Multivariate Data Analysis

(MGT513, BAT531, TIM711)

Lecture 2

UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
2009

Chapter 4

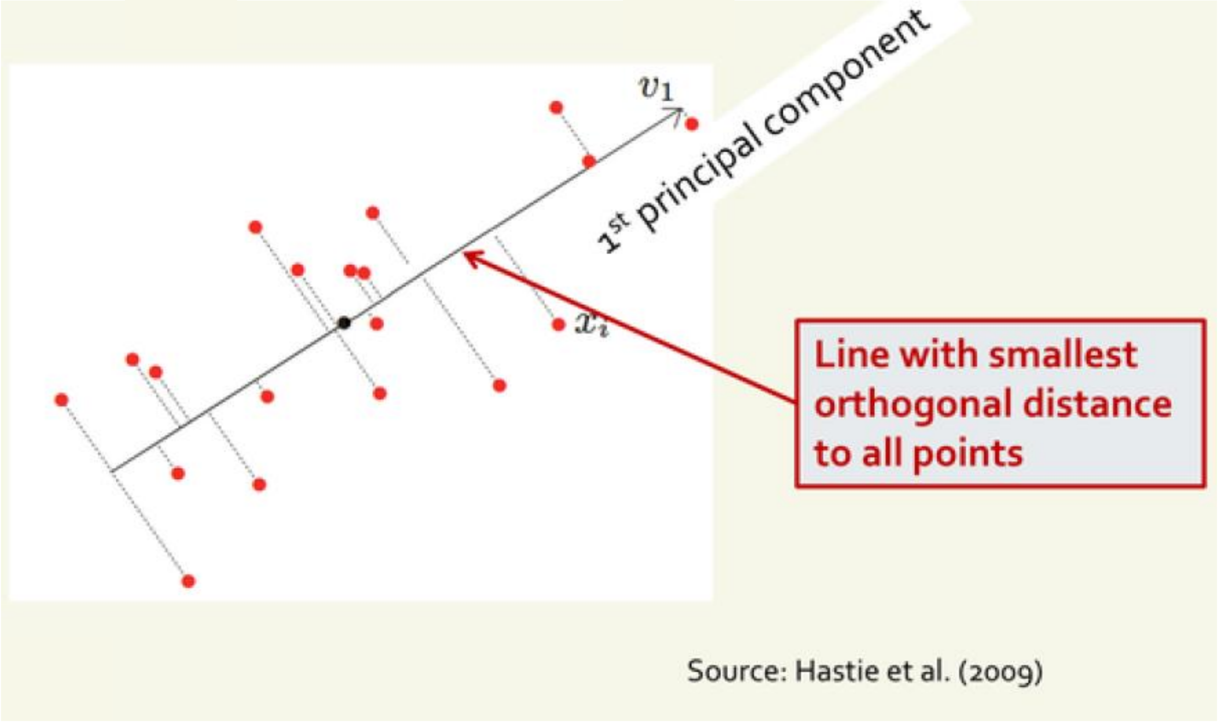
Principal Component Analysis (PCA)

Intuition

- Principal Components Analysis (***PCA***) is the method to find a linear combination of the original variables with maximum variance to extract most information from data
  - The first Principal Component ($\boldsymbol{PC_1}$) has the largest variance and $\boldsymbol{PC_2}$ has the second largest variance and so forth.
  - Principal Components are orthogonal to each other

