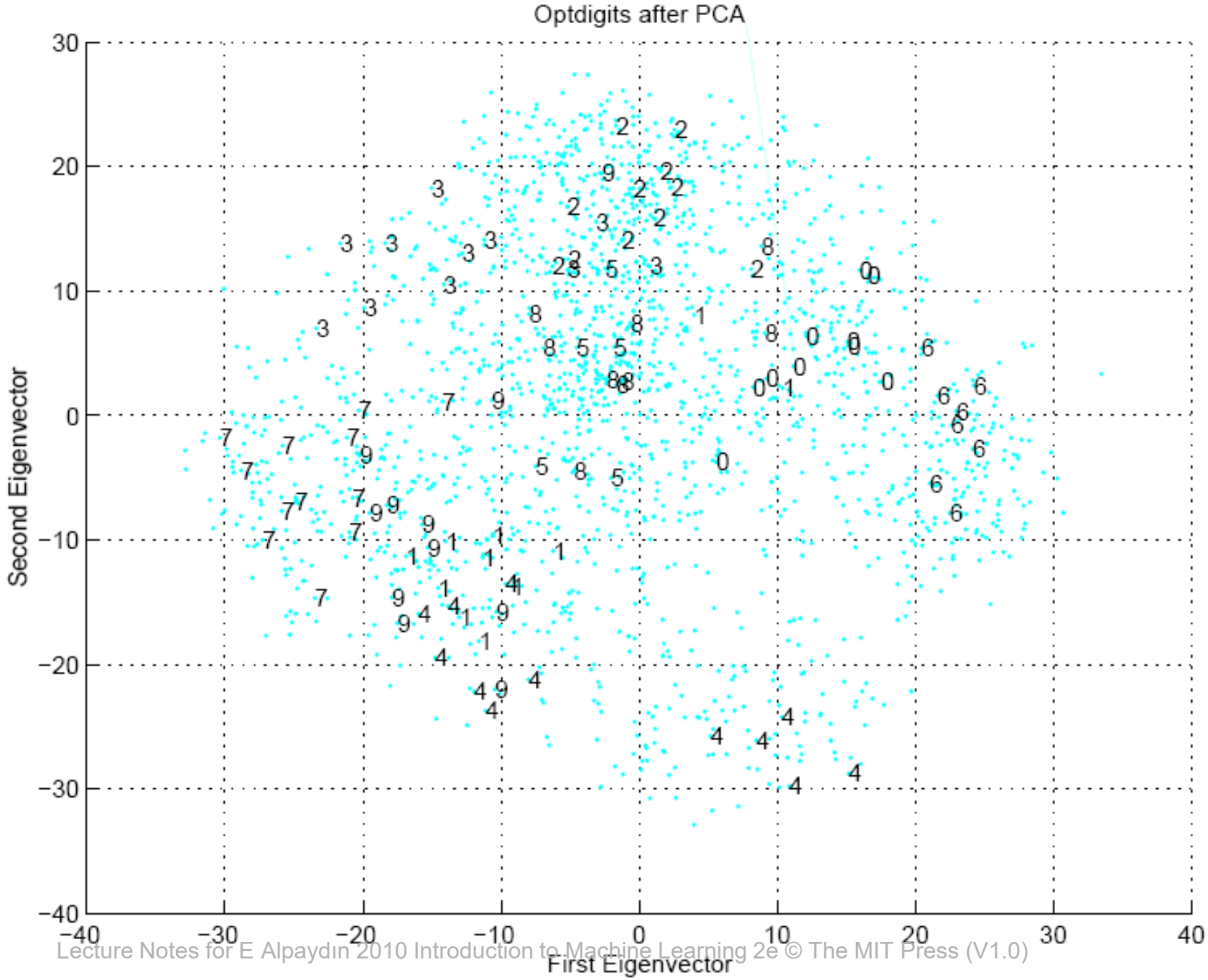
- Proportion of Variance (PoV) explained

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

  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
−30
−20
−10
0
10
20
30
40

Factor Analysis

- In PCA from the original dimensions x_i we form a new set of variables z that are linear combinations of x_i's.
- In factor analysis (FA), we assume that there is a set of unobservable latent factors z_j , j = 1, . . . , k, which when acting in combination generate x.
- The goal is to characterize the dependency among the observed variables by means of a smaller number of factors
- Though factor analysis always partitions the variables into factor clusters, whether the factors mean anything, or really exist, is open to question

