Page-1

Dimension Reduction using Principal Components Analysis (PCA)

Page-2

Application of dimension reduction

- Computational advantage for other algorithms
- Face recognition— image data (pixels) along new axes works better for recognizing faces
- Image compression

Page-3

Data for 25 undergraduate programs at business schools in US universities in 1995.

Use PCA to:

1) Reduce # columns Additional benefits:

2) Identify relation between columns

3) Visualize universities in 2D

Univ	SAT	Top10	Accept	SFRatio	Expenses	GradRate
Brown	1310	89	22	13	22,704	94
CalTech	1415	100	25	6	63,575	81
CMU	1260	62	59	9	25,026	72
Columbia	1310	76	24	12	31,510	88
Cornell	1280	83	33	13	21,864	90
Dartmouth	1340	89	23	10	32,162	95
Duke	1315	90	30	12	31,585	95
Georgetown	1255	74	24	12	20,126	92
Harvard	1400	91	14	11	39,525	97
JohnsHopkins	1305	75	44	7	58,691	87
MIT	1380	94	30	10	34,870	91
Northwestern	1260	85	39	11	28,052	89
NotreDame	1255	81	42	13	15,122	94
PennState	1081	38	54	18	10,185	80
Princeton	1375	91	14	8	30,220	95
Purdue	1005	28	90	19	9,066	69
Stanford	1360	90	20	12	36,450	93
TexasA&M	1075	49	67	25	8,704	67
UCBerkeley	1240	95	40	17	15,140	78
UChicago	1290	75	50	13	38,380	87
UMichigan	1180	65	68	16	15,470	85
UPenn	1285	80	36	11	27,553	90
UVA	1225	77	44	14	13,349	92
UWisconsin	1085	40	69	15	11,857	71
Yale	1375	95	19	11	43,514	96

PAGE-4

(PCA)

Input → Output

Univ	SAT	Top10	Accept	SFRatio	Expenses	GradRate
Brown	1310	89	22	13	22,704	94
CalTech	1415	100	25	6	63,575	81
CMU	1260	62	59	9	25,026	72
Columbia	1310	76	24	12	31,510	88
Cornell	1280	83	33	13	21,864	90
Dartmouth	1340	89	23	10	32,162	95
Duke	1315	90	30	12	31,585	95
Georgetown	1255	74	24	12	20,126	92
Harvard	1400	91	14	11	39,525	97
JohnsHopkins	1305	75	44	7	58,691	87

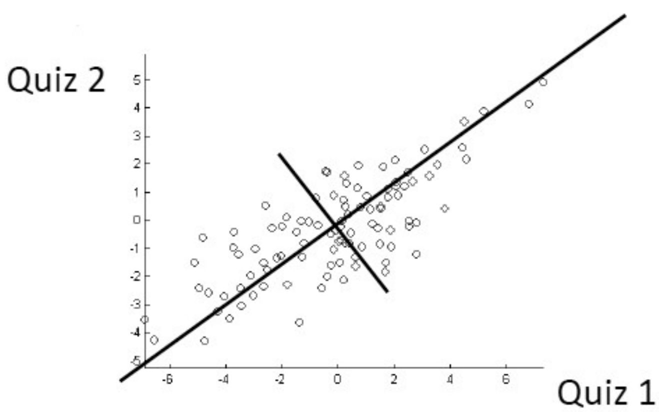
PC_1	PC_2	PC_3	PC_4	PC_5	PC_6

Hope is that a fewer columns may capture most of the information from the original dataset

Reduce the number of columns to fewer columns so that those fewer column will capture most of the information from the original data set

PAGE-5

The Primitive Idea – Intuition First



How to compress the data losing the least amount of information?

PAGE-6

Input == PCA ==>	Output
• p measurements/ original columns • Correlated	• p principal components (= p weighted averages of original measurements) • Uncorrelated • Ordered by variance • Keep top principal components; drop rest

PAGE-7

Mechanism

Univ	SAT	Top10	Accept	SFRatio	Expenses	GradRate
Brown	1310	89	22	13	22,704	94
CalTech	1415	100	25	6	63,575	81
CMU	1260	62	59	9	25,026	72
Columbia	1310	76	24	12	31,510	88
Cornell	1280	83	33	13	21,864	90
Dartmouth	1340	89	23	10	32,162	95
Duke	1315	90	30	12	31,585	95
Georgetown	1255	74	24	12	20,126	92
Harvard	1400	91	14	11	39,525	97
JohnsHopkins	1305	75	44	7	58,691	87

