Dimensionality Reduction: Principal Component Analysis (주성분 분석)

Table of Contents

I Feature Extraction

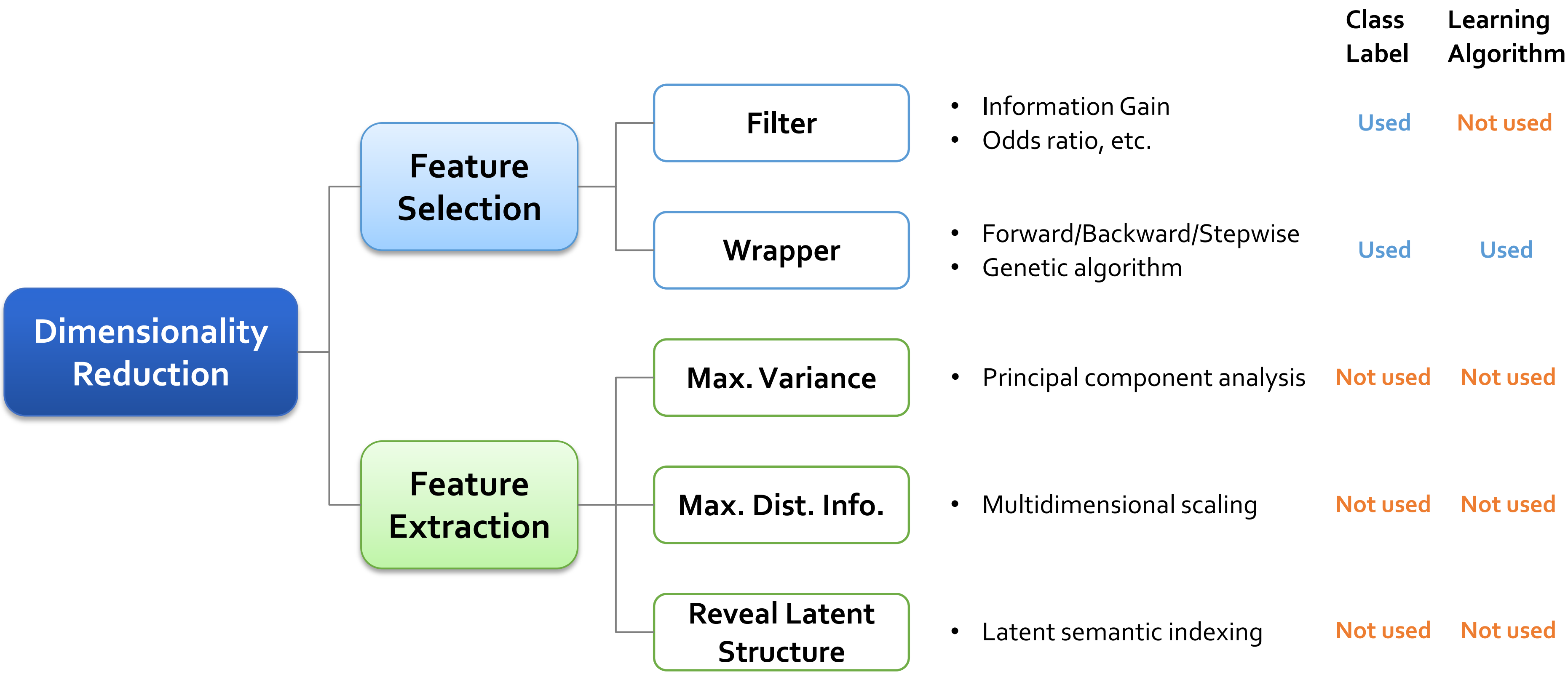
II Principal Component Analysis

Dimensionality Reduction

- **Why is dimensionality reduction necessary?**
  - To make large problems **computationally efficient** (conserving computation, storage and network resources)
  - To **improve the quality** of data mining results
    - Improve classification accuracy or clustering modularity
    - Reduce the amount of training data needed to obtain a desired level of performance

Dimensionality Reduction

- A simplified taxonomy of dimensionality reduction techniques



Dimensionality Reduction

Feature selection vs. feature extraction

- **Feature selection**: select a small subset of original variables
- **Feature extraction**: construct/extract a new set of features based on the original variables

