



Chapter 7: Principal Component Analysis and Canonical Correlation Analysis

Jingyuan Liu
Department of Statistics, School of Economics
Wang Yanan Institute for Studies in Economics
Xiamen University

Outline

1 Introduction to Principal Component Analysis (PCA)

- Intuition and Basic Idea

2 Population PCA

- Deriving Population Principal Components
- Population PCA with Standardized Variables

3 Sample PCA

- Deriving Sample Principal Components
- Geometric Interpretation of Sample PCA
- Deciding How Many Components to Retain
- Application Example with R Implementation

4 Canonical Correlation Analysis

- Canonical Variates and Canonical Correlations


Outline

- 1 Introduction to Principal Component Analysis (PCA)
 - Intuition and Basic Idea

Intuition: Example

The test scores of 52 students on 6 subjects: Y_1 =math, Y_2 =physics, Y_3 =chemistry, Y_4 =Chinese, Y_5 =history, Y_6 =English, were recorded. Part of data are as follows.

1	Y1	Y2	Y3	Y4	Y5	Y6
2	65	61	72	84	81	79
3	77	77	76	64	70	55
4	67	63	49	65	67	57
5	78	84	75	62	71	64
6	66	71	67	52	65	57
7	83	100	79	41	67	50
8	86	94	97	51	63	55
9	67	84	53	58	66	56
10	69	56	67	75	94	80
11	77	90	80	68	66	60
12	84	67	75	60	70	63
13	62	67	83	71	85	77
14	91	74	97	62	71	66
15	82	70	83	68	77	85
16	66	61	77	62	73	64
17	90	78	78	59	72	66

- 
- We want to construct an informative index of the overall examination performance, which provides a measure of examination success that maximally discriminate among students.
 - An average score is certainly one way to provide a single scale on which to compare the students.
 - However, with unequal weights we may be able to spread the students out further and obtain a better ranking.
 - In order to maximally spread the students out, the informative index should retain as much variation in the data as possible.

Principal component analysis (PCA) will provide a useful tool for answering such a question. It seeks for a linear combination of the scores on the different subjects such that it has the largest variability among all possible linear combinations.



Intuition: Remarks

- Recall that discriminant analysis is to seek for the linear combination of variables with the maximum transformed distance between two or more groups.
- In principal component analysis (PCA), however, we only have one population, and we aim to find the linear combination of variables with the maximum variance.

Idea of PCA

- PCA tries to describe the variation in a set of correlated variables, Y_1, \dots, Y_p , in terms of a new **uncorrelated** set of variables, Z_1, \dots, Z_p , called **principal components** of Y_1, \dots, Y_p , each of which is a linear combination of the y -variables.
- Z_j 's are derived in decreasing order of “importance”:
 - 1 Z_1 accounts for as much of the variation in the original data among all linear combinations of y -variables.
 - 2 Z_2 accounts for as much as possible of the remaining variation, subject to being uncorrelated with Z_1 .
 - 3 Z_3 accounts for as much as possible of the remaining variation, subject to being uncorrelated with Z_1 and Z_2 .
 - 4

Usage of the Idea of PCA

- Sometimes we could use a smaller number $k < p$ of the principal components to account for most of the data variability - **data reduction**.
- E.g. In regression, if the number of independent variables is large relative to the sample size, or the independent variables are highly correlated, the PCA can first be conducted before regression analysis.



**Reduce The
Fat In Your
Data**



Outline

2 Population PCA

- Deriving Population Principal Components
- Population PCA with Standardized Variables

Population PCA: Notations

- Let $\mathbf{y} = (Y_1, \dots, Y_p)'$ have the covariance matrix $\mathbf{\Sigma}$.
- The linear combinations Z_1, \dots, Z_p can be described as

$$Z_1 = \mathbf{a}'_1 \mathbf{y} = a_{11} Y_1 + a_{12} Y_2 + \dots + a_{1p} Y_p$$

$$Z_2 = \mathbf{a}'_2 \mathbf{y} = a_{21} Y_1 + a_{22} Y_2 + \dots + a_{2p} Y_p$$

$$\vdots$$

$$Z_p = \mathbf{a}'_p \mathbf{y} = a_{p1} Y_1 + a_{p2} Y_2 + \dots + a_{pp} Y_p$$

with the variances and covariances

$$\text{var}(Z_j) = \mathbf{a}'_j \mathbf{\Sigma} \mathbf{a}_j, \quad \text{cov}(Z_j, Z_k) = \mathbf{a}'_j \mathbf{\Sigma} \mathbf{a}_k, \quad j, k = 1, \dots, p.$$

Population PCA: Objective

According to the idea of PCA, the principal components (PC) should be defined as

First PC = linear combination $\mathbf{a}'_1 \mathbf{y}$ that maximizes $var(\mathbf{a}'_1 \mathbf{y})$
subject to $\mathbf{a}'_1 \mathbf{a}_1 = 1$

Second PC = linear combination $\mathbf{a}'_2 \mathbf{y}$ that maximizes $var(\mathbf{a}'_2 \mathbf{y})$
subject to $\mathbf{a}'_2 \mathbf{a}_2 = 1$ and $cov(\mathbf{a}'_1 \mathbf{y}, \mathbf{a}'_2 \mathbf{y}) = 0$

\vdots

j th PC = linear combination $\mathbf{a}'_j \mathbf{y}$ that maximizes $var(\mathbf{a}'_j \mathbf{y})$
subject to $\mathbf{a}'_j \mathbf{a}_j = 1$ and $cov(\mathbf{a}'_k \mathbf{y}, \mathbf{a}'_j \mathbf{y}) = 0, k < j$