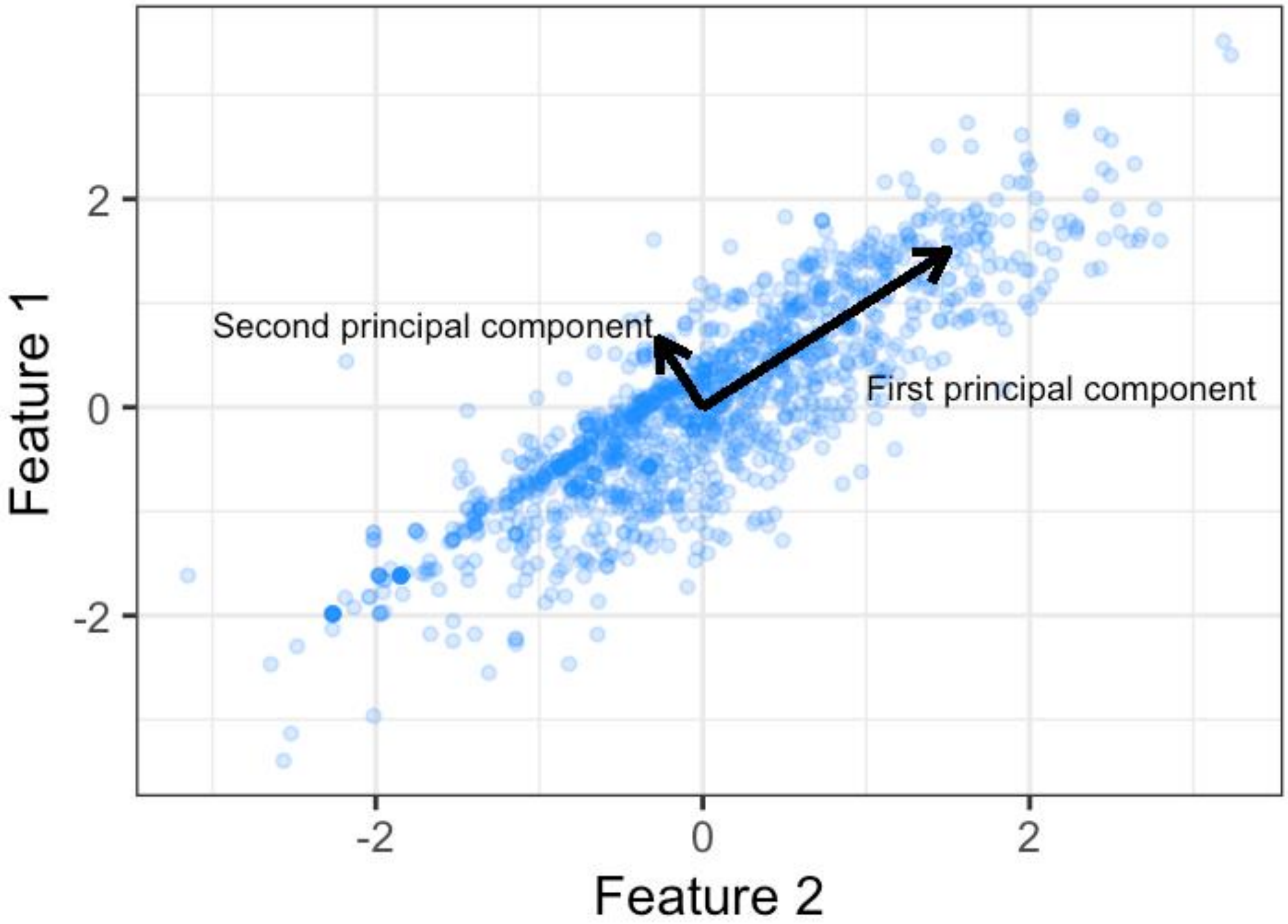
2D illustration

- The first Principal Component ($\boldsymbol{PC_1}$) is the direction with maximum variance. It is the direction that minimizes the sum of squared residuals.