V_1	V_2	V_3	…	V_n
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

Feature Selection

V_1	V_5	V_8
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…

Feature Extraction

Z_1	Z_2	Z_3
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…

$$Z_1 = V_1 + 0.2*V_2$$

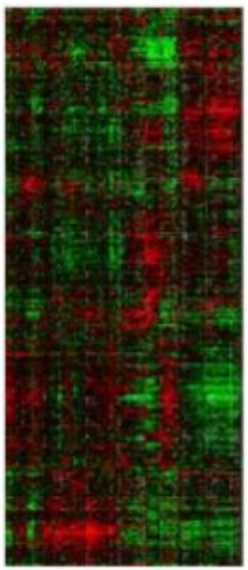
$$Z_2 = V_3 - 2*V_5$$

$$Z_3 = V_4 + V_6 - V_9$$

Feature Extraction

Why feature extraction than feature selection?

- Features in a data space are highly correlated and every feature contains relevant information
- The intrinsic dimension may be small



Gene expression

Face images

Handwritten digits

Table of Contents

I Feature Extraction

II Principal Component Analysis

Principal Component Analysis: PCA

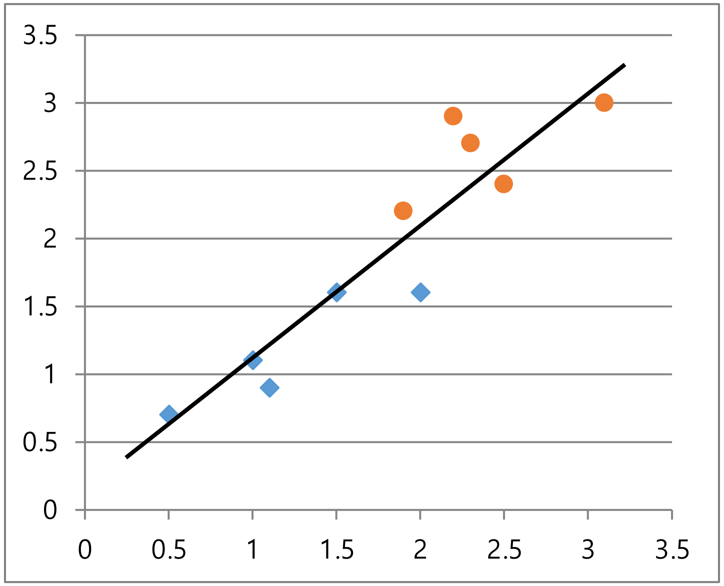
Why PCA?

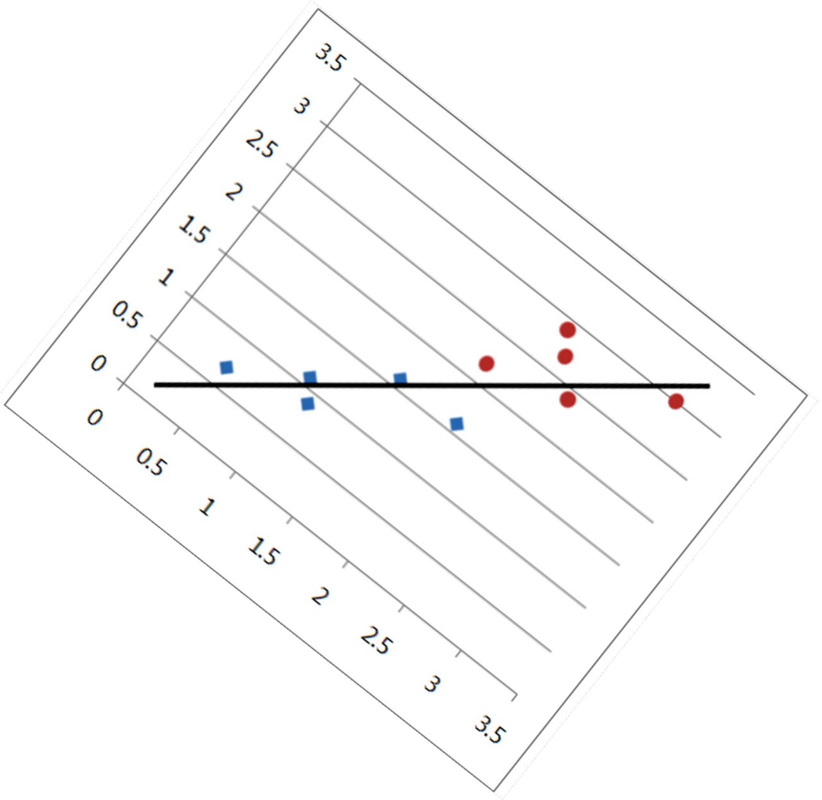
- **Feature Extraction**
  - Extract useful features among a large number of features
- **Data Compression**
  - Efficient storage and retrieval
- **Visualization**
  - Projection of high-dimensional data onto 2D or 3D