Population PCA: Result

Theorem (Population PCA)

Denote $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$ to be the eigenvalue-eigenvector pairs for $\mathbf{\Sigma}$, where $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ and $\mathbf{e}_1, \dots, \mathbf{e}_p$ are standardized. Then the j th principal component of Y_1, \dots, Y_p is given by

$$Z_j = \mathbf{e}_j' \mathbf{y} = e_{j1} Y_1 + e_{j2} Y_2 + \dots + e_{jp} Y_p, j = 1, \dots, p,$$

with $\text{var}(Z_j) = \mathbf{e}_j' \mathbf{\Sigma} \mathbf{e}_j = \lambda_j$, and $\text{cov}(Z_j, Z_k) = \mathbf{e}_j' \mathbf{\Sigma} \mathbf{e}_k = 0$, $j, k = 1, \dots, p, j \neq k$. Furthermore,

$$\sum_{j=1}^p \text{var}(Z_j) = \sum_{j=1}^p \text{var}(Y_j)$$

PCA for Multivariate Normal Population

- We do not require multivariate normality for the PCA.
- However, if $\mathbf{y} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, the density contours

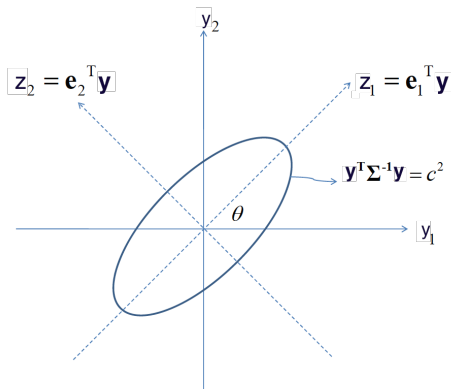
$$\mathbf{y}'\mathbf{\Sigma}^{-1}\mathbf{y} = c^2$$

can be written, in terms of principal components, as

$$\frac{Z_1^2}{\lambda_1} + \frac{Z_2^2}{\lambda_2} + \dots + \frac{Z_p^2}{\lambda_p} = c^2.$$

- So the PCA rotates the coordinate system of y 's to a new one in the direction of the axes of the density contour ellipsoids of Y_1, \dots, Y_p .

For instance, when $p = 2$, the rotation of axes by the principal components can be depicted as follows:



The constant density ellipse and the principle components for bivariate normal distribution.

Population PCA: Remarks

- The proportion of total variance due to the k th principal component is

$$\frac{\lambda_k}{\sum_{j=1}^p \lambda_j}, \quad k = 1, \dots, p.$$

- If most (e.g. 80%) of the total population variance, for large p , can be attributed to the first several principal components, then these components can “replace” the original p variables without much loss of information.
- The correlation between the principal component Z_j and original variable Y_k is

$$\rho_{Z_j, Y_k} = \frac{e_{jk} \sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}, \quad j, k = 1, 2, \dots, p,$$

where σ_{kk} are the k th diagonal of Σ .

Outline

2 Population PCA

- Deriving Population Principal Components
- Population PCA with Standardized Variables

PCA with Standardized Variables

- If $\mathbf{y} = (Y_1, \dots, Y_p)'$ are measured on scales with widely differing ranges or if the units are not commensurate, the principal components of $\mathbf{\Sigma}$ will be dominated by the variables with large variances.
- In this case, we can standardize each component Y_j by $W_j = (Y_j - \mu_j) / \sqrt{\sigma_{jj}}$, $j = 1, \dots, p$. That is,

$$\mathbf{w} = \mathbf{D}_s^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\mathbf{w} = (W_1, \dots, W_p)'$, $\mathbf{D}_s = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}})$.

- Then $\text{COV}(\mathbf{w}) = \mathbf{P}$, the correlation matrix of \mathbf{y} . Hence conducting PCA on W_1, \dots, W_p is equivalent to that on Y_1, \dots, Y_p from \mathbf{P} .

Theorem (Population PCA from \mathbf{P})

Denote $(\tilde{\lambda}_1, \tilde{\mathbf{e}}_1), \dots, (\tilde{\lambda}_p, \tilde{\mathbf{e}}_p)$ to be the eigen pairs for \mathbf{P} , where $\tilde{\lambda}_1 \geq \dots \geq \tilde{\lambda}_p \geq 0$ and $\tilde{\mathbf{e}}_1, \dots, \tilde{\mathbf{e}}_p$ are standardized. Then the j th principal component of standardized variables $\mathbf{w} = (W_1, \dots, W_p)'$, with $\text{COV}(\mathbf{w}) = \mathbf{P}$, is given by

$$V_j = \tilde{\mathbf{e}}_j' \mathbf{w} = \tilde{\mathbf{e}}_j' \mathbf{D}_s^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \tilde{e}_{j1} W_1 + \tilde{e}_{j2} W_2 + \dots + \tilde{e}_{jp} W_p.$$

Furthermore,

$$\sum_{j=1}^p \text{var}(V_j) = \sum_{j=1}^p \text{var}(W_j) = p, \text{ and } \rho_{V_j, W_k} = \tilde{e}_{jk} \sqrt{\tilde{\lambda}_j}.$$

Remarks:

- The principal components from \mathbf{P} are scale invariant.
- The eigen pairs derived from $\mathbf{\Sigma}$ are different from those derived from \mathbf{P} , thus the direction and scale of the two sets of principal components are different, and one set of principal components is not a simple function of the other.
- If the variables are measured in quite different scales, they should probably be standardized for a more balanced representation.