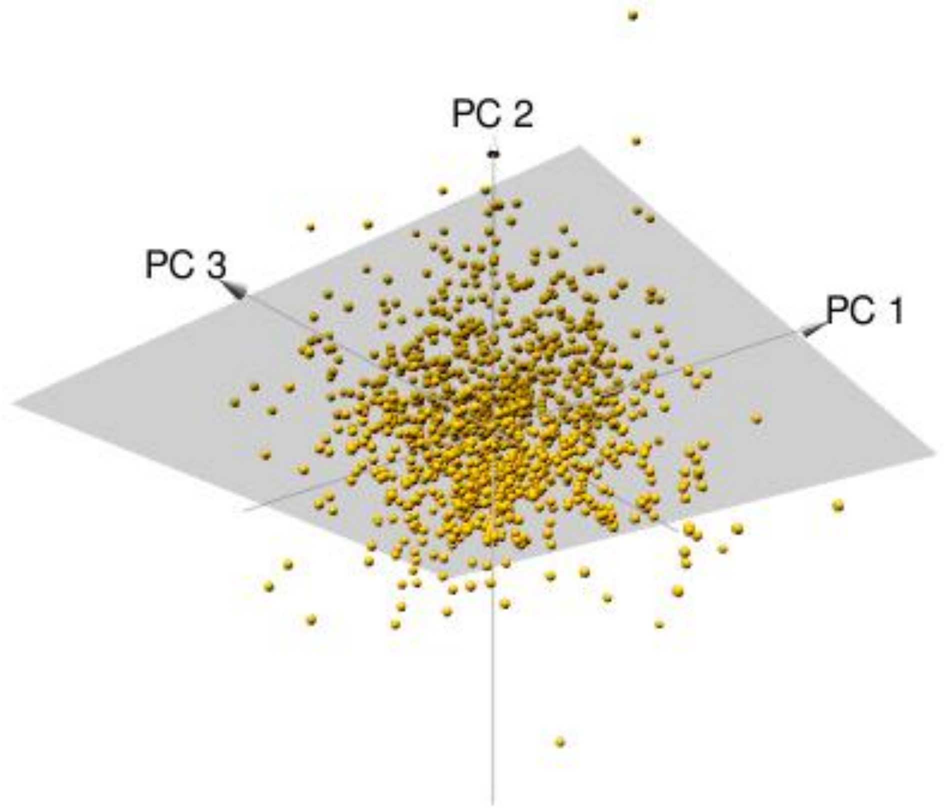
Source: Fabio Sigrist

2D illustration



source:https://bradleyboehmke.github.io/HOML/pca.html

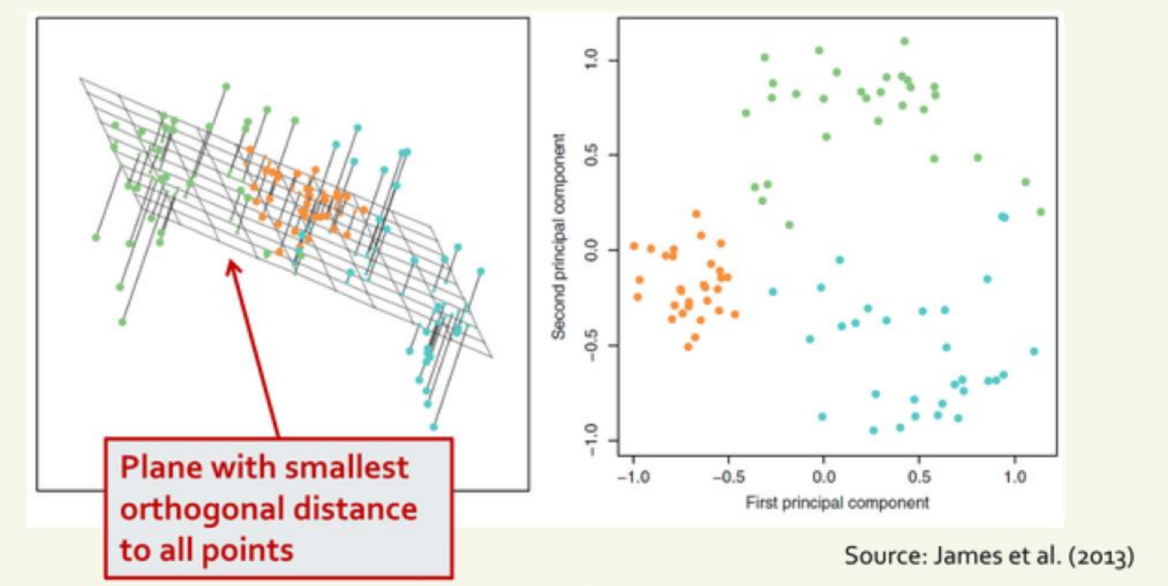
3D illustration



source:https://bradleyboehmke.github.io/HOML/pca.html

3D illustration

- The 1st and the 2nd Principal Components ($\boldsymbol{PC_1}, \boldsymbol{PC_2}$) form a plane such that projection on the plane yields maximum variance. Equivalently, they form a plane such that sum of squared residuals (orthogonal to the plane) is minimal.



Source: Fabio Sigrist

Goal: Dimension Reduction of Data

- Re-Expressing Data
- Lower dimensions of data account for as much as information of the original data

Potential Applications: Example

- *Example : Dimension Reduction* Individual Need for Cognition

Potential Applications: Example

- Table 4.1: 18 question for cognition

TABLE 4.1 Eighteen items used in measuring a survey respondent's "need for cognition"

Item	Response
C_1	I prefer complex to simple problems.
C_2	I like to have the responsibility of handling a situation that requires a lot of thinking.
C_3	Thinking is not my idea of fun. (R)
C_4	I would rather do something requiring little thought than something that is sure to challenge my thinking abilities. (R)
C_5	I try to anticipate and avoid situations where there is a likely chance that I will have to think in depth about something. (R)
C_6	I find satisfaction in deliberating hard for long hours.
C_7	I only think as hard as I have to. (R)
C_8	I prefer to think about small daily projects to long-term ones. (R)
C_9	I like tasks that require little thought once I've learned them. (R)

Potential Applications: Example

- Table 4.1: 18 question for cognition

C_{10}	The idea of relying on thought to make my way to the top appeals to me.
C_{11}	I really enjoy a task that involves coming up with new solutions to problems.
C_{12}	Learning new ways to think doesn't excite me much. (R)
C_{13}	I prefer my life to be filled with puzzles that I must solve.
C_{14}	The notion of thinking abstractly is appealing to me.
C_{15}	I prefer tasks that are intellectual, difficult, and important to ones that do not require much thought.
C_{16}	I feel relief rather than satisfaction after completing a task that required a lot of mental effort. (R)
C_{17}	It's enough for me that something gets the job done; I don't care how or why it works. (R)
C_{18}	I usually end up deliberating about issues even when they do not affect me personally.

(R) indicates a reverse-coded item.
Source: Cacioppo, Petty, and Kao (1984).

Potential Applications: Example

- Table 4.2: Correlation matrix

TABLE 4.2 Correlation matrix for 18 items measuring need for cognition (n = 201)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}
C_2	0.445																
C_3	−0.239	−0.454															
C_4	−0.270	−0.375	0.365														
C_5	−0.326	−0.487	0.421	0.558													
C_6	0.248	0.274	−0.187	−0.187	−0.250												
C_7	−0.268	−0.338	0.319	0.345	0.343	−0.235											
C_8	−0.270	−0.355	0.228	0.306	0.340	−0.165	0.314										
C_9	−0.320	−0.328	0.389	0.415	0.310	−0.174	0.221	0.312									
C_{10}	0.289	0.375	−0.411	−0.338	−0.336	0.279	−0.268	−0.299	−0.380								
C_{11}	0.364	0.516	−0.325	−0.405	−0.384	0.297	−0.262	−0.148	−0.338	0.520							
C_{12}	−0.366	−0.415	0.258	0.373	0.555	−0.275	0.310	0.132	0.245	−0.261	−0.392						
C_{13}	0.393	0.382	−0.245	−0.288	−0.322	0.220	−0.172	−0.245	−0.250	0.321	0.418	−0.350					
C_{14}	0.341	0.376	−0.257	−0.250	−0.295	0.291	−0.231	−0.331	−0.145	0.338	0.310	−0.397	0.454				
C_{15}	0.268	0.354	−0.228	−0.185	−0.165	0.186	−0.066	−0.181	−0.177	0.223	0.236	−0.199	0.274	0.230			
C_{16}	−0.280	−0.192	0.166	0.320	0.242	−0.155	0.212	0.150	0.214	−0.147	−0.242	0.192	−0.058	−0.132	−0.071		
C_{17}	−0.273	−0.425	0.282	0.412	0.332	−0.232	0.251	0.364	0.373	−0.226	−0.361	0.284	−0.114	−0.090	−0.174	0.331	
C_{18}	0.126	0.166	−0.140	−0.069	−0.105	0.162	−0.131	−0.042	−0.007	−0.021	0.100	−0.119	0.029	0.209	0.084	−0.194	−0.087

Potential Applications: Example

- Table 4.3: the first PC

TABLE 4.3 Results from principal components analysis of need for cognition data in Table 4.2

	$\mathbf{u}_1$		$\mathbf{u}_1$
C_1	0.251	C_{10}	0.259
C_2	0.309	C_{11}	0.282
C_3	−0.253	C_{12}	−0.259
C_4	−0.275	C_{13}	0.232
C_5	−0.289	C_{14}	0.230
C_6	0.183	C_{15}	0.169
C_7	−0.227	C_{16}	−0.164
C_8	−0.206	C_{17}	−0.229
C_9	−0.234	C_{18}	0.087

Eigenvalue $\lambda_1 = 5.7794$
Proportion of variance accounted for 32.1 percent.