Principal Component Analysis: PCA

Principal Component Analysis: PCA

- To find a set orthogonal bases to preserve the variance of the original data

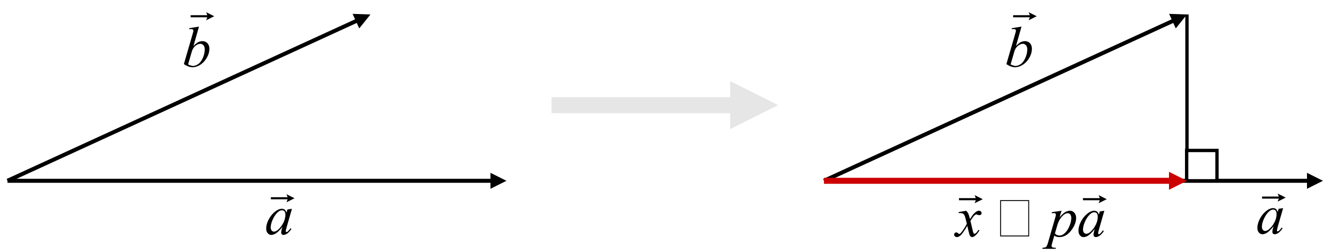




Principal Component Analysis: PCA

Principal Component Analysis: PCA

- Projection onto a basis



$$(\vec{b} - p\vec{a})^{\mathrm{T}}\vec{a} = 0 \quad \Rightarrow \quad \vec{b}^{\mathrm{T}}\vec{a} - p\vec{a}^{\mathrm{T}}\vec{a} = 0 \quad \Rightarrow \quad p = \frac{\vec{b}^{\mathrm{T}}\vec{a}}{\vec{a}^{\mathrm{T}}\vec{a}}$$

$$\vec{x} = p\vec{a} = \frac{\vec{b}^{\mathrm{T}}\vec{a}}{\vec{a}^{\mathrm{T}}\vec{a}}\vec{a}$$

If $\vec{a}$ is unit vector

$$p = \vec{b}^{\mathrm{T}}\vec{a} \quad \Rightarrow \quad \vec{x} = p\vec{a} = (\vec{b}^{\mathrm{T}}\vec{a})\vec{a}$$

Principal Component Analysis: PCA

Principal Component Analysis: PCA

- Covariance
  - $\mathbf{X}$: a data set (m by n, m: # of variables, n: # of records)

$$Cov(\mathbf{X}) \square \frac{1}{n}(\mathbf{X}-\overline{\mathbf{X}})(\mathbf{X}-\overline{\mathbf{X}})^{\mathrm{T}}$$

  - $\mathrm{Cov}(\mathbf{X})_{ij} = \mathrm{Cov}(\mathbf{X})_{ji}$
  - Total variance of the data set

$= \mathrm{tr}[\mathrm{Cov}(\mathbf{X})]$

$= \mathrm{Cov}(\mathbf{X})_{11} + \mathrm{Cov}(\mathbf{X})_{22} + \mathrm{Cov}(\mathbf{X})_{33} + \ldots + \mathrm{Cov}(\mathbf{X})_{mm}$

Principal Component Analysis: PCA

Principal Component Analysis: PCA

- Eigenvalue and eigenvector

When matrix $\mathbf{A}$ is given, scalar value λ and vector $\mathbf{x}$ that satisfy

$$\mathbf{Ax} = \lambda\mathbf{x} \text{ or } (\mathbf{A}-\lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

are called eigenvalue and eigenvector, respectively.