Outline

3 Sample PCA

- Deriving Sample Principal Components
- Geometric Interpretation of Sample PCA
- Deciding How Many Components to Retain
- Application Example with R Implementation

Sample PCA: Example

Back to the test score example, where we want a “better” combination of Y_1, \dots, Y_6 than the unweighted average to distinguish the students - It is actually a sample-level PCA.

1	Y1	Y2	Y3	Y4	Y5	Y6
2	65	61	72	84	81	79
3	77	77	76	64	70	55
4	67	63	49	65	67	57
5	78	84	75	62	71	64
6	66	71	67	52	65	57
7	83	100	79	41	67	50
8	86	94	97	51	63	55
9	67	84	53	58	66	56
10	69	56	67	75	94	80
11	77	90	80	68	66	60
12	84	67	75	60	70	63
13	62	67	83	71	85	77
14	91	74	97	62	71	66
15	82	70	83	68	77	85
16	66	61	77	62	73	64
17	90	78	78	59	72	66

Sample PCA: Introduction

- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$, where $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})'$, represent n iid copies of the p -variate random vector \mathbf{y} , with sample mean vector $\bar{\mathbf{y}}$, sample covariance matrix \mathbf{S} and sample correlation matrix \mathbf{R} .
- Our objective is to construct some uncorrelated linear combinations of the measured \mathbf{y} that account for as much of the variation in the sample as possible.
- The linear combination of each observation vector can be described by $z_i = \mathbf{a}'\mathbf{y}_i$, with the sample mean $\bar{z} = \mathbf{a}'\bar{\mathbf{y}}$ and sample variance $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$. If two linear combinations $z_{i1} = \mathbf{a}'_1\mathbf{y}_i$ and $z_{i2} = \mathbf{a}'_2\mathbf{y}_i$ are involved, their sample covariance is $s_{z_1 z_2} = \mathbf{a}'_1\mathbf{S}\mathbf{a}_2$.

Sample PCA: Objective

The sample principal components (PC) should be defined as

First PC = linear combination $\mathbf{a}'_1 \mathbf{y}_i$ that maximizes $\mathbf{a}'_1 \mathbf{S} \mathbf{a}_1$
subject to $\mathbf{a}'_1 \mathbf{a}_1 = 1$

Second PC = linear combination $\mathbf{a}'_2 \mathbf{y}_i$ that maximizes $\mathbf{a}'_2 \mathbf{S} \mathbf{a}_2$
subject to $\mathbf{a}'_2 \mathbf{a}_2 = 1$ and $\mathbf{a}'_1 \mathbf{S} \mathbf{a}_2 = 0$

\vdots

j th PC = linear combination $\mathbf{a}'_j \mathbf{y}_i$ that maximizes $\mathbf{a}'_j \mathbf{S} \mathbf{a}_j$
subject to $\mathbf{a}'_j \mathbf{a}_j = 1$ and $\mathbf{a}'_k \mathbf{S} \mathbf{a}_j = 0, k < j$

Sample PCA: Result

Theorem (Sample PCA)

Denote $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), \dots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$ to be the eigenvalue-eigenvector pairs for \mathbf{S} , where $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p \geq 0$ and $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_p$ are standardized. Then the j th sample principal component from the i th sample is given by

$$z_{ij} = \hat{\mathbf{e}}_j' \mathbf{y}_i = \hat{e}_{j1}y_{i1} + \hat{e}_{j2}y_{i2} + \dots + \hat{e}_{jp}y_{ip}, j = 1, \dots, p, i = 1, \dots, n,$$

with $s_{z_j}^2 = \hat{\mathbf{e}}_j' \mathbf{S} \hat{\mathbf{e}}_j = \hat{\lambda}_j$, and $s_{z_j z_k} = \hat{\mathbf{e}}_j' \mathbf{S} \hat{\mathbf{e}}_k = 0, j, k = 1, \dots, p, j \neq k$. Furthermore, let s_{jj} be the j th diagonal of \mathbf{S} , then

$$\text{total sample variance} = \sum_{j=1}^p s_{jj} = \sum_{j=1}^p \hat{\lambda}_j.$$

Sample PCA: Remarks

- The sample correlation between Z_j and Y_k is

$$r_{z_j y_k} = \frac{\hat{e}_{jk} \sqrt{\hat{\lambda}_j}}{\sqrt{s_{kk}}}, \quad j, k = 1, 2, \dots, p.$$

- The same argument as the population PCA about the principal components constructed from the covariance or the correlation matrix applies to the sample PCA.
- The observations \mathbf{y}_i 's are often centered by $\mathbf{y}_i - \bar{\mathbf{y}}$. This has no effects on the sample covariance or correlation matrix, hence will not change the coefficient vector $\hat{\mathbf{e}}_j$'s.

