

# **Lecture 11: Principal component and factor analyses**

**BTBI3008I**

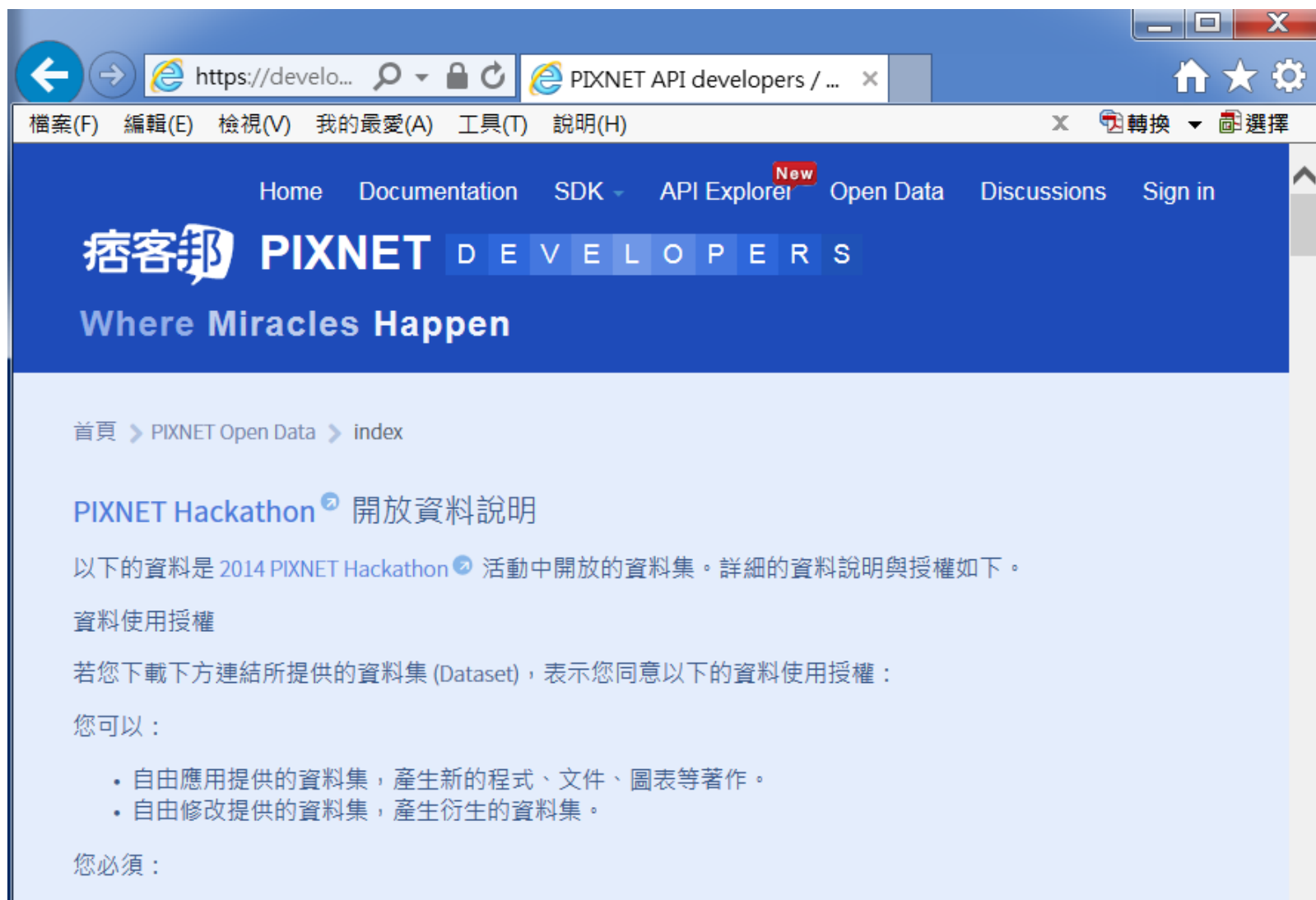
**統計應用方法** Applied Methods in Statistics

**2025/4/30**

# Aspects of multivariate analysis

- Data include simultaneous measurements on many variables- multivariate analysis
- Multivariate methods include the following:
  1. Hypothesis construction and testing
  2. Investigation of the dependence among variables
  3. Data reduction or structure simplification (principal component and factor analysis)
  4. Sorting and grouping (clustering)
  5. Prediction (classification)