- If a matrix A is an m by m *diagonalizable* matrix,
  - There exist m eigenvalues and eigenvectors
  - Eigenvectors are orthogonal to each other
  - $\mathrm{tr}(\mathbf{A}) = \lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_m$

PCA Procedure

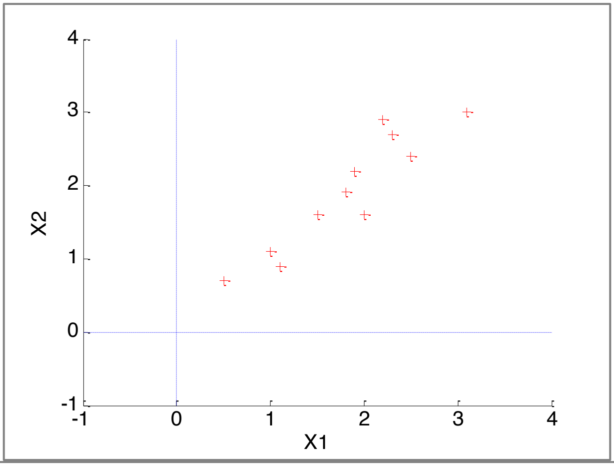
1

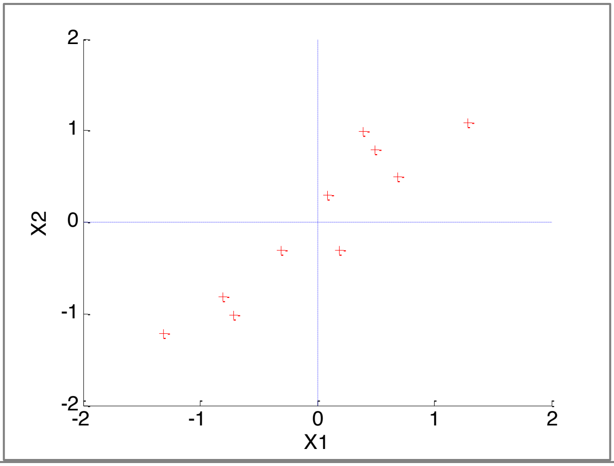
Step 1: data centering

- **Make the mean of the variables equal to 0**

x_1	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
x_2	2.4	0.7	2.9	2.2	3	2.7	1.6	1.1	1.6	0.9

x_1	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
x_2	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01





PCA Procedure

2

Step 2: Formulate the optimization problem

- If a vector x is projected onto a basis w, then the variance after the projection becomes

$$\mathbf{V} = (\mathbf{w}^{\mathrm{T}}\mathbf{X})(\mathbf{w}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}} = \mathbf{w}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{w} = n\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}$$

$\mathbf{S}$ is the sample covariance matrix where $\mathbf{x}$ is normalized.

✓ The purpose of PCA is to maximize the variance V after projection

$$\begin{aligned} &\max \quad \mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w} \\ &\text{s.t.} \quad \mathbf{w}^{\mathrm{T}}\mathbf{w} = 1 \end{aligned} \qquad S = \begin{pmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{pmatrix}$$

PCA Procedure

3

Step 3: Obtain the solution

- By employing Lagrangian multiplier,

$$\max \quad \mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}$$

$$\text{s.t.} \quad \mathbf{w}^{\mathrm{T}}\mathbf{w} = 1$$

$$L = \mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w} - \lambda(\mathbf{w}^{\mathrm{T}}\mathbf{w} - 1),$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{S}\mathbf{w} - \lambda\mathbf{w} = 0 \quad \Rightarrow \quad (\mathbf{S} - \lambda\mathbf{I})\mathbf{w} = 0$$

$$\mathit{Eigenvectors} = \begin{pmatrix} -0.7352 & 0.6779 \\ 0.6779 & 0.7352 \end{pmatrix}$$

$$\mathit{Eigenvalues} = [0.0491 \quad 1.2840]$$

PCA Procedure

4

Step 4: Find the base set of bases

- Sort the eigenvectors in a descending order of eigenvalues

$$FeatureVector = (eig_1, eig_2, \cdots eig_n)$$

$$FeatureVector = \begin{pmatrix} 0.6779 & -0.7352 \\ 0.7352 & 0.6779 \end{pmatrix}$$

✓ One basis can preserve 96% of the original variance in this example (1.2840/(0.0491+1.2840))