Outline

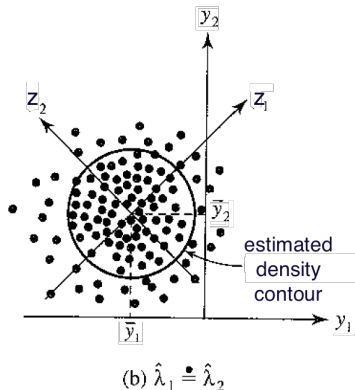
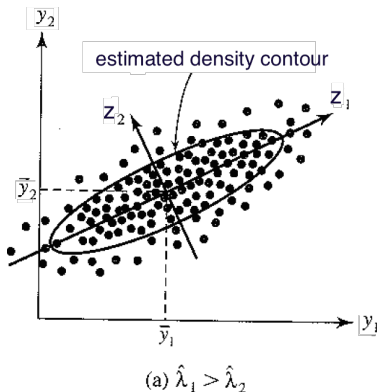
3 Sample PCA

- Deriving Sample Principal Components
- Geometric Interpretation of Sample PCA
- Deciding How Many Components to Retain
- Application Example with R Implementation

Sample PCA: Geometric Interpretation

In the p -variate scatter plot of data, the centered principal components, $z_{ij} = \hat{\mathbf{e}}_j'(\mathbf{y}_i - \bar{\mathbf{y}})$, $j = 1, \dots, p$, $i = 1, \dots, n$, can be viewed as shifting the origin of the original coordinate system to $\bar{\mathbf{y}}$ and rotating the coordinate axes until they pass through the scatter in the directions of maximum variances.

For instance, when $p = 2$, the following two plots depict the rotation and shifting of axes by sample principal components:



Geometric Interpretation: Example

The following table consists of part of the head measurements on first and second sons (Frets 1921) from 25 families.

First Son		Second Son	
Head Length	Head Breadth	Head Length	Head Breadth
y_1	y_2	x_1	x_2
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151
179	158	186	148
183	147	174	147
174	150	185	152
190	159	195	157

To illustrate the sample principal components as a rotation when $p = 2$, we use two variables from the data: Y_1 is the head length and Y_2 is head width for the first son. The mean vector and covariance matrix are

$$\bar{\mathbf{y}} = \begin{pmatrix} 185.7 \\ 151.1 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 95.29 & 52.87 \\ 52.87 & 54.36 \end{pmatrix}.$$

The eigenvalues and eigenvectors of \mathbf{S} are

$$\begin{aligned} \hat{\lambda}_1 &= 131.52, & \hat{\mathbf{e}}_1 &= (0.825, 0.565)', \\ \hat{\lambda}_2 &= 18.14, & \hat{\mathbf{e}}_2 &= (-0.565, 0.825)'. \end{aligned}$$

The symmetry of $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ is due to orthogonality: $\hat{\mathbf{e}}_1' \hat{\mathbf{e}}_2 = 0$.

Thus the centered principal components are

$$z_1 = 0.825(y_1 - 185.7) + 0.565(y_2 - 151.1)$$

$$z_2 = -0.565(y_1 - 185.7) + 0.825(y_2 - 151.1).$$

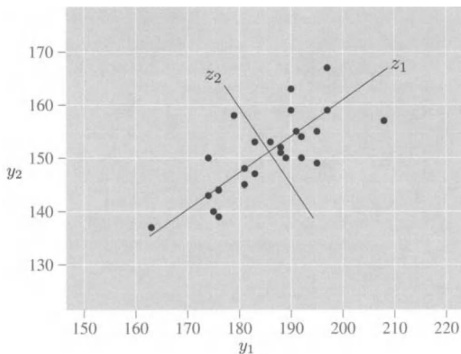
The axes of the sample principal components can be obtained by setting z_1 and z_2 to be zero.

■ Major axis: Set $z_2 = 0$

$$y_2 - 151.1 = \frac{0.565}{0.825}(y_1 - 185.7)$$

■ Minor axis: Set $z_1 = 0$

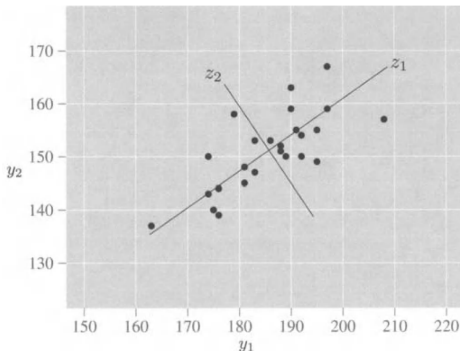
$$y_2 - 151.1 = -\frac{0.825}{0.565}(y_1 - 185.7)$$



The major axis is the line passing through $\bar{\mathbf{y}} = (185.7, 151.1)'$ in the direction $\hat{\mathbf{e}}_1 = (0.825, 0.565)'$, with slope $0.565/0.825$.

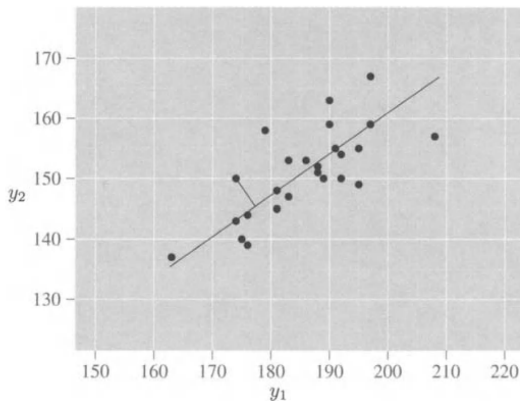
The First Principal Component

In the above example, the major axis of the principal components looks very much like a regression line. Is it really the case?