PC_1	PC_2	PC_3	PC_4	PC_5	PC_6

The ith principal component is a weighted average of original measurements/columns:

$PC_i = a_{i1}X_1 + a_{i2}X_2 + \ldots\ldots + a_{ip}X_p$

Weights (aij) are chosen such that:

1. PCs are ordered by their variance (PC1 has largest variance, followed by PC2, PC3, and so on)
2. Pairs of PCs have correlation = 0
3. For each PC, sum of squared weights =1

PAGE-8

$PC_i = a_{i1}X_1 + a_{i2}X_2 + \ldots\ldots + a_{ip}X_p$

Demystifying weight computation:

Main idea: high variance = lots of information

$Var(PC_i) = a_{i1}^2 Var(X_1) + a_{i2}^2 Var(X_2) + \ldots + a_{ip}^2 Var(X_p) + 2 a_{i1} a_{i2} Cov(X_1, X_2) + \ldots + 2 a_{ip-1} a_{ip} Cov(X_{p-1}, X_p)$

Also want, $CoVar(PC_i, PC_j) = 0$ when $i \neq j$

• Goal: Find weights aij that maximize variance of PCi, while keeping PCi uncorrelated to other PCs.

• The covariance matrix of the X's is needed.

PAGE-9

Standardize the inputs:

Why?

• variables with large variances will have bigger influence on result

Solution

• Standardize before applying PCA

Univ	Z_SAT	Z_Top10	Z_Accept	Z_SFRatio	Z_Expenses	Z_GradRate
Brown	0.4020	0.6442	-0.8719	0.0688	-0.3247	0.8037
CalTech	1.3710	1.2103	-0.7198	-1.6522	2.5087	-0.6315
CMU	-0.0594	-0.7451	1.0037	-0.9146	-0.1637	-1.6251
Columbia	0.4020	-0.0247	-0.7705	-0.1770	0.2858	0.1413
Cornell	0.1251	0.3355	-0.3143	0.0688	-0.3829	0.3621
Dartmouth	0.6788	0.6442	-0.8212	-0.6687	0.3310	0.9141
Duke	0.4481	0.6957	-0.4664	-0.1770	0.2910	0.9141
Georgetown	-0.1056	-0.1276	-0.7705	-0.1770	-0.5034	0.5829
Harvard	1.2326	0.7471	-1.2774	-0.4229	0.8414	1.1349
JohnsHopkins	0.3559	-0.0762	0.2433	-1.4063	2.1701	0.0309
MIT	1.0480	0.9015	-0.4664	-0.6687	0.5187	0.4725
Northwestern	-0.0594	0.4384	-0.0101	-0.4229	0.0460	0.2517
NotreDame	-0.1056	0.2326	0.1419	0.0688	-0.8503	0.8037
PennState	-1.7113	-1.9800	0.7502	1.2981	-1.1926	-0.7419
Princeton	1.0018	0.7471	-1.2774	-1.1605	0.1963	0.9141
Purdue	-2.4127	-2.4946	2.5751	1.5440	-1.2702	-1.9563
Stanford	0.8634	0.6957	-0.9733	-0.1770	0.6282	0.6933
TexasA&M	-1.7667	-1.4140	1.4092	3.0192	-1.2953	-2.1771
UCBerkeley	-0.2440	0.9530	0.0406	1.0523	-0.8491	-0.9627
UChicago	0.2174	-0.0762	0.5475	0.0688	0.7620	0.0309
UMichigan	-0.7977	-0.5907	1.4599	0.8064	-0.8262	-0.1899
UPenn	0.1713	0.1811	-0.1622	-0.4229	0.0114	0.3621
UVA	-0.3824	0.0268	0.2433	0.3147	-0.9732	0.5829
UWisconsin	-1.6744	-1.8771	1.5106	0.5606	-1.0767	-1.7355
Yale	1.0018	0.9530	-1.0240	-0.4229	1.1179	1.0245

Excel: =standardize(cell, average(column), stdev(column))

PAGE-10

Standardization shortcut for PCA

• Rather than standardize the data manually, you can use correlation matrix instead of covariance matrix as input s different results!