Let w_1 be one of the eigenvectors and λ_1 be the corresponding eigenvalue.
The variation of the samples projected onto w_1 is
$(\mathrm{w}_1{}^{\mathrm{T}}\mathrm{X})(\mathrm{w}_1{}^{\mathrm{T}}\mathrm{X})^{\mathrm{T}} = \mathrm{w}_1{}^{\mathrm{T}}\mathrm{X}\mathrm{X}^{\mathrm{T}}\mathrm{w}_1 = \mathrm{w}_1{}^{\mathrm{T}}\mathrm{S}\mathrm{w}_1$
Since $\mathrm{S}\mathrm{w}_1 = \lambda_1 \mathrm{w}_1$,
$\mathrm{w}_1{}^{\mathrm{T}}\mathrm{S}\mathrm{w}_1 = \mathrm{w}_1{}^{\mathrm{T}}\lambda_1\mathrm{w}_1 = \lambda_1\mathrm{w}_1{}^{\mathrm{T}}\mathrm{w}_1 = \lambda_1$

PCA Procedure

5

Step 5: Extract new features

- Project the original data onto the selected bases

- If each eigenvector has elements e_{ik}:

$$\mathbf{e}_1 = \begin{bmatrix} e_{11} \\ e_{21} \\ \vdots \\ e_{p1} \end{bmatrix}, \mathbf{e}_2 = \begin{bmatrix} e_{12} \\ e_{22} \\ \vdots \\ e_{p2} \end{bmatrix}, \ldots, \mathbf{e}_p = \begin{bmatrix} e_{1p} \\ e_{2p} \\ \vdots \\ e_{pp} \end{bmatrix}$$

- Then the principal components are formed by:

$$\begin{gathered} Y_1 = e_{11}X_1 + e_{21}X_2 + \ldots + e_{p1}X_p \\ Y_2 = e_{12}X_1 + e_{22}X_2 + \ldots + e_{p2}X_p \\ \vdots \\ Y_p = e_{1p}X_1 + e_{2p}X_2 + \ldots + e_{pp}X_p \end{gathered}$$