# PIXNET Hackathon 開放資料



The screenshot shows a web browser window with the URL <https://developers.pixnet.tw>. The page has a blue header with navigation links: Home, Documentation, SDK, API Explorer (marked as 'New'), Open Data, Discussions, and Sign in. The main title is '痞客邦 PIXNET DEVELOPERS' with the tagline 'Where Miracles Happen'. The breadcrumb trail is '首頁 > PIXNET Open Data > index'. The main heading is 'PIXNET Hackathon 開放資料說明'. The text below states: '以下的資料是 2014 PIXNET Hackathon 活動中開放的資料集。詳細的資料說明與授權如下。' followed by '資料使用授權'. A paragraph says: '若您下載下方連結所提供的資料集 (Dataset)，表示您同意以下的資料使用授權：' followed by '您可以：'. A bulleted list follows: '• 自由應用提供的資料集，產生新的程式、文件、圖表等著作。' and '• 自由修改提供的資料集，產生衍生的資料集。'. The final line says '您必須：'.

Home Documentation SDK API Explorer <sup>New</sup> Open Data Discussions Sign in

痞客邦 PIXNET DEVELOPERS

Where Miracles Happen

首頁 > PIXNET Open Data > index

## PIXNET Hackathon 開放資料說明

以下的資料是 2014 PIXNET Hackathon 活動中開放的資料集。詳細的資料說明與授權如下。

### 資料使用授權

若您下載下方連結所提供的資料集 (Dataset)，表示您同意以下的資料使用授權：

您可以：

- 自由應用提供的資料集，產生新的程式、文件、圖表等著作。
- 自由修改提供的資料集，產生衍生的資料集。

您必須：

## Datasets

### 1. 分類熱門照片 (含 EXIF)

下載連結：[JSON format](#) | [CSV format](#)

資料格式：

ZIP 檔案內有 24 個 JSON 檔案，檔案名稱為 "相簿分類 ID"，分類名稱請參考頁面上方「資料使用須知」。所有的資料一定有 EXIF 與位置資訊。

### 2. 去識別化後的照片 EXIF 資訊

下載連結：[JSON format](#) | [CSV format](#)

資料格式：

ZIP 檔案內有 5 個 JSON 檔案，檔案名稱為 YYYY-W.json，每個檔案內含一週的資料，分類名稱請參考頁面上方「資料使用須知」。所有的資料一定有 EXIF 欄位。

### 3. 分類人氣部落格相關資料

下載連結：[JSON format](#) | [CSV format](#)

資料格式：

ZIP 檔案內有 41 個 JSON 檔案，檔案名稱為 "部落格分類ID"。部落格分類名稱請參考頁面上方「資料使用須知」。

### 4. 去識別化後的部落格訪客資料

下載連結：[JSON format](#) | [CSV format](#)

資料格式：

ZIP 檔案內有 1 個 JSON 檔案，檔案名稱為 `YYYY-mm.json`，表示該月份的訪客資訊。部落格文章分類名稱請參考頁面上方「資料使用須知」。

# **Principal component analysis**

# Principal component analysis (PCA)

- A PCA is concerned with explaining the variance-covariance structure of a set of variables through a few “linear” combinations of these variables.
- Objectives of a PCA:
  - data reduction: the total variability of  $p$  variables can be accounted for by  $k$  principle components, where  $p > k$ .
  - interpretation: can reveal relationship that were not previously suspected.

# PCA

- Principle components depend solely on the covariance matrix.
- Development of principle components dose **NOT** require a multivariate normal assumption.
- However, a multivariate normal assumption is useful for inference of the principle components.

# Principal components (PCs)

- Observed data:  $\mathbf{X} = (X_1, X_2, \dots, X_p)$

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

- $$Y_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$
$$\vdots$$

$$Y_k = a_{k1}X_1 + a_{k2}X_2 + \dots + a_{kp}X_p$$

- $k < p$



# PCs

- Find  $a_{11}, a_{12}, \dots, a_{1p}$  such that  $\text{Var}(Y_1)$  has the maximum variance among all linear combination of  $X \rightarrow Y_1$  is the 1st principal component
- Find  $a_{21}, a_{22}, \dots, a_{2p}$  such that  $\text{Var}(Y_2)$  has the maximum variance among all linear combination of  $X$  with  $\text{Cov}(Y_1, Y_2) = 0 \rightarrow Y_2$  is the 2nd principal component
- Find  $a_{k1}, a_{k2}, \dots, a_{kp}$  such that  $\text{Var}(Y_k)$  has the maximum variance among all linear combination of  $X$  with  $\text{Cov}(Y_i, Y_k) = 0 \forall i < k \rightarrow Y_k$  is the  $k$ th principal component