Principal Components and Regression

Consider the perpendicular distance from each point to the major axis:



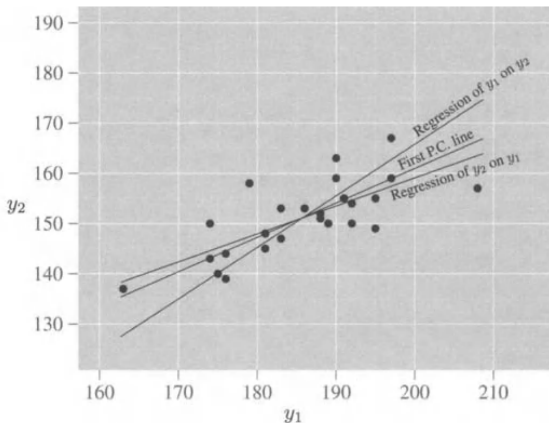
- When $p = 2$, the perpendicular distance from the i th point to the line is simply $z_{i2} = \hat{\mathbf{e}}_2'(\mathbf{y}_i - \bar{\mathbf{y}})$.
- Hence the sum of squares of perpendicular distances is

$$\sum_{i=1}^n z_{i2}^2 = \sum_{i=1}^n \hat{\mathbf{e}}_2'(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' \hat{\mathbf{e}}_2 = (n - 1)\lambda_2,$$

which is a minimum when $p = 2$.

- Therefore, the major axis of principal components minimizes the total sum of squared perpendicular distances from the points to the line.

Recall that the regression line minimizes the “vertical” distances from the points to the line.



Outline

3 Sample PCA

- Deriving Sample Principal Components
- Geometric Interpretation of Sample PCA
- Deciding How Many Components to Retain
- Application Example with R Implementation

Deciding How Many Components

In every application, a decision must be made on how many principal components should be retained in order to effectively summarize the data. The following criteria are proposed:

- (1) **Percentage cutoff:** Retain sufficient number of principal components to account for a specified percentage of the total variance, say 80%.
- (2) **Average cutoff:** Retain the principal components whose eigenvalues are greater than the average of eigenvalues, $\sum_{j=1}^p \lambda_j / p$. For a correlation matrix, this average is 1.
- (3) **scree graph:** Use a plot of λ_j versus j , and look for a natural break between the “large” eigenvalues and the “small” eigenvalues.

Deciding How Many Components: Percentage Cutoff

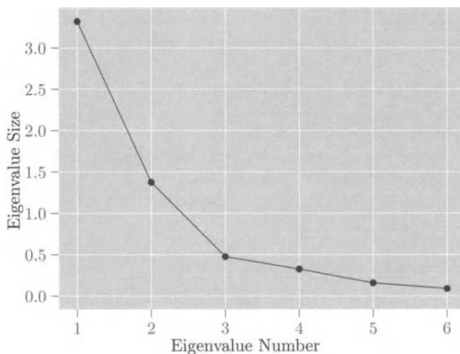
- This criterion is intuitive and has easy interpretation for the selected principal components.
- However, the challenge lies in selecting an appropriate threshold percentage.
 - Intuitively, we should find a relatively high percentage.
 - But if we aim too high, we run the risk of including components that are “sample specific”.

Deciding How Many Components: Average Cutoff

- This criterion is widely used and is the default in many software packages.
- The average eigenvalue is also the average variance of the individual variables. Thus this criterion retains those components that account for more variance than the average variance of the variables.
- In cases where the data can be successfully summarized in a relatively small number of dimensions, there is often a wide gap between the two eigenvalues that fall on both sides of the average.

Deciding How Many Components: Scree Graph

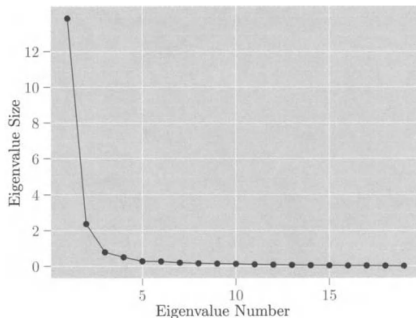
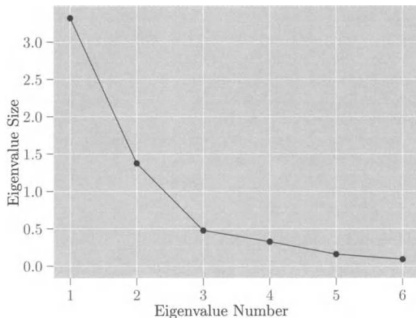
The scree graph is named for its similarity in appearance to a cliff with rocky debris at its bottom.



Scree Graph

- In most cases, the first several eigenvalues form a steep curve, followed by a bend and then a straight-line trend with shallow slope.
- The recommendation is to retain those eigenvalues in the steep curve **before** the first one on the straight line.
- In practice, however, the turning point between the steep curve and the straight line may not be distinct, or there may be more than one discernible bend. In such cases, this approach is not as conclusive.

Scree Graph: Example



- In the left scree graph, the first two principal components should be retained.
- In the right scree graph, the rule suggests two, but possibly four is better.

The Last Few Principal Components

- Up to this point, we have focused on using the first few principal components to summarize and simplify the data.
- However, the last few components may carry useful information in some applications.
- If, for example, the last eigenvalue is near zero, it signifies the presence of a collinearity.

Outline

3 Sample PCA

- Deriving Sample Principal Components
- Geometric Interpretation of Sample PCA
- Deciding How Many Components to Retain
- Application Example with R Implementation

Test Score Example

Back to the test score example, where the scores of 52 students on 6 subjects were collected.