PCA Procedure

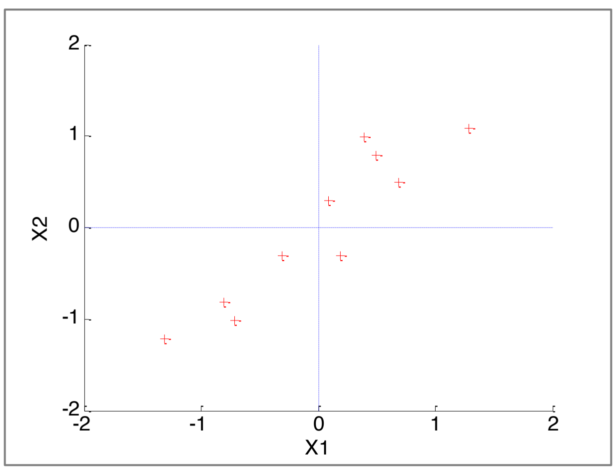
Step 5: Extract new features

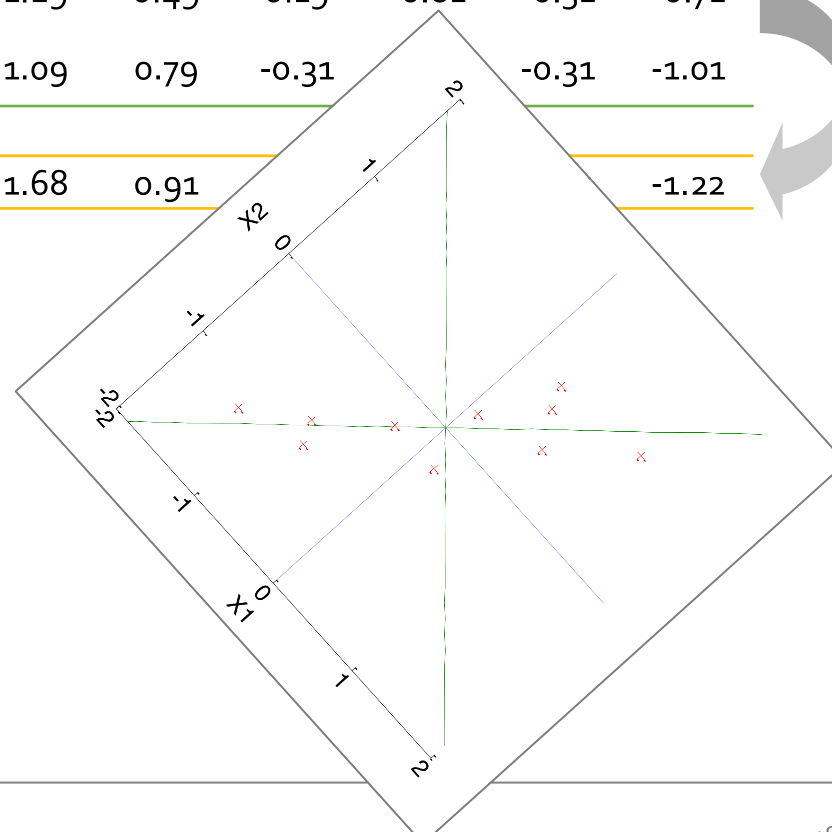
5

- Project the original data onto the selected bases

x_1	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
x_2	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31			-0.31	-1.01
z_1	0.83	-1.78	0.99	0.27	1.68	0.91				-1.22

$Z\_1=(X\_1, X\_2)\cdot(0.677, 0.735)$





PCA Procedure

PCA Example

- **원래 데이터**

시리얼 이름	제조업체명	유형	칼로리	단백질	지방	나트륨	식이섬유	복합탄수화물	설탕	칼륨	비타민
100% Bran	N	C	70	4	1	130	10	5	6	280	25
100% Natural Bran	Q	C	120	3	5	15	2	8	8	135	0
All-Bran	K	C	70	4	1	260	9	7	5	320	25
All-Bran with Extra Fiber	K	C	50	4	0	140	14	8	0	330	25
Almond Delight	R	C	110	2	2	200	1	14	8		25
Apple Cinnamon Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25
Apple Jacks	K	C	110	2	0	125	1	11	14	30	25
Basic 4	G	C	130	3	2	210	2	18	8	100	25
Bran Chex	R	C	90	2	1	200	4	15	6	125	25
Bran Flakes	P	C	90	3	0	210	5	13	5	190	25
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25
Cheerios	G	C	110	6	2	290	2	17	1	105	25
Cinnamon Toast Crunch	G	C	120	1	3	210	0	13	9	45	25
Clusters	G	C	110	3	2	140	2	13	7	105	25
Cocoa Puffs	G	C	110	1	1	180	0	12	13	55	25
Corn Chex	R	C	110	2	0	280	0	22	3	25	25
Corn Flakes	K	C	100	2	0	290	1	21	2	35	25
Corn Pops	K	C	110	1	0	90	1	13	12	20	25
Count Chocula	G	C	110	1	1	180	0	12	13	65	25
Cracklin' Oat Bran	K	C	110	3	3	140	4	10	7	160	25