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

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

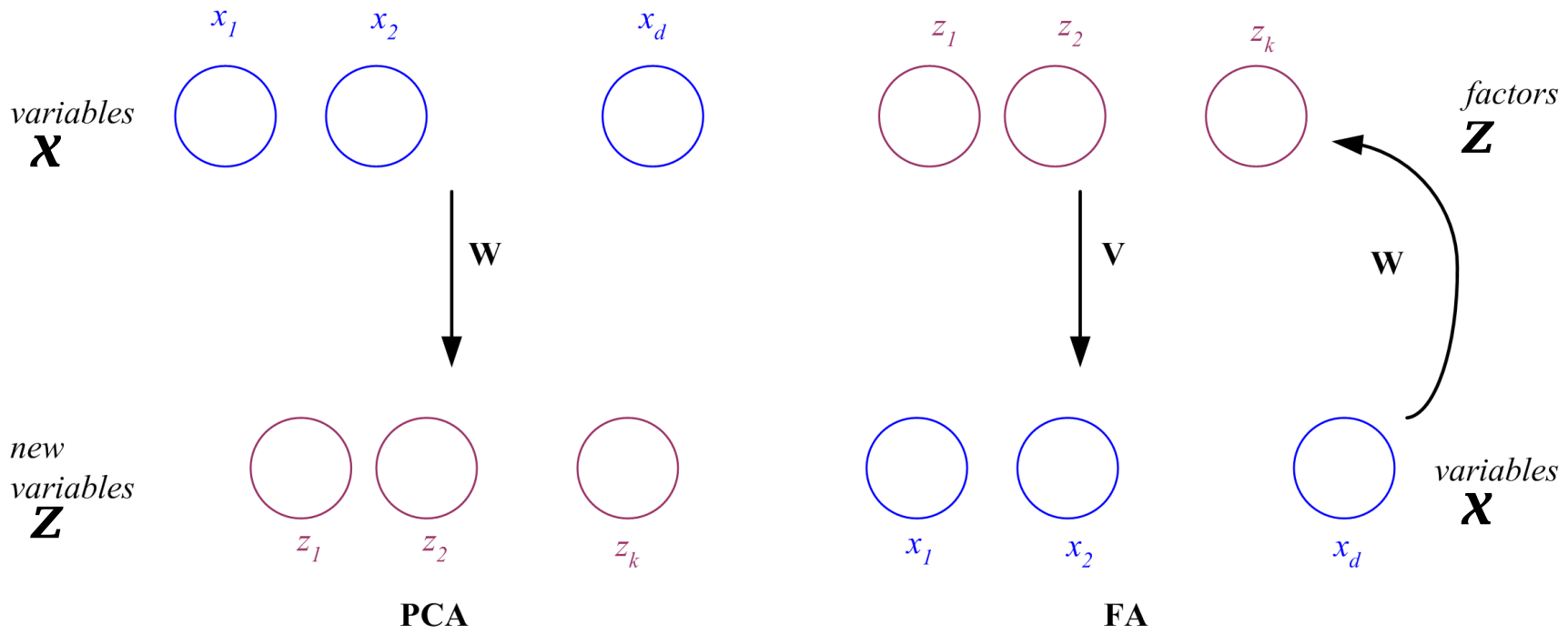
and v_{ij} are the factor loadings

- This example is obtained from Andrew Ng Lecture Notes

PCA vs FA

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

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



$$x_i - \mu_i = \sum_{j=1}^{m} v_{ij} z_j + \varepsilon_i, \qquad i = 1,2,....,l$$

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

We assume that $\boldsymbol{\mu} = \mathbf{0}$ and given that Var(z_j)=1, Cov(z_i, z_j)=0,

$$E[\mathbf{x}\mathbf{x}^T] = \mathbf{V}E[\mathbf{z}\mathbf{z}^T]\mathbf{V}^T + \Sigma_{\boldsymbol{\varepsilon}} = \Sigma_x = \mathbf{V}\mathbf{V}^T + \Sigma_{\boldsymbol{\varepsilon}} = \mathbf{V}\mathbf{V}^T + \Psi$$

There are different methods to obtain V, We will further assume that latent variables can be obtained linear combinations of the observations

$$\mathbf{z} = \mathbf{W}\mathbf{x}$$

- Post multiplying with x and taking expectation gives

$$E[\mathbf{z}\mathbf{x}^T] = E[\mathbf{z}\mathbf{z}^T\mathbf{V}^\mathbf{T}] + E[\mathbf{z}\varepsilon^T] = \mathbf{V}^\mathbf{T}$$