1	Y1	Y2	Y3	Y4	Y5	Y6
2	65	61	72	84	81	79
3	77	77	76	64	70	55
4	67	63	49	65	67	57
5	78	84	75	62	71	64
6	66	71	67	52	65	57
7	83	100	79	41	67	50
8	86	94	97	51	63	55
9	67	84	53	58	66	56
10	69	56	67	75	94	80
11	77	90	80	68	66	60
12	84	67	75	60	70	63
13	62	67	83	71	85	77
14	91	74	97	62	71	66
15	82	70	83	68	77	85
16	66	61	77	62	73	64
17	90	78	78	59	72	66

We could first check the sample correlation matrix of the original 6 variables:

```
> test<-read.table("/Users/jingyuan/快盘/Teaching/Multivariate  
Analysis/R code/Chap7/test_score.csv", sep=";", header=T)  
> (R<-round(cor(test), 3)) # sample correlation matrix
```

	Y1	Y2	Y3	Y4	Y5	Y6
Y1	1.000	0.647	0.696	-0.561	-0.456	-0.439
Y2	0.647	1.000	0.573	-0.503	-0.351	-0.458
Y3	0.696	0.573	1.000	-0.380	-0.274	-0.244
Y4	-0.561	-0.503	-0.380	1.000	0.813	0.835
Y5	-0.456	-0.351	-0.274	0.813	1.000	0.819
Y6	-0.439	-0.458	-0.244	0.835	0.819	1.000

We can see that the original variables are highly correlated.

The sample PCA can then be conducted, via the sample correlation matrix \mathbf{R} , i.e. based on the standardized variables:

```
> test_PCA<-princomp(test, cor=T) # sample PCA  
> summary(test_PCA, loadings=T)
```

Importance of components:


	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.9261112	1.1236019	0.66395522	0.52009785	0.41172308	0.38309295
Proportion of Variance	0.6183174	0.2104135	0.07347275	0.04508363	0.02825265	0.02446003
Cumulative Proportion	0.6183174	0.8287309	0.90220369	0.94728732	0.97553997	1.00000000

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Y1	-0.412	-0.376	0.216	0.788		0.145
Y2	-0.381	-0.357	-0.806	-0.118	0.212	-0.141
Y3	-0.332	-0.563	0.467	-0.588		
Y4	0.461	-0.279			0.599	0.590
Y5	0.421	-0.415	-0.250		-0.738	0.205
Y6	0.430	-0.407	0.146	0.134	0.222	-0.749

Note that if the sample covariance matrix \mathbf{S} is used, we could omit $cor=T$, or specify $cor=F$.

- The “standard deviation” provides the square-roots of the eigenvalues of \mathbf{R} , e.g. $\sqrt{\tilde{\lambda}_1} = 1.926$, $\sqrt{\tilde{\lambda}_2} = 1.124$, where $\tilde{\lambda}_j$ is the j th eigenvalue of \mathbf{R} .
- The “proportion of variance” reports the contribution of each principal component to the total variance of data $\tilde{\lambda}_j / \sum_{j=1}^6 \tilde{\lambda}_j = \tilde{\lambda}_j / p$ since \mathbf{R} is used.
- The “cumulative proportion” cumulates the sequential contributions of the principal components. From this quantity, it seems that the first two principal components are already sufficient to capture most (82.9%) of the total variation in the data.




The “loadings” of principal components are formed by the orthogonal eigenvectors of \mathbf{R} . Each column corresponds to each eigenvector, so it provides the coefficients of principal components. E.g. The first two principal components are

$$z_1 = -0.412y_1 - 0.381y_2 - 0.332y_3 + 0.461y_4 + 0.421y_5 + 0.430y_6$$

$$z_2 = -0.376y_1 - 0.357y_2 - 0.563y_3 - 0.279y_4 - 0.415y_5 - 0.407y_6$$

Note that the y_j 's are the standardized version.



How to understand the first two principal components from their coefficients?

To interpret the meaning of the principal components, recall that Y_1 =math, Y_2 =physics, Y_3 =chemistry, Y_4 =Chinese, Y_5 =history, Y_6 =English. That is, Y_1 , Y_2 and Y_3 are science subjects, and Y_4 , Y_5 and Y_6 are liberal arts.

- z_1 consists of opposite signs between (Y_1, Y_2, Y_3) and (Y_4, Y_5, Y_6) , with similar magnitude. This implies the groupwise discrepancy between science and liberal arts. The typical students are 6,7,45,30,49.

```
> test[c(6,7,45,30,49),]
```

	Y1	Y2	Y3	Y4	Y5	Y6
6	83	100	79	41	67	50
7	86	94	97	51	63	55
45	99	100	99	53	63	60
30	64	61	49	100	99	95
49	52	62	65	100	96	100

- z_2 consists of the same sign for all Y_j 's, reflecting the balanced characteristic among all subjects, such as students 26,33,8.


```
> test[c(26,33,8),]
```

	Y1	Y2	Y3	Y4	Y5	Y6
26	87	84	100	74	81	76
33	86	78	92	87	87	77
8	67	84	53	58	66	56

We may further explore the sample principal component scores of those typical students:

```
> # sample principal components of the typical students
> samplePC<-(round(test_PCA$scores,3))[c(6,7,45,30,49,26,33,8),]
> rownames(samplePC)<-c(6,7,45,30,49,26,33,8)
> samplePC
```

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
6	-3.518	0.820	-1.072	-0.156	-0.763	-0.166
7	-3.516	-0.104	0.101	-0.574	-0.011	0.080
45	-3.975	-1.054	0.147	0.349	0.252	-0.049
30	4.490	-0.693	-0.620	0.832	-0.054	0.029
49	4.622	-0.997	-0.236	-0.724	0.289	-0.465
26	-0.841	-2.117	0.544	-0.156	-0.070	0.192
33	0.345	-2.187	0.414	0.225	0.018	0.854
8	-0.982	2.326	-1.292	-0.017	0.093	-0.052

- 
- The first principal components of students 6,7 and 45 are negative with large absolute values, indicating their performances in science are significant better than that in art. Opposite conclusion is drawn for students 30 and 49.
 - The second principal components of students 26 and 33 are negative with large absolute values, thus they perform well in all subjects. While student 8 performs poorly in all.