# PCs (cont'd)

- $\text{Var}(Y_1)$  = the largest eigenvalue of **the covariance matrix  $\text{Cov}(\mathbf{X})$** , and  $(a_{11}, a_{12}, \dots, a_{1p})$  is its corresponding eigenvector
- $\text{Var}(Y_2)$  = the 2nd largest eigenvalue of **the covariance matrix  $\text{Cov}(\mathbf{X})$** , and  $(a_{21}, a_{22}, \dots, a_{2p})$  is its corresponding eigenvector
- $\text{Var}(Y_k)$  = the  $k$ th largest eigenvalue of **the covariance matrix  $\text{Cov}(\mathbf{X})$** , and  $(a_{k1}, a_{k2}, \dots, a_{kp})$  is its corresponding eigenvector

# 分類熱門照片

	A	B	C	D	E	F	G	H	I
1	類別	ISO	光圈	快門	焦距	相機型號	上傳時間	累計人氣	
2	1	50	2.4	0.042	3	1	13.88202186	3876	
3	1	80	2.4	0.042	3	1	13.88202189	1650	
4	1	75	2.8	0.042	3	1	12.55760815	1184	
5	1	640	2.4	0.067	4	1	13.52986414	1178	
6	1	71	2.8	0.042	3	1	12.55760838	957	
7	1	73	2.8	0.059	3	1	12.55760834	913	
8	1	72	2.8	0.059	3	1	12.55760825	825	
9	1	74	2.8	0.05	3	1	12.5576083	805	
10	1	72	2.8	0.033	3	1	12.55760819	778	
11	1	1000	3.5	0.02	18	2	13.02962592	708	
12	1	79	2.8	0.05	3	1	12.65794278	594	
13	1	86	2.8	0.067	3	1	12.65794276	568	
14	1	220	2.8	0.067	3	1	12.65794273	405	
15	1	80	2.8	0.033	3	1	12.65794275	388	
16	1	1016	2.8	0.1	3	1	12.79276768	385	
17	1	640	2.8	0	28	2	12.55523557	358	

(RMD\_example 11.1)

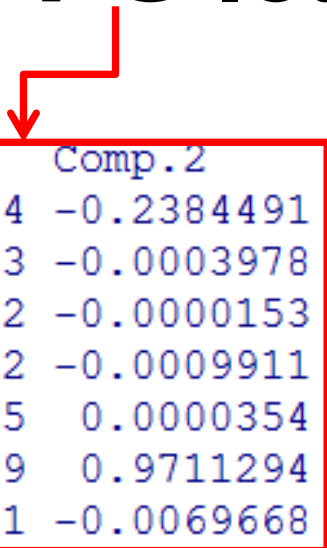
Variable	Description
類別	1=女生個人, 2=男生個人
ISO	照片的感光度，數值越高，對光越敏感
光圈	光圈值
快門	快門速度 (秒)
焦距	毫米 (mm)
相機型號	0, 1, 2
上傳時間	UNIX 時間戳，與1970年1月1日00:00:00的秒差 $\times 10^{-8}$
累計人氣	Hits

# 分類熱門照片 PC loadings

- $Y_1 = 0.97(\text{ISO}) - 0.00015(\text{光圈}) + 0.000019(\text{快門}) + 0.000083(\text{焦距}) + 0.000001(\text{相機型號}) + 0.24(\text{上傳時間}) - 0.0095(\text{累積人氣})$
- $Y_2 = -0.24(\text{ISO}) - 0.00040(\text{光圈}) - 0.000015(\text{快門}) - 0.000099(\text{焦距}) + 0.000035(\text{相機型號}) + 0.97(\text{上傳時間}) - 0.0070(\text{累積人氣})$

# PC loadings

Loadings:



	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
ISO	0.9711254	-0.2384491	0.0075776	0.0003176	-0.0000632
光圈	-0.0001543	-0.0003978	-0.0000032	-0.0276181	-0.9981672
快門	0.0000192	-0.0000153	0.0000031	0.0006682	-0.0038335
焦距	0.0000832	-0.0009911	0.0002883	-0.9996118	0.0273896
相機型號	0.0000015	0.0000354	0.0001029	-0.0034608	0.0538259
上傳時間	0.2383799	0.9711294	0.0090343	-0.0009512	-0.0003986
累計人氣	-0.0095131	-0.0069668	0.9999304	0.0002946	-0.0000126