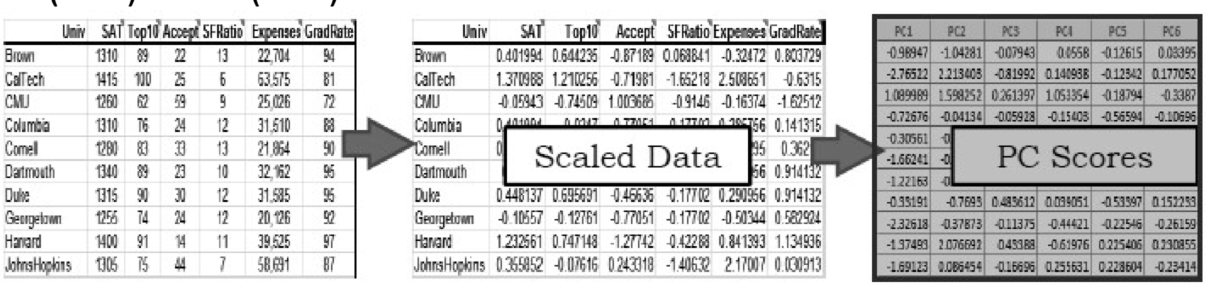
Univ	SAT	Top10	Accept	SFRatio	Expenses	GradRate
Brown	0.401994	0.644235	-0.87189	0.068841	-0.32472	0.803729
CalTech	1.370988	1.210256	-0.71981	-1.65218	2.508651	-0.6315
CMU	-0.05943	-0.74509	1.003685	-0.9146	-0.16374	-1.62512
Columbia	0.401994	-0.0247	-0.77051	-0.17702	0.285756	0.141315
Cornell	0.125139	0.335496	-0.31429	0.068841	-0.38295	0.36212
Dartmouth	0.67885	0.644235	-0.8212	-0.66874	0.330956	0.914132
Duke	0.448137	0.695691	-0.46636	-0.17702	0.290956	0.914132
Georgetown	-0.10557	-0.12761	-0.77051	-0.17702	-0.50344	0.582924
Harvard	1.232561	0.747148	-1.27742	-0.42288	0.841393	1.134936
JohnsHopkins	0.355852	-0.07616	0.243318	-1.40632	2.17007	0.030913

						6
SAT	-0.45775	0.03968	-0.18704	0.13124	0.020646	-0.85805
Top10	-0.42714	-0.19993	-0.49781	0.374896	0.482016	0.396075
Accept	0.424308	0.320893	0.156279	0.061287	0.801094	-0.21693
SFRatio	0.390648	-0.43256	-0.60608	-0.50739	0.076824	-0.17205
Expenses	-0.36252	0.634486	-0.20474	-0.6234	0.072548	0.173763
GradRate	-0.3794	-0.51555	0.532473	-0.43863	0.33811	0.003538

Variances						
	1	2	3	4	5	6
Variance	4.612085	0.786816	0.286562	0.16378	0.124306	0.026451
Variance Pe	76.86808	13.1136	4.776031	2.729668	2.07177	0.440844
Cumulative	76.86808	89.98169	94.75772	97.48739	99.55916	100

PCs are uncorrelated

- Var(PC1) > Var (PC2) > ...



$PC_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$

PAGE-12

Computing principal scores

- For each record, we can compute their score on each PC.
- Multiply each weight (aij) by the appropriate Xij
- Example for Brown University (using Standardized numbers):

• PC Score1 for Brown University = (–0.458) (0.40) + (–0.427) (.64) + (0.424) (–0.87) + (0.391) (.07) + (–0.363) (–0.32) + (–0.379) (.80) = –0.989

PAGE-13

R Code for PCA (Assignment)

OPTIONAL R Code

```
install.packages("gdata") ## for reading xls files
install.packages("xlsx") ## ” for reading xlsx files
mydata<-read.xlsx("University Ranking.xlsx",1) ## use read.csv for csv files
mydata ## make sure the data is loaded correctly
help(princomp) ## to understand the api for princomp
pcaObj<-princomp(mydata[1:25,2:7], cor = TRUE, scores = TRUE, covmat = NULL)
## the first column in mydata has university names
## princomp(mydata, cor = TRUE) not_same_as prcomp(mydata, scale=TRUE); similar, but different
summary(pcaObj)
loadings(pcaObj)
plot(pcaObj)
biplot(pcaObj)
pcaObj$loadings
pcaObj$scores
```

PAGE-14

Goal #1: Reduce data dimension

• PCs are ordered by their variance (=information)

• Choose top few PCs and drop the rest!

Example:

• PC1 captures most 76.86% of the information.