Definition of PCA

- Model Setup
  - n: number of observations
  - p: number of variables (number of columns in $\mathbf{X}$)
  - $\mathbf{X} = (X_1, X_2, \cdots, X_p)$: *n by p matrix of variables* where $X_i = \begin{pmatrix} \mathrm{X}_{1i} \\ \vdots \\ \mathrm{X}_{ni} \end{pmatrix}$ $(i$=1,...,$p)$
  - **U**= $(\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_p})$: *p by p* matrix of eigenvectors where $\mathbf{u}_{\boldsymbol{i}} = \begin{pmatrix} u_{1i} \\ \vdots \\ u_{pi} \end{pmatrix}$

    $(i$=1,...,$p)$
  - $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_p)$: the i^{th}principal component $Z_i = \begin{pmatrix} Z_{1i} \\ \vdots \\ Z_{ni} \end{pmatrix}$ $(i$=1,...,$p)$
- In a matrix notation: **Z=XU**

Definition of PCA

- Linear Combinations
  - The i^{th}principal component Z_i is the *normalized* linear combination

$$
\begin{aligned}
Z_i &= \mathbf{X}\mathbf{u}_{\boldsymbol{i}} \\
&= u_{1i}X_1 + u_{2i}X_2 + \cdots + u_{pi}X_p
\end{aligned}
$$

  - With n data points

$$
\begin{pmatrix} z_{1\mathrm{i}} \\ \vdots \\ z_{n\mathrm{i}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = u_{1i}\begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} + u_{2i}\begin{pmatrix} x_{12} \\ \vdots \\ x_{n2} \end{pmatrix} + \cdots + u_{pi}\begin{pmatrix} x_{1p} \\ \vdots \\ x_{np} \end{pmatrix}
$$

- Question: How should we find $\mathbf{u}_{\boldsymbol{i}}$?

Definition of PCA

1. Normalization of $\mathbf{u}_{\boldsymbol{i}}$ <u>(must)</u>:

$$\sum_{j=1}^{p} u_{ij}^2 = 1$$

   - Since principal components are uncorrelated:

$$\mathbf{u}_{\boldsymbol{l}}^{\boldsymbol{T}} \mathbf{u}_{\boldsymbol{m}} = \begin{cases} 1 \text{ if } l = m \\ 0 \text{ if } l \neq m \end{cases}$$

2. Mean centering of X_i <u>(must):</u>
   - Make the mean of all the variables to be zero ($E(X_i) = 0\ for\ all\ i$)
   - For each variable X_i, subtract the mean from the raw values

Definition of PCA

3. Scaling of X_i (optional):
   - Make the variance of all the variables to be 1 $(Var(X_i) = 1 \text{ for } all\ i)$
   - For each variable X_i, divide the raw values by the sample standard deviation
   - Scaling decision
     1. Different measurement units: Need to scale data
     2. Same measurement units: Not necessary
   - Mean centering + Scaling = Standardization
     - After standardization, sample correlation matrix ($\mathbf{R}$) *can be used instead of sample covariance matrix* (**C**) [as in LCG]

Computation of PCA

- Goal: To Find A Linear combination of the Original Variables with Maximum Variance

  - Let's find $\mathbf{U}$ to maximize the variance of $\mathbf{Z} = \mathbf{XU}$ for

$$\mathrm{cov}(Z) = \frac{1}{n-1}\mathbf{U}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{XU} = \mathbf{U}^{\mathbf{T}}\mathbf{CU}$$

where $\mathbf{C}$= $\frac{1}{n-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}$ (sample covariance matrix of $\mathbf{X}$)

Computation of PCA

- Steps for finding Principal Components
  - The first Principal Component ($\boldsymbol{PC_1}$) is the normalized linear combination ($\mathrm{Z}_1 = \mathbf{X}\mathbf{u_1}$) that maximizes $\mathrm{var}(\mathrm{Z}_1) = \mathbf{u_1}^T\mathbf{C}\mathbf{u_1}$ subject to $\mathbf{u_1}^\mathbf{T}\mathbf{u_1}$ =1
  - The second Principal Component ($\boldsymbol{PC_2}$) is the normalized linear combination ($\mathrm{Z}_2 = \mathbf{X}\mathbf{u_2}$) that maximizes $\mathrm{var}(\mathrm{Z}_2) = \mathbf{u_2}^\mathbf{T}\mathbf{C}\mathbf{u_2}$ subject to $\mathbf{u_2}^\mathbf{T}\mathbf{u_2}$ =1 and $\mathrm{cov}(\mathrm{Z}_1, \mathrm{Z}_2) = 0$

…

  - The i^{th} Principal Component ($\boldsymbol{PC_i}$) is the normalized linear combination ($\mathrm{Z_i} = \mathbf{X}\mathbf{u_i}$) that maximizes $\mathrm{var}(\mathrm{Z_i}) = \mathbf{u_i}^\mathbf{T}\mathbf{C}\mathbf{u_i}$ subject to $\mathbf{u_i}^\mathbf{T}\mathbf{u_i}$ =1 and $\mathrm{cov}(\mathrm{Z_i}, \mathrm{Z_j}) = 0$ for all *j* <*i*

Computation of PCA

- Solution: Lagrange Multiplier

$$L = \mathbf{u}^{\mathrm{T}}\mathbf{C}\mathbf{u} - \lambda\left(\mathbf{u}^{\mathrm{T}}\mathbf{u} - 1\right)$$

$$\Rightarrow \frac{\vartheta L}{\vartheta \mathbf{u}} = 2\mathbf{C}\mathbf{u} - 2\lambda\mathbf{u} = 0$$

gives

$$\mathbf{C}\mathbf{u} = \lambda\mathbf{u} \text{ or } (\mathbf{C} - \lambda \mathrm{I})\mathbf{u} = 0$$