The above principal component scores can also be obtained by predicting the current sample:

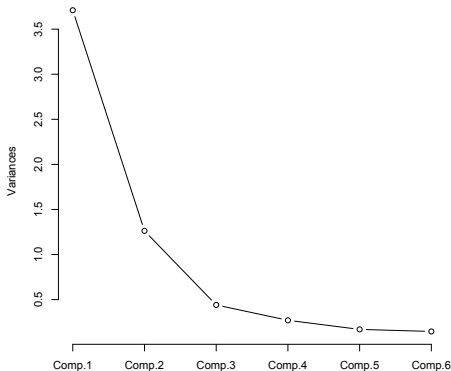
```
> # another way to obtain the sample principal components
> samplePC2<-round(predict(test_PCA),3) [c(6,7,45,30,49,26,33,8),]
> rownames(samplePC2)<-c(6,7,45,30,49,26,33,8)
> samplePC2
```

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
6	-3.518	0.820	-1.072	-0.156	-0.763	-0.166
7	-3.516	-0.104	0.101	-0.574	-0.011	0.080
45	-3.975	-1.054	0.147	0.349	0.252	-0.049
30	4.490	-0.693	-0.620	0.832	-0.054	0.029
49	4.622	-0.997	-0.236	-0.724	0.289	-0.465
26	-0.841	-2.117	0.544	-0.156	-0.070	0.192
33	0.345	-2.187	0.414	0.225	0.018	0.854
8	-0.982	2.326	-1.292	-0.017	0.093	-0.052

Of course, the “*predict*” function can be used to predict new observation, in addition to the current sample.

As was aforementioned, the first two principal components seem to be sufficient for describing the data. This can further be verified by the scree graph:

```
> screeplot(test_PCA, type="lines") # scree graph
```





Remark:

- In addition to “*princomp*” function, some more advanced functions from package “*psych*” are available.

Outline

- 4 Canonical Correlation Analysis
 - Canonical Variates and Canonical Correlations

Introduction to Canonical Correlations

- PCA aims to explain a set of correlated variables by the uncorrelated linear combinations of them, to account for the maximal variation among the original variables.
- In canonical correlation analysis, we try to explain two sets of variables by two sets of linear combinations, to account for the maximal correlation between the original two sets.
- Canonical correlations measure the strength of association between the two sets of variables, such as relating college “performance” variables with precollege “achievement” variables.

Idea of Canonical Correlation Analysis

- The idea is first to find the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair, and so on. The pairs of linear combinations are called the **canonical variates/variables** and their correlations are called **canonical correlations**.
- Details of this section refer to Chapter 10, “*Applied Multivariate Statistical Analysis*”.

- Specifically, let $\mathbf{y}_1 = (Y_1, \dots, Y_p)'$ denote the first group of p variables with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_{11}$. Let $\mathbf{y}_2 = (Y_{p+1}, \dots, Y_{p+q})'$ denote the second group of q variables with $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_{22}$. The covariance matrix between the two vectors is $COV(\mathbf{y}_1, \mathbf{y}_2) = \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21}$.
- Set $U = \mathbf{a}'\mathbf{y}_1$ and $V = \mathbf{b}'\mathbf{y}_2$ for some pair of coefficient vectors \mathbf{a} and \mathbf{b} . We have

$$\text{corr}(U, V) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{(\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a})(\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b})}}.$$

The canonical variates are defined as following:

- The first pair of canonical variables $U_1 = \mathbf{a}'_1 \mathbf{y}_1$ and $V_1 = \mathbf{b}'_1 \mathbf{y}_2$, with unit variances, are defined by

$$(\mathbf{a}_1, \mathbf{b}_1) = \operatorname{argmax}_{\mathbf{a}, \mathbf{b}} \frac{\mathbf{a}' \boldsymbol{\Sigma}_{12} \mathbf{b}}{\sqrt{(\mathbf{a}' \boldsymbol{\Sigma}_{11} \mathbf{a})(\mathbf{b}' \boldsymbol{\Sigma}_{22} \mathbf{b})}}.$$

- For $k \geq 2$, the k th pair of canonical variables is the pair of U_k, V_k , having unit variances, which maximize the above correlation among all choices that are uncorrelated with the previous $k - 1$ canonical variable pairs, i.e. for all $l < k$,

$$\operatorname{corr}(U_k, U_l) = \operatorname{corr}(V_k, V_l) = \operatorname{corr}(U_k, V_l) = \operatorname{corr}(V_k, U_l) = 0.$$

Canonical Correlation: Result

Theorem (Canonical variates and canonical correlations)