6

PCA Procedure

PCA Example

- Eigenvectors eigenvalues for each principal component

6

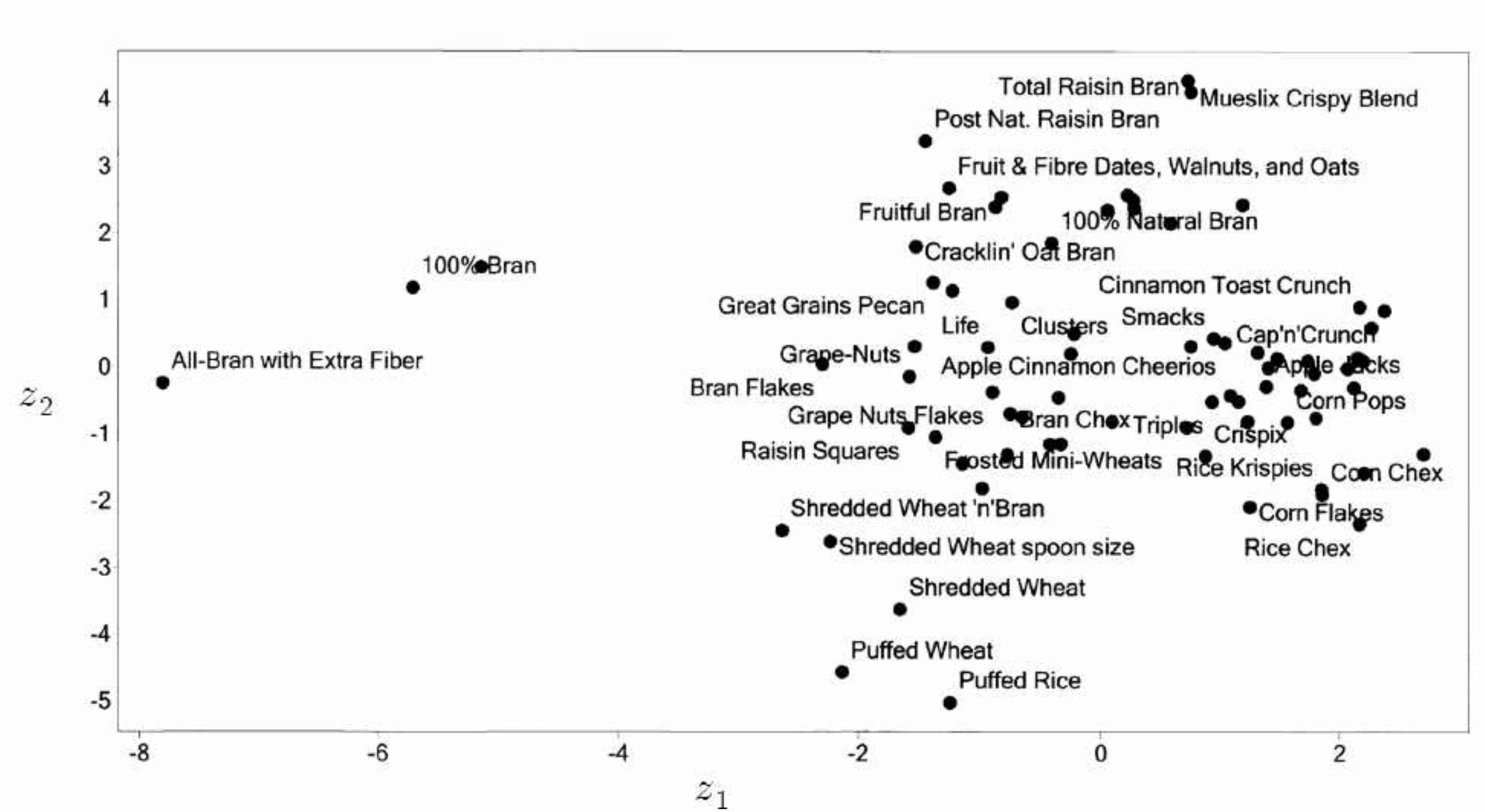
변수이름	1	2	3	4	5	6	7
calories	0.2995424	0.39314792	0.11485746	0.20435865	0.20389892	-0.25590625	-0.02559552
protein	-0.30735639	0.16532333	0.27728197	0.30074316	0.319749	0.120752	0.28270504
fat	0.03991544	0.34572428	-0.20489009	0.18683317	0.58689332	0.34796733	-0.05115468
sodium	0.18339655	0.13722059	0.38943109	0.12033724	-0.33836424	0.66437215	-0.28370309
fiber	-0.45349041	0.17981192	0.06976604	0.03917367	-0.255119	0.0642436	0.11232537
carbo	0.19244903	-0.14944831	0.56245244	0.0878355	0.18274252	-0.32639283	-0.26046798
sugars	0.22806853	0.35143444	-0.35540518	-0.02270711	-0.31487244	-0.15208226	0.22798519
potass	-0.40196434	0.30054429	0.06762024	0.09087842	-0.14836049	0.02515389	0.14880823
vitamins	0.11598022	0.1729092	0.38785872	-0.6041106	-0.04928682	0.12948574	0.29427618
shelf	-0.17126338	0.26505029	-0.00153102	-0.63887852	0.32910112	-0.05204415	-0.17483434
weight	0.05029929	0.45030847	0.24713831	0.15342878	-0.22128329	-0.39877367	0.01392053
cups	0.29463556	-0.21224795	0.13999969	0.04748911	0.12081645	0.09946091	0.74856687
rating	-0.43837839	-0.25153893	0.1818424	0.0383162	0.05758421	-0.18614525	0.06344455

분산	3.63360572	3.1480546	1.90934956	1.01947618	0.98935974	0.72206175	0.67151642
분산비(%)	27.95081329	24.21580505	14.6873045	7.84212446	7.61045933	5.55432129	5.16551113
누적분산비(%)	27.95081329	52.16661835	66.85391998	74.69604492	82.3065033	87.86082458	93.02633667