In a linear transformation,

λ: eigenvalue

$\mathbf{u}$: eigenvector

Computation of PCA

- Spectral Decomposition (Eigen-decomposition) of $\mathbf{C}$

$\mathbf{C}$ is p by p symmetric matrix of rank p. Then there exists p by p orthogonal matrix $\mathbf{U}$ so that $\mathbf{U}^{\mathbf{T}}\mathbf{U}$= $\mathbf{U}\mathbf{U}^{\mathbf{T}}$= $\mathbf{I}_p$ and

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathbf{T}} = \sum_{i=1}^{p} \lambda_i \mathbf{u_i} \mathbf{u_i^T}$$

where

λ_i: i^{th} eigenvalue (i=1,...,p)

$\mathbf{U} = (\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_p})$: eigenvector matrix

$$\mathbf{D} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)\ \{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p\} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

Computation of PCA

- Variance of PCs
  - Let $\mathbf{Z}$=$\mathbf{XU}$, then

$$\mathrm{cov}(\mathbf{Z}) = \mathrm{cov}(\mathbf{XU})=\mathbf{U^TCU} = \mathbf{U^TUDU^TU} = \mathbf{D}$$

  - For the first Principal Component ($\boldsymbol{PC_1}$), $\mathrm{var}(\mathrm{Z}_1) = \mathrm{var}(\mathbf{Xu_1}) = \mathbf{u_1}^\mathbf{T}\mathbf{Cu_1}=\lambda_1$
  - For the second Principal Component ($\boldsymbol{PC}_2$), $\mathrm{var}(\mathrm{Z}_2) = \mathrm{var}(\mathbf{Xu_2}) = \mathbf{u_2}^\mathbf{T}\mathbf{Cu_2}=\lambda_2$
  - For the i^{th} Principal Component ($\boldsymbol{PC_i}$), $\mathrm{var}(\mathrm{Z_i}) = \mathrm{var}(\mathbf{Xu_i}) = \mathbf{u_i}^\mathbf{T}\mathbf{Cu_i}=\lambda_i$

  - Since $\mathrm{Z_i}$ are uncorrelated, for $k \leq p$

$$\mathrm{var}(\mathrm{Z}_1 + \mathrm{Z}_2 + \cdots + \mathrm{Z_k}) = \mathrm{var}(\mathrm{Z}_1) + \mathrm{var}(\mathrm{Z}_2) + \cdots + \mathrm{var}(\mathrm{Z_k}) = \lambda_1 + \lambda_2 + \cdots + \lambda_\mathrm{k}$$

For $k \leq p$,

$$0 \leq \frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} \leq 1$$

Alternative way to conduct PCA

- Singular Value Decomposition (SVD)
  - Since $\mathbf{Z}=\mathbf{XU}$, the standardized matrix of principal components is
  $$\mathbf{Z}_s = \mathbf{Z}\mathbf{D}^{-1/2} = \mathbf{X}\mathbf{U}\mathbf{D}^{-1/2}$$
  - Postmultiplying $\mathbf{D}^{1/2}$ and then $\mathbf{U}^{\mathrm{T}}$ gives
  $$\mathbf{X} = \mathbf{Z}_s\mathbf{D}^{1/2}\mathbf{U}^{\mathrm{T}}$$
  - This is SVD (Singular Value Decomposition of $\mathbf{X}$)

- PCA can be obtained by
  1) Spectral Decomposition of $\mathbf{C}$ or
  2) Singular Value Decomposition of $\mathbf{X}$

Principal Components Loadings

- Principal component scores: $\mathbf{Z}$ [similar to fitted values in regression analysis]

- Principal components loadings: *correlations* between principal component scores ($\mathbf{Z}$) and [standardized] original variable ($\mathbf{X}$)

$$\mathbf{F} = \mathrm{corr}(\mathbf{X}, \mathbf{Z}) = \frac{1}{(n-1)} \mathbf{X}^{\mathbf{T}} \mathbf{Z}_{\mathbf{s}} = \frac{1}{(n-1)} \mathbf{X}^{\mathbf{T}} \mathbf{X} \mathbf{U} \mathbf{D}^{-\mathbf{1}/\mathbf{2}} = \mathbf{U} \mathbf{D} \mathbf{U}^{\mathbf{T}} \mathbf{U} \mathbf{D}^{-\mathbf{1}/\mathbf{2}} = \mathbf{U} \mathbf{D}^{\mathbf{1}/\mathbf{2}}$$

  - Variance accounted for in variable X_i by the first c principal components:

$$\sum_{j=1}^{\mathrm{c}} f_{ij}^2$$

where f_{ij} is the correlation between X_i and Z_j from matrix $\mathbf{F}$

  - When $c = p$, $\sum_{j=1}^{\mathrm{c}} f_{ij}^2 = 1$

Principal Components Loadings

TABLE 4.7 Principal component loadings: Correlations between principal components and original data

	Principal Component Loadings		
	Z_1	Z_2	Z_3
X_1	0.9279	−0.0798	−0.3641
X_2	0.7255	0.6696	0.1590
X_3	0.8222	−0.5008	0.2706

Sample Problem: Gross State Product(GSP)

- Measured in millions of dollars for 13 industries of USA in 1996

1. agriculture, forestry, and fishing
2. mining
3. construction
4. manufacturing (durable goods)
5. manufacturing (nondurable goods)
6. transportation
7. communications
8. electricity, gas, and sanitation
9. wholesale trade
10. retail trade
11. fiduciary, insurance, and real estate
12. services
13. government