Let $\rho_1^{*2} \geq \dots \geq \rho_s^{*2}$ be the common (largest) eigenvalues of $\mathbf{A} = \mathbf{\Sigma}_{11}^{-1/2} \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1/2}$ and $\mathbf{B} = \mathbf{\Sigma}_{22}^{-1/2} \mathbf{\Sigma}_{21} \mathbf{\Sigma}_{11}^{-1} \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1/2}$, where $s = \min(p, q)$. Denote $\mathbf{e}_1, \dots, \mathbf{e}_s$ and $\mathbf{f}_1, \dots, \mathbf{f}_s$ are the corresponding eigenvectors of \mathbf{A} and \mathbf{B} . Then

$$\max_{\mathbf{a}, \mathbf{b}} \text{corr}(U, V) = \rho_1^*,$$

attained by the first canonical variate pair

$$U_1 = \mathbf{a}'_1 \mathbf{y}_1 = \mathbf{e}'_1 \mathbf{\Sigma}_{11}^{-1/2} \mathbf{y}_1 \text{ and } V_1 = \mathbf{b}'_1 \mathbf{y}_2 = \mathbf{f}'_1 \mathbf{\Sigma}_{22}^{-1/2} \mathbf{y}_2.$$

And the k th pair of canonical variates, $k = 1, \dots, s$,

$$U_k = \mathbf{a}'_k \mathbf{y}_1 = \mathbf{e}'_k \mathbf{\Sigma}_{11}^{-1/2} \mathbf{y}_1 \text{ and } V_k = \mathbf{b}'_k \mathbf{y}_2 = \mathbf{f}'_k \mathbf{\Sigma}_{22}^{-1/2} \mathbf{y}_2$$

maximizes $\text{corr}(U_k, V_k) = \rho_k^*$ among those linear combinations uncorrelated with the preceding $k - 1$ pairs of variables.

Canonical Correlation: Remarks

- If the original variables are standardized, then the new canonical variates for the standardized variables are obtained by replacing Σ with \mathbf{P} .
- The result of sample canonical correlation analysis can be modified by replacing Σ with \mathbf{S} .
- The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pairs of canonical variables.

Canonical Correlation: Example

Example: A health club collected three health indices, Y_1 = weight, Y_2 = waist size and Y_3 = pulse, from 20 mid-age male. Three training indices are also recorded: Y_4 = number of pull-ups, Y_5 = number of sit-ups, and Y_6 = number of leaps.

1	Y1	Y2	Y3	Y4	Y5	Y6
2	191	36	50	5	162	60
3	189	37	52	2	110	60
4	193	38	58	12	101	101
5	162	35	62	12	105	37
6	189	35	46	13	155	58
7	182	36	56	4	101	42
8	211	38	56	8	101	38
9	167	34	60	6	125	40
10	176	31	74	15	200	40
11	154	33	56	17	251	250
12	169	34	50	17	120	38
13	166	33	52	13	210	115
14	154	34	64	14	215	105
15	247	46	50	1	50	50
16	193	36	46	6	70	31
17	202	37	62	12	210	120
18	176	37	54	4	60	25
19	157	32	52	11	230	80
20	156	33	54	15	225	73
21	138	33	68	2	110	43

To investigate the association between the health indices $\mathbf{y}_1 = (Y_1, Y_2, Y_3)'$ and the training indices $\mathbf{y}_2 = (Y_4, Y_5, Y_6)'$:

```
> health<-read.table("/Users/jingyuan/快盘/Teaching/Multivariate Analysis/R code/
Chap7/health.csv",sep=",", header=T)
> (R<-round(cor(health),3))
```

	Y1	Y2	Y3	Y4	Y5	Y6
Y1	1.000	0.870	-0.366	-0.390	-0.493	-0.226
Y2	0.870	1.000	-0.353	-0.552	-0.646	-0.191
Y3	-0.366	-0.353	1.000	0.151	0.225	0.035
Y4	-0.390	-0.552	0.151	1.000	0.696	0.496
Y5	-0.493	-0.646	0.225	0.696	1.000	0.669
Y6	-0.226	-0.191	0.035	0.496	0.669	1.000

```
> R11=R[1:3,1:3]
> R12=R[1:3,4:6]
> R21=R[4:6,1:3]
> R22=R[4:6,4:6]
> A<-solve(R11)%*%R12%*%solve(R22)%*%R21 # matrix for the first group Y1,Y2,Y3
> ev<-eigen(A)$values # common eigenvalues of both groups
> round(sqrt(ev),3) # the canonical correlations
[1] 0.796 0.200 0.071
```

We could also directly apply the “*cancor*” function, here via the sample correlation matrix **R** with standardized variables:

```
> health.std=scale(health) # standardize the original data
> ca=cancor(health.std[,1:3],health.std[,4:6]) # canonical correlation analysis
via R
> ca$cor      # canonical correlations
[1] 0.79560815 0.20055604 0.07257029
> ca$xcoef    # the loadings (coefficients) of the first group
      [,1]      [,2]      [,3]
Y1 -0.17788841 -0.43230348  0.04381432
Y2  0.36232695  0.27085764 -0.11608883
Y3 -0.01356309 -0.05301954 -0.24106633
> ca$ycoef    # the loadings (coefficients) of the second group
      [,1]      [,2]      [,3]
Y4 -0.08018009 -0.08615561  0.29745900
Y5 -0.24180670  0.02833066 -0.28373986
Y6  0.16435956  0.24367781  0.09608099
```

For instance, the first pair of canonical variates is obtained with the standardized Y_1, \dots, Y_6 :

$$U_1 = -0.178Y_1 + 0.362Y_2 - 0.014Y_3, \quad V_1 = -0.08Y_4 - 0.24Y_5 + 0.164Y_6.$$

Summary and Take-home Messages

- What is the objective of PCA?
- How to derive the population and sample principal components?
- How to interpret the principal components geometrically?
- What is the difference between the PCA via covariance matrix and that via correlation matrix?
- How to decide how many principal components to retain?
- What are canonical variates and canonical correlations?
And how to obtain them?