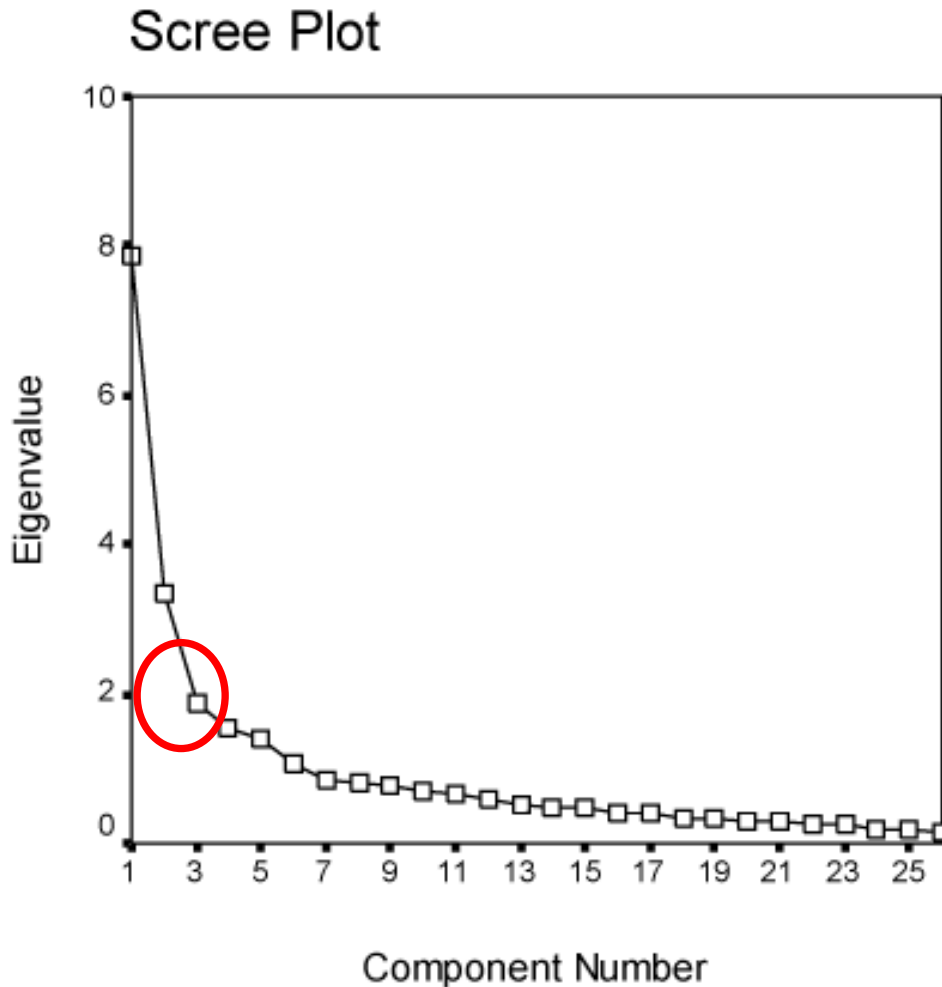
	Comp.6	Comp.7
ISO	0.0000109	0.0000227
光圈	0.0537373	0.0034113
快門	-0.0073856	-0.9999652
焦距	-0.0049467	-0.0007364
相機型號	0.9985155	-0.0075836
上傳時間	-0.0000176	-0.0000093
累計人氣	-0.0001011	0.0000040

(RMD\_example 11.2)

# The number of PCs

- The amount of total variance explained
  - If most (e.g., 80% to 90%) of the total population variance can be explained by the first  $k$  principle components ( $k < p$ ), then these  $k$  principle components can replace the original  $p$  variables without much loss of information (total variance).
- The relative size of eigenvalues (**variances**)
  - Scree plot
- Subject-matter interpretation of the components

# Scree plot



- The point at which the remaining eigenvalues are relatively small and all about the same size (an elbow (bend) in the scree plot).
- In the above plot, an elbow occurs at about  $k=3 \rightarrow$  choose **TWO** principle components.