Sample Problem: Gross State Product(GSP)

- Table 4.9:Raw Data – Highly correlated

TABLE 4.9 Correlations among 13 different measures of economic activity (raw data in millions of dollars)

	AGRICUL-TURE	*MINING*	*CONSTRUC*	*MFR_DUR*	*MFR_NON*	*TRANS-PORT*	*COMMUN*	*UTILITIES*	*WHOLE-SALE*	*RETAIL*	*FIDUCIARY*	*SERVICE*
MINING	0.248											
CONSTRUC	0.804	0.415										
MFR_DUR	0.749	0.262	0.873									
MFR_NON	0.662	0.391	0.879	0.841								
TRANSPORT	0.813	0.458	0.976	0.862	0.896							
COMMUN	0.716	0.356	0.930	0.742	0.835	0.923						
UTILITIES	0.674	0.525	0.951	0.837	0.907	0.942	0.903					
WHOLESALE	0.815	0.346	0.977	0.877	0.883	0.973	0.953	0.930				
RETAIL	0.848	0.343	0.984	0.886	0.862	0.965	0.929	0.920	0.984			
FIDUCIARY	0.740	0.190	0.902	0.790	0.793	0.876	0.946	0.851	0.944	0.933		
SERVICES	0.804	0.269	0.955	0.846	0.829	0.932	0.950	0.894	0.979	0.978	0.983	
GOVT	0.814	0.344	0.972	0.843	0.868	0.949	0.957	0.915	0.974	0.982	0.949	0.978

Sample Problem: Gross State Product(GSP)

- Table 4.10: Share Data - Original data divided by the total GSP
- Rank of the 13 × 13 correlation matrix=12

TABLE 4.10 Correlations among 13 different measures of economic activity (data expressed as share of total GSP)

	AGRICUL-TURE	MINING	CONSTRUC	MFR_DUR	MFR_NON	TRANS-PORT	COMMUN	UTILITIES	WHOLE-SALE	RETAIL	FIDUCIARY	SERVICES
MINING	−0.064											
CONSTRUC	0.085	−0.021										
MFR_DUR	0.032	−0.424	−0.130									
MFR_NON	−0.145	−0.138	−0.318	0.204								
TRANSP	0.279	0.612	0.075	−0.357	−0.176							
COMMUN	−0.184	−0.193	−0.023	−0.317	−0.100	−0.049						
UTILITIES	0.043	0.390	0.013	−0.051	0.071	−0.056	−0.169					
WHOLESALE	0.245	−0.553	−0.087	0.271	0.039	−0.214	0.330	−0.267				
RETAIL	0.095	−0.396	0.401	0.195	−0.121	−0.148	0.125	0.030	0.166			
FIDUCIARY	−0.301	−0.406	−0.253	−0.182	−0.133	−0.503	0.120	−0.379	0.040	−0.309		
SERVICES	−0.322	−0.460	0.324	−0.159	−0.458	−0.422	0.309	−0.314	0.239	0.202	0.519	
GOVT	0.110	0.231	0.181	−0.411	−0.237	0.428	0.193	0.045	−0.343	0.287	−0.351	−0.180

Sample Problem: Gross State Product(GSP)

- Results (Raw Data: Table 4.11)

TABLE 4.11 Results from principal components analysis of *GSP_RAW* data: Eigenvalues and loadings

	1	2	3	4	5	6	7	8	9	10	11	12	13
Eigenvalue	10.9443	0.9794	0.4001	0.3427	0.1392	0.0694	0.0401	0.0333	0.0251	0.0105	0.0075	0.0061	0.0022
Cumulative	0.8419	0.9172	0.9480	0.9743	0.9851	0.9904	0.9935	0.9960	0.9980	0.9988	0.9994	0.9998	1.0000

	Loadings		
	Z_1	Z_2	Z_3
AGRICULTURE	0.82452	−0.14508	0.51730
MINING	0.39706	0.90347	0.08516
CONSTRUC	0.98718	0.03238	0.00568
MFR_DUR	0.88799	−0.08722	0.13649
MFR_NON	0.90347	0.09041	−0.12308
TRANSPORT	0.98010	0.08473	0.03860
COMMUN	0.94977	−0.02280	−0.19607
UTILITIES	0.95031	0.19483	−0.14271
WHOLESALE	0.99204	−0.05292	−0.01980
RETAIL	0.98999	−0.06360	0.05374
FIDUCIARY	0.93650	−0.21819	−0.15567
SERVICES	0.97547	−0.14711	−0.05229
GOVT	0.98436	−0.05950	−0.02916

Variance accounted for by		
Z_1	Z_2	Z_3
10.9443	0.9794	0.4001

Sample Problem: Gross State Product(GSP)

- Results (Share Data: Table 4.12)

TABLE 4.12 Results from principal components analysis of *GSP_SHARE* data: Eigenvalues and Loadings

	1	2	3	4	5	6	7	8	9	10	11	12	13
Eigenvalue	3.2355	2.2365	1.9598	1.3603	1.1574	0.8683	0.7245	0.6158	0.3182	0.2354	0.1517	0.1365	0.0000
Cumulative	0.2489	0.4209	0.5717	0.6763	0.7654	0.8321	0.8879	0.9352	0.9597	0.9778	0.9895	1.0000	1.0000