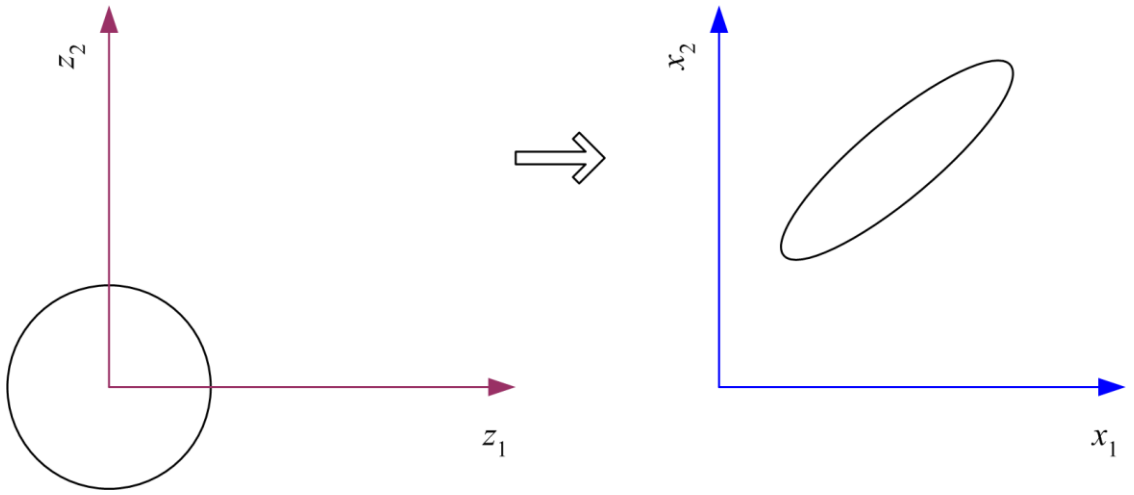
$$E[\mathbf{z}\mathbf{x}^T] = E[\mathbf{W}\mathbf{x}\mathbf{x}^T] = \mathbf{W}\boldsymbol{\Sigma}_\mathbf{x}$$

$$\mathbf{W}\boldsymbol{\Sigma}_\mathbf{x} = \mathbf{V}^\mathbf{T}, \qquad and \qquad \mathbf{W} = \mathbf{V}^\mathbf{T}\boldsymbol{\Sigma}_\mathbf{x}^{-1}$$

$$\mathbf{z} = \mathbf{W}\mathbf{x} \Rightarrow \mathbf{z} = \mathbf{V}^\mathbf{T}\boldsymbol{\Sigma}_\mathbf{x}^{-1}\mathbf{x}$$