• The first 2 PCs capture 89.98%

• Data reduction: use only two variables instead of 6

Principal Components						
Feature\Cc	1	2	3	4	5	6
SAT	-0.45775	0.03968	-0.18704	0.13124	0.020646	-0.85805
Top10	-0.42714	-0.19993	-0.49781	0.374896	0.482016	0.396075
Accept	0.424308	0.320893	0.156279	0.061287	0.801094	-0.21693
SFRatio	0.390648	-0.43256	-0.60608	-0.50739	0.076824	-0.17205
Expenses	-0.36252	0.634486	-0.20474	-0.6234	0.072548	0.173763
GradRate	-0.3794	-0.51555	0.532473	-0.43863	0.33811	0.003538

Variances						
	1	2	3	4	5	6
Variance	4.612085	0.786816	0.286562	0.16378	0.124306	0.026451
Variance Pe	76.86808	13.1136	4.776031	2.729668	2.07177	0.440844
Cumulative	76.86808	89.98169	94.75772	97.48739	99.55916	100

PAGE-15

Matrix Transpose

$$\begin{bmatrix} 1 & 2 \end{bmatrix}^T = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

OPTIONAL: R code

```
help(matrix)
A<-matrix(c(1,2),nrow=1,ncol=2,byrow=TRUE)
A
t(A)
B<-matrix(c(1,2,3,4),nrow=2,ncol=2,byrow=TRUE)
B
t(B)
C<matrix(c(1,2,3,4,5,6),nrow=3,ncol=2,byrow=TRUE)
C
t(C)
```

PAGE-16

Matrix Multiplication:

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mp} \end{pmatrix}$$

$$\mathbf{AB} = \begin{pmatrix} (AB)_{11} & (AB)_{12} & \cdots & (AB)_{1p} \\ (AB)_{21} & (AB)_{22} & \cdots & (AB)_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ (AB)_{n1} & (AB)_{n2} & \cdots & (AB)_{np} \end{pmatrix} \quad (AB)_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}.$$

$$A_{3x2} \cdot B_{2x4} = C_{3x4} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix} =$$

$$= \begin{pmatrix} 1\cdot1+2\cdot5 & 1\cdot2+2\cdot6 & 1\cdot3+2\cdot7 & 1\cdot4+2\cdot8 \\ 3\cdot1+4\cdot5 & 3\cdot2+4\cdot6 & 3\cdot3+4\cdot7 & 3\cdot4+4\cdot8 \\ 5\cdot1+6\cdot5 & 5\cdot2+6\cdot6 & 5\cdot3+6\cdot7 & 5\cdot4+6\cdot8 \end{pmatrix} = \begin{pmatrix} 11 & 14 & 17 & 20 \\ 23 & 30 & 37 & 44 \\ 35 & 46 & 57 & 68 \end{pmatrix}$$

$$AB \neq BA$$

OPTIONAL R Code

```
A<-
matrix(c(1,2,3,4,5,6),nrow=3,ncol=2,byrow=
TRUE)
A
B<-
matrix(c(1,2,3,4,5,6,7,8),nrow=2,ncol=4,byro
w=TRUE)
B
C<-A%*%B
D<-t(B)%*%t(A) ## note, B%*%A is not
possible;
how does D look like?
AB ≠BA
```

PAGE-17

Matrix Inverse:

If, $A B I$,identity matrix, Then, B = A^{-1}

$$\begin{bmatrix} 1\,0 \dots 0\,0 \\ 0\,1 \dots 0\,0 \\ \dots\dots\dots \\ 0\,0 \dots 1\,0 \\ 0\,0 \dots 0\,1 \end{bmatrix}$$

OPTIONAL R Code

```
## How to create nxn
Identity matrix?
help(diag)
A<-diag(5)
## find inverse of a matrix
solve(A)
```

PAGE-18

Data Compression:

$[\text{PCScores}]_{Nxp} = [\text{ScaledData}]_{Nxp} \text{ x } [\text{PrincipalComponents}]_{pxp}$

$[\text{ScaledData}]_{Nxp} = [\text{PCScores}]_{Nxp} \text{ x } [\text{PrincipalComponents}]^{-1}{}_{pxp}$

$= [\text{PCScores}]_{Nxp} \text{ x } [\text{PrincipalComponents}]^{T}{}_{pxp}$

Approximation:

$[\text{ApproximatedScaledData}]_{Nxp} = [\text{PCScore}]_{Nxc} \text{ X }$
$[\text{PrincipalComponent}]^{T}{}_{cxp}$

c = Number of components kept; c <=p

PAGE-19

Goal #2: Learn relationships with PCA by interpreting the weights

- ai1,…, aip are the coefficients for PCi.
- They describe the role of original X variables in computing PCi.
- Useful in providing context-specific interpretation of each PC.