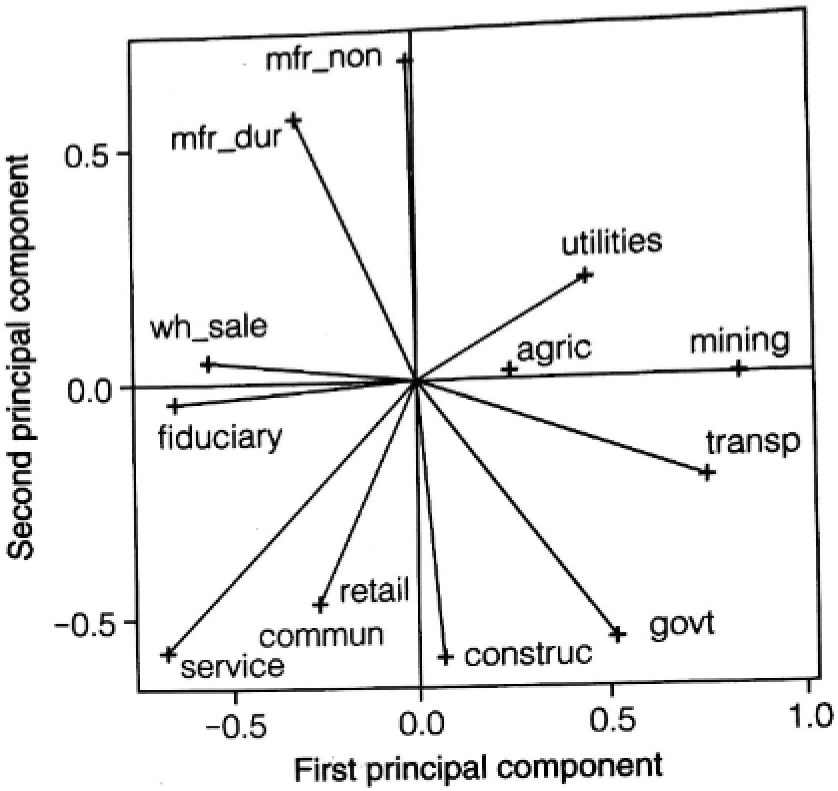
	Loadings		
	Z_1	Z_2	Z_3
AGRICULTURE	0.24251	−0.01116	0.53899
MINING	0.84487	−0.00222	−0.36357
CONSTRUC	0.06347	0.58840	0.36005
MFR_DUR	−0.32981	−0.56192	0.52553
MFR_NON	−0.01746	−0.68617	0.04997
TRANSPORT	0.75273	0.21978	0.00890
COMMUN	−0.27324	0.47225	−0.11488
UTILITIES	0.44418	−0.20634	0.09641
WHOLESALE	−0.56709	−0.04233	0.40618
RETAIL	−0.16213	0.39039	0.71010
FIDUCIARY	−0.65308	0.04526	−0.62523
SERVICES	−0.68331	0.57414	−0.17846
GOVT	0.51955	0.55114	0.11951

Variance accounted for		
Z_1	Z_2	Z_3
3.2355	2.2365	1.9598

Sample Problem: Gross State Product(GSP)

- Factor Loading Plot of Z1 VS. Z2 (Figure 4.12) - Displaying "Similarities"

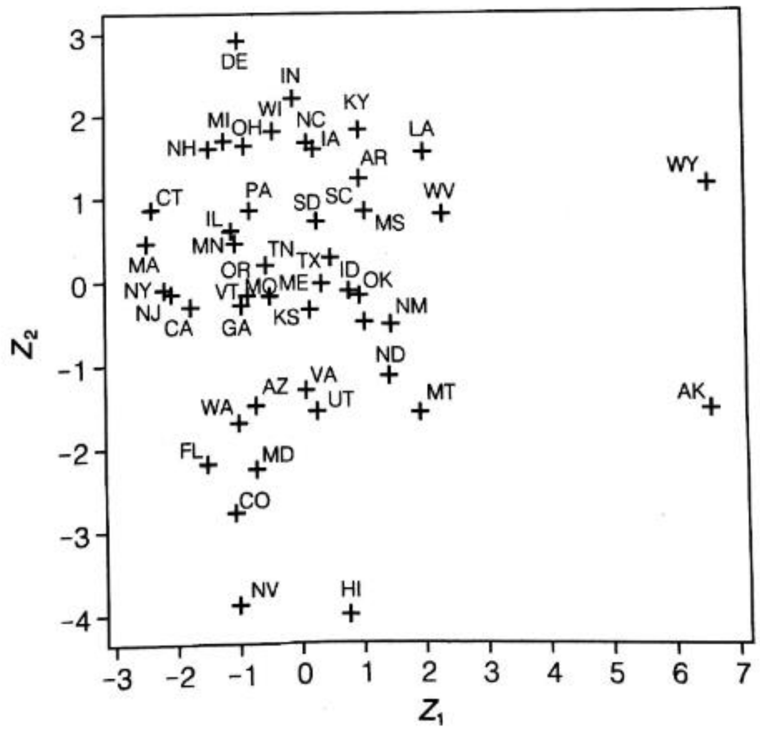
FIGURE 4.12
Plot of factor loadings for first two principal components from *GSP_SHARE* data



Sample Problem: Gross State Product(GSP)

- PC Scores Plot of Z1 VS. Z2 (Figure 4.13) - Displaying “Outliers”

FIGURE 4.13 Plot of principal component scores for first two principal components from *GSP_SHARE* data



Questions Regarding the Application of PC: When Is It Appropriate to Use PC?

- Up to now, we assumed all conditions are valid for using Principal components.
- If the variables are largely independent of one another, then principal components may be not appropriate.
- Let $\mathbf{R}$ the correlation matrix true population of $\mathbf{X}$
  - Let's consider

$$H_0: \mathbf{R} = \mathbf{I} \quad H_1: \mathbf{R} \neq \mathbf{I}$$

  - Determinant of the correlation matrix

$$|\mathbf{R}| = \prod_{j=1}^{p} \lambda_j$$

Questions Regarding the Application of PC: When Is It Appropriate to Use PC?