Factor Analysis

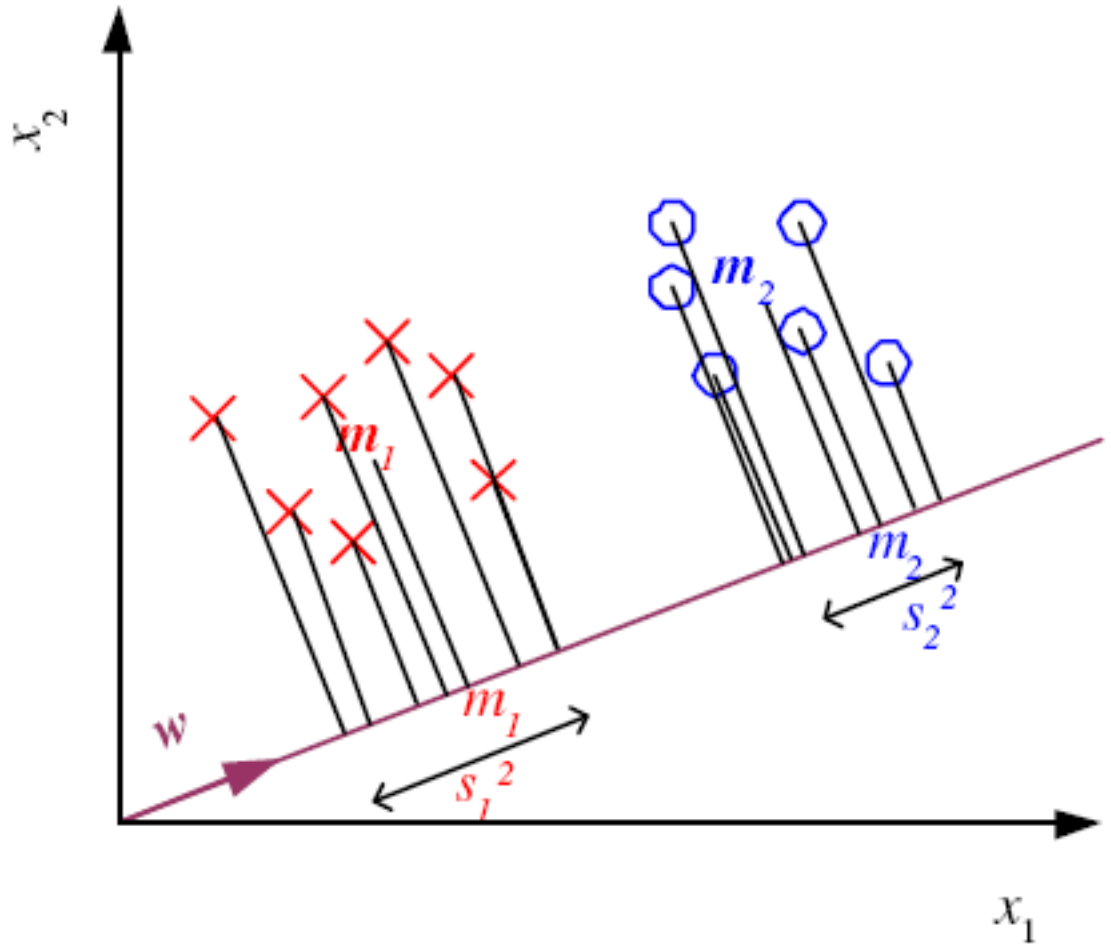
- In FA, factors z_j are stretched, rotated and translated to generate $\mathbf{x}$



- For dimensionality reduction, FA offers no advantage over PCA except the interpretability of factors allowing the identification of common causes, a simple explanation, and knowledge extraction

Linear Discriminant Analysis

- Find a low-dimensional space such that when ***x*** is projected, classes are well-separated.
- Find ***w*** that maximizes



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

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

- Between-class scatter:

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

- Within-class scatter:

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

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

$$s_1^2 + s_1^2 = \mathbf{w}^T\mathbf{S}_W\mathbf{w} \text{ where } \mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

Fisher's Linear Discriminant

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

- Take the derivative of J w.r.t. to w and set it equal to 0.

$$\frac{(2\mathbf{S}_B w)\mathbf{w}^T \mathbf{S}_W \mathbf{w} - (2\mathbf{S}_w w)\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{(\mathbf{w}^T \mathbf{S}_W \mathbf{w})^2} = 0$$

$$\mathbf{w}^T\mathbf{S}_W\mathbf{w}(\mathbf{S}_B\mathbf{w}) - \mathbf{w}^T\mathbf{S}_B\mathbf{w}(\mathbf{S}_w\mathbf{w}) = 0$$

$$\frac{\mathbf{w}^T\mathbf{S}_W\mathbf{w}(\mathbf{S}_B w) - \mathbf{w}^T\mathbf{S}_B\mathbf{w}(\mathbf{S}_w w)}{(\mathbf{w}^T\mathbf{S}_W\mathbf{w})} = 0$$

$$(\mathbf{S}_B\mathbf{w}) - \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}(\mathbf{S}_w\mathbf{w})}{(\mathbf{w}^T\mathbf{S}_W\mathbf{w})} = 0, \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{(\mathbf{w}^T\mathbf{S}_W\mathbf{w})} = \lambda$$

$$\mathbf{S}_B\mathbf{w} - \lambda(\mathbf{S}_w\mathbf{w}) = 0, \qquad \mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w} \qquad \mathbf{S}_w^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}$$

$$\mathbf{S}_B w = (m_1 - m_2)(m_1 - m_2)^T w = (m_1 - m_2)\alpha$$

$$\mathbf{S}_B w = (m_1 - m_2)(m_1 - m_2)^T w = (m_1 - m_2)\alpha$$

$\mathbf{S}_B w$ for any vector w points in the same direction as $(m_1 - m_2)$

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

Fisher's Linear Discriminant

- Find $\mathbf{w}$ that max $$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} = \frac{\left|\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)\right|^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$
- LDA soln: $$\mathbf{w} = c \cdot \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$$
- Parametric soln: $$\mathbf{w} = \Sigma^{-1} (\mu_1 - \mu_2)$$ when $p(\mathbf{x} | C_i) \sim \mathcal{N}(\mu_i, \Sigma)$