PCA Procedure

PCA Example

- In a 2-dimensional space



6

PCA Procedure

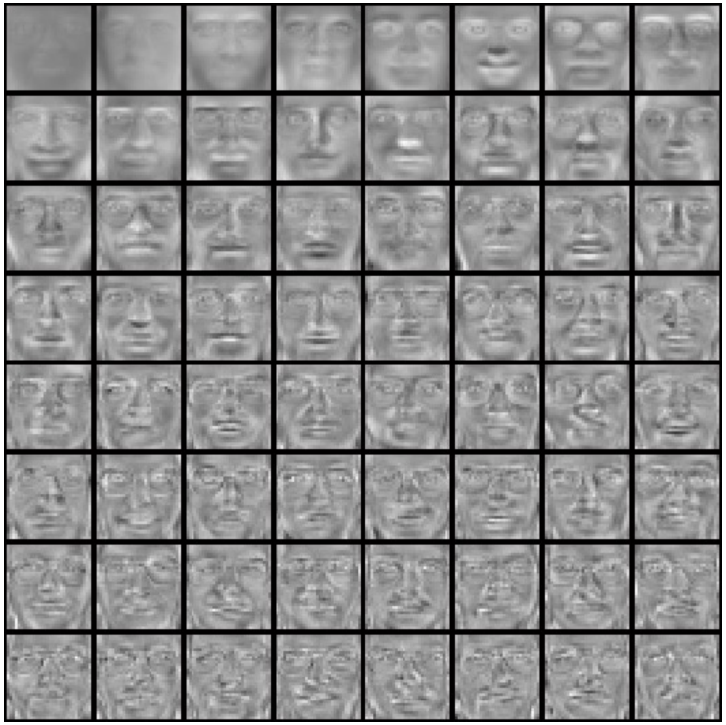
6

PCA Example: Face Image Data

Average

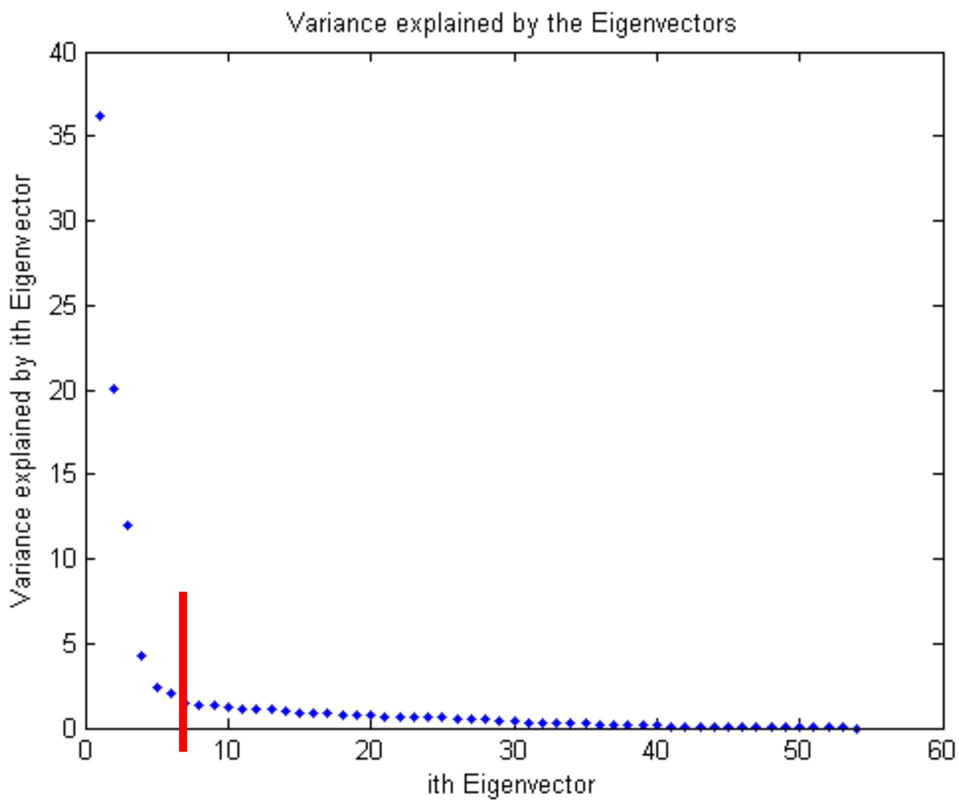
Original Image

Eigenvectors



PCA Procedure

- **How to determine the optimal number of features**
  - Use the scree plot
  - In general, select the principal component around the saddle point



Limitation

- **May not perform well for some classification task**
  - Lose the class information by preserving the variance as much as possible