Principal Components						
Feature\Co	1	2	3	4	5	6
SAT	-0.45775	0.03968	-0.18704	0.13124	0.020646	-0.85805
Top10	-0.42714	-0.19993	-0.49781	0.374896	0.482016	0.396075
Accept	0.424308	0.320893	0.156279	0.061287	0.801094	-0.21693
SFRatio	0.390648	-0.43256	-0.60608	-0.50739	0.076824	-0.17205
Expenses	-0.36252	0.634486	-0.20474	-0.6234	0.072548	0.173763
GradRate	-0.3794	-0.51555	0.532473	-0.43863	0.33811	0.003538

Variances						
	1	2	3	4	5	6
Variance	4.612085	0.786816	0.286562	0.16378	0.124306	0.026451
Variance Pe	76.86808	13.1136	4.776031	2.729668	2.07177	0.440844
Cumulative	76.86808	89.98169	94.75772	97.48739	99.55916	100

PAGE-20

PC1 Scores
(choose one or more)

Feature\Co	1	2
SAT	-0.45775	0.03968
Top10	-0.42714	-0.19993
Accept	0.424308	0.320893
SFRatio	0.390648	-0.43256
Expenses	-0.36252	0.634486
GradRate	-0.3794	-0.51555

1. Are approximately a simple average of the 6 variables
2. Measure the degree of high Accept & SFRatio, but low Expenses, GradRate, SAT, and Top10

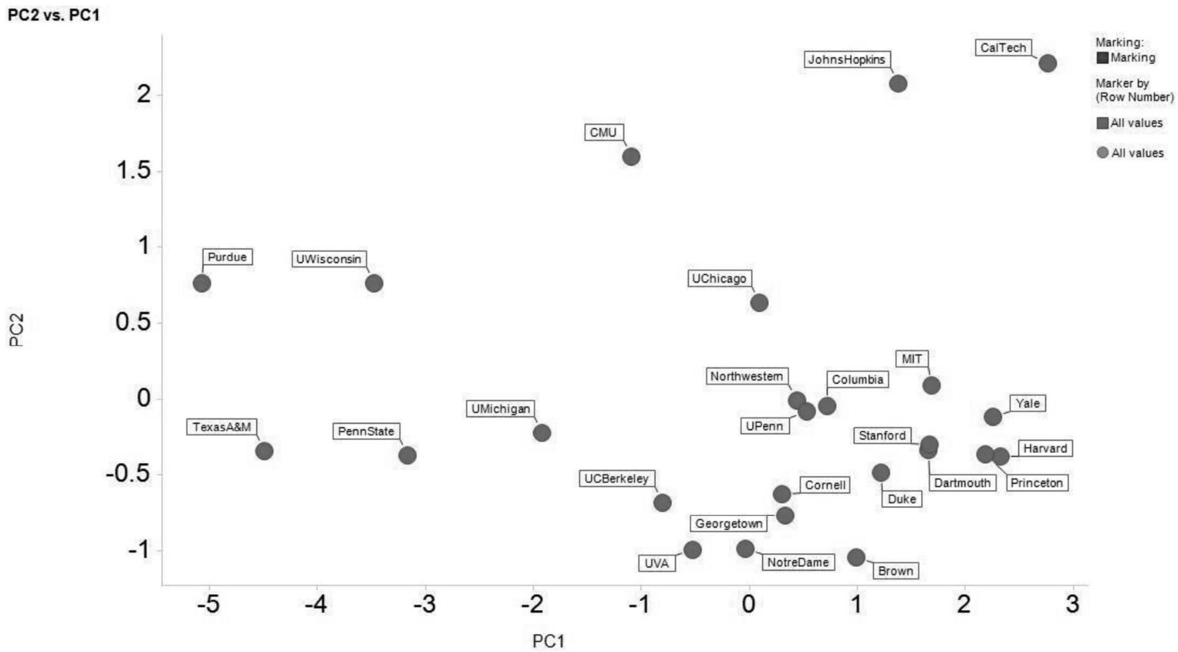
PAGE-21

Goal #3: Use PCA for visualization

• The first 2 (or 3) PCs provide a way to project the data from a p-dimensional space onto a 2D (or 3D) space

PAGE-22

Scatter Plot: PC2 vs. PC1 scores



PAGE-23

Monitoring batch processes using PCA

- Multivariate data at different time points
- Historical database of successful batches are used
- Multivariate trajectory data is projected to low-dimensional space

>>> Simple monitoring charts to spot outlier

PAGE-24

Your Turn!

1. If we use a subset of the principal components, is this useful for prediction? for explanation?
2. What are advantages and weaknesses of PCA compared to choosing a subset of the variables?
3. PCA vs. Clustering