K>2 Classes

- Within-class scatter:

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

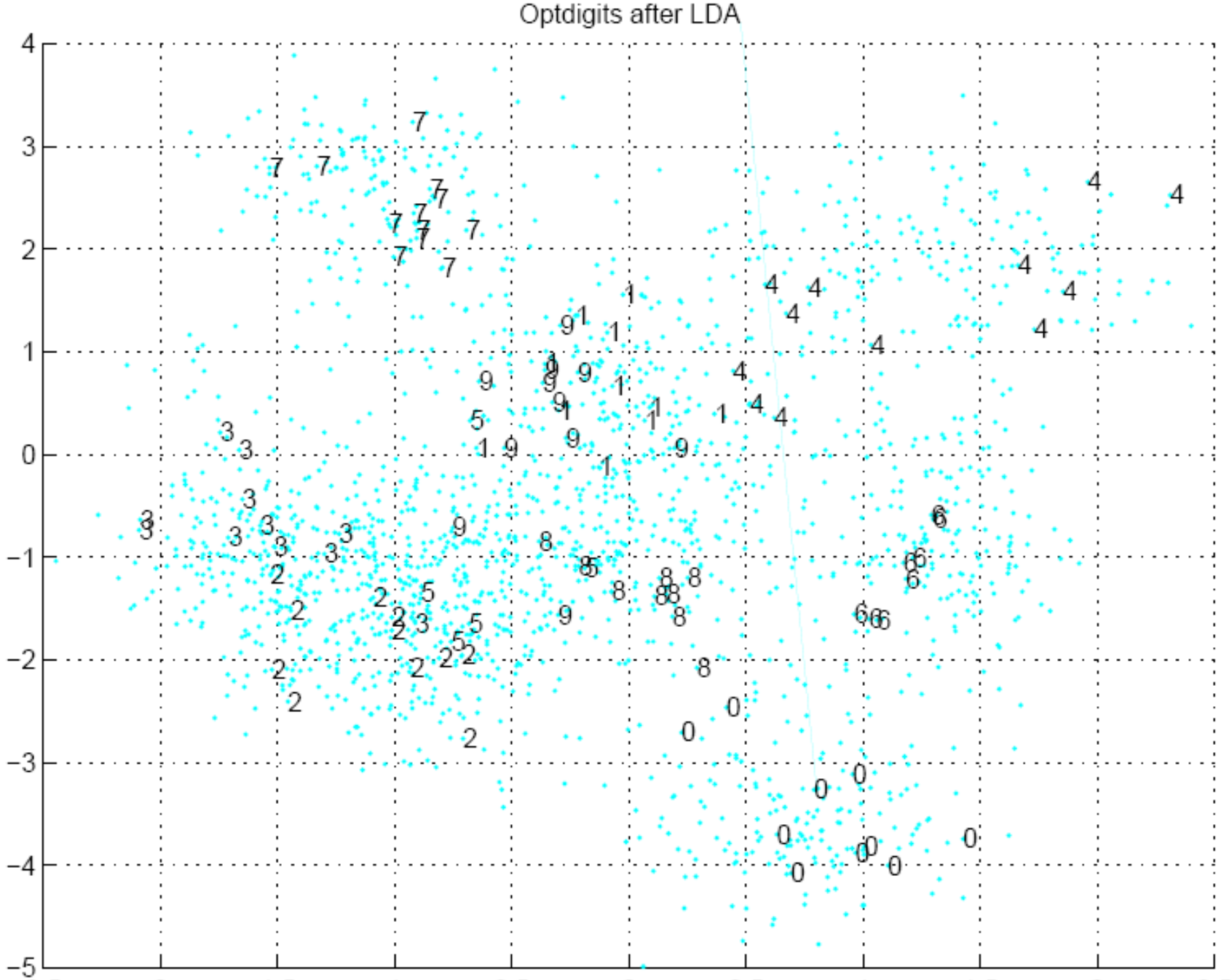
- Between-class scatter:

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

- Find $\mathbf{W}$ 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
4
3
2
1
0
−1
−2
−3
−4
−5
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5