Bartlett's Sphericity Test

$$\chi_B^2 = -\left[(n-1) - \frac{(2p+5)}{6}\right] ln|\mathbf{R}| \sim \chi^2\left[\frac{(p^2-p)}{6}\right]$$

- Under H_0, $ln|\mathbf{R}| \sim 0$ => $\chi_B^2 \rightarrow 0$
- Under H_1, $ln|\mathbf{R}| < 0$ => χ_B^2 gets larger

- Bartelett's Sphericity Test is for reference, not deterministic!

How Should the Data Be Scaled?

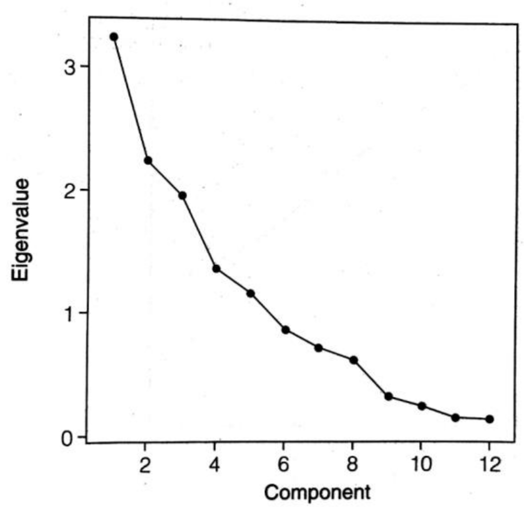
- The result of PCA depends on the scale (e.g. *m* vs. *cm*)
- The variables with large variances might dominate the 1st PCs.
- With different measurement units, use correlation matrix because standardizing ensures that the data are expressed in comparable units
- With same measurement units, use covariance matrix; for instance, a market survey questionnaire

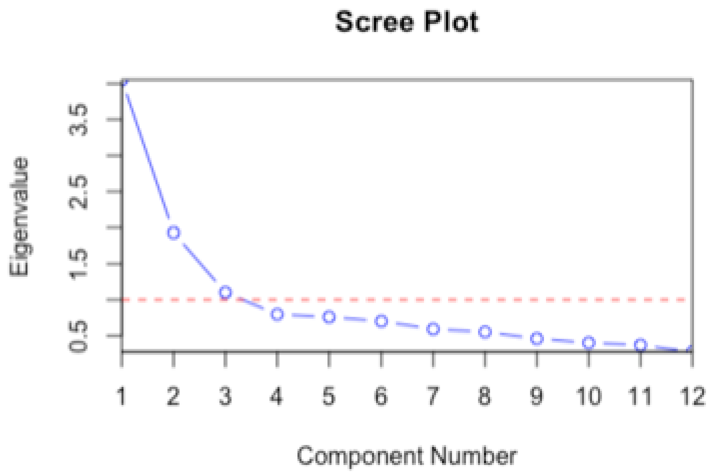
How Many PCs Should Be Retained?

- Rules of thumb
  1. Cumulative percentage of total explained variance

     (e.g. $\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} > 0.8$)
  2. Scree Plot *Cattell* (1996): Elbow Criterion

FIGURE 4.14
Scree plot for
GSP_SHARE data



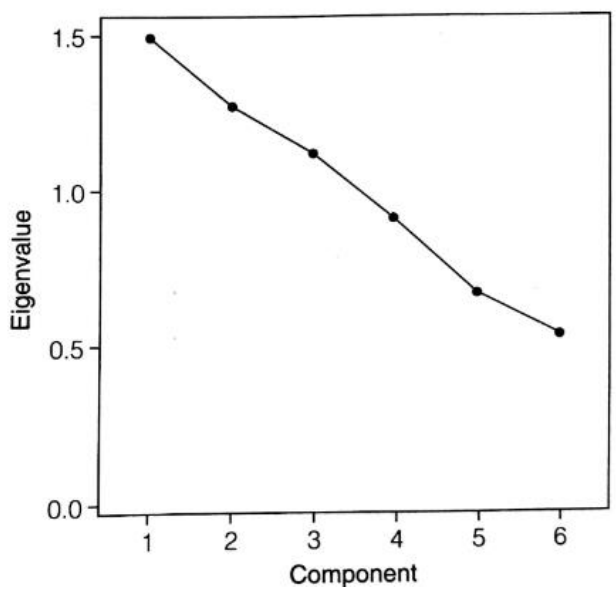


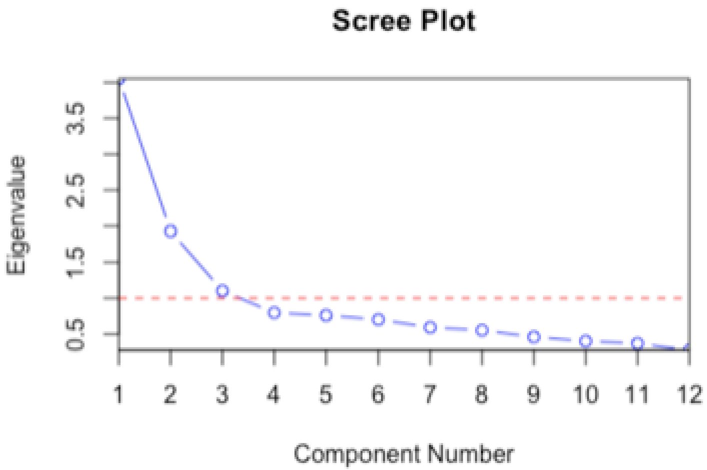
Source: Wiki

How Many PCs Should Be Retained?

- Rules of thumb
  3. *Kaiser*'s Rule (1959): $\lambda > 1$

FIGURE 4.15
Scree plot for
Burke data





Source: Wiki