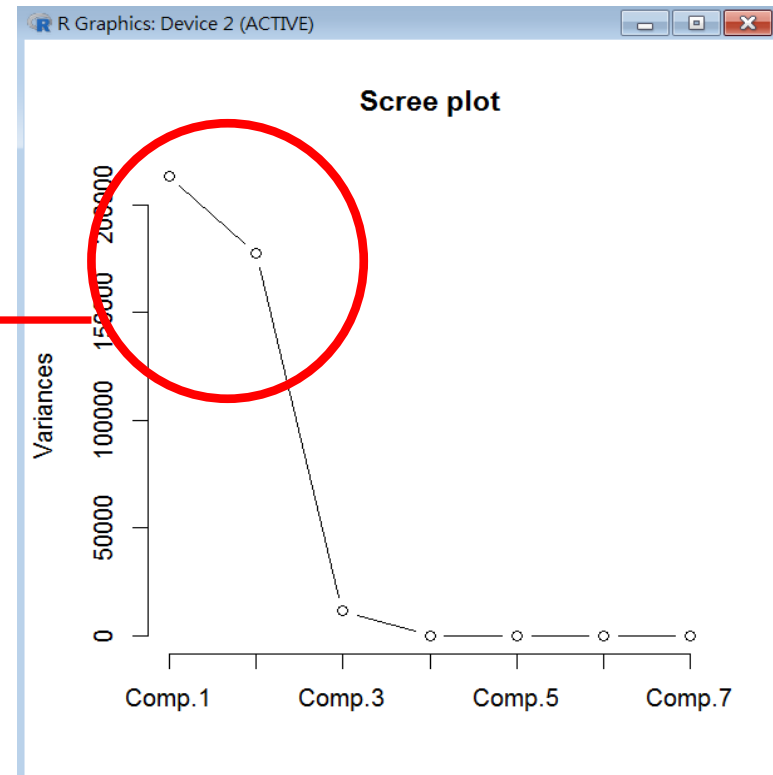
# 分類熱門照片:# of PCS

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	461.89174	421.5308688	106.98632088	5.857071e+00
Proportion of Variance	0.53003	0.4414473	0.02843653	8.522776e-05
Cumulative Proportion	0.53003	0.9714774	0.99991388	9.999991e-01
	Comp.5	Comp.6	Comp.7	
Standard deviation	5.228968e-01	2.88975e-01	3.199553e-02	
Proportion of Variance	6.792850e-07	2.07463e-07	2.543307e-09	
Cumulative Proportion	9.999998e-01	1.000000e+00	1.000000e+00	

2 PCs with the  
amount of total  
variance  
explained > 90%

Scree plot with an  
elbow occurs at  
about  $k=3$



# PCA from standardized variables

- Variables should probably be standardized **if they are measures on scales with widely differing ranges.**
- Example:
  - $X_1$ : annual sales in the \$10,000 to \$350,000 range
  - $X_2$ : the ratio (net annual income/total assets) that falls in the 0.01 to 0.6 range
- No standardization: the 1st principle component will have a heavy weighting of  $X_1$
- Standardization:  $X_2$  will play a larger role

# PCA from standardized variables

- Standardizing observed data:  $Z_1 = \frac{X_1 - \mu_1}{\sigma_1}$ ,  $Z_2 = \frac{X_2 - \mu_2}{\sigma_2}$ , ...

$$W_1 = b_{11}Z_1 + b_{12}Z_2 + \cdots + b_{1p}Z_p$$

- $W_2 = b_{21}Z_1 + b_{22}Z_2 + \cdots + b_{2p}Z_p$

$\vdots$

$$W_k = b_{k1}Z_1 + b_{k2}Z_2 + \cdots + b_{kp}Z_p$$

# PCA from standardized variables

- $\text{Var}(W_1)$  = the largest eigenvalue of **the correlation matrix  $\text{Cor}(\mathbf{X})$** , and  $(b_{11}, b_{12}, \dots, b_{1p})$  is its corresponding eigenvector
- $\text{Var}(W_2)$  = the 2nd largest eigenvalue of **the correlation matrix  $\text{Cor}(\mathbf{X})$** , and  $(b_{21}, b_{22}, \dots, b_{2p})$  is its corresponding eigenvector
- $\text{Var}(W_k)$  = the  $k$ th largest eigenvalue of **the correlation matrix  $\text{Cor}(\mathbf{X})$** , and  $(b_{k1}, b_{k2}, \dots, b_{kp})$  is its corresponding eigenvector

# 分類熱門照片 standardized PCs

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.2154626	1.1551005	1.0297473	1.0079490	0.9408723
Proportion of Variance	0.2110499	0.1906082	0.1514828	0.1451373	0.1264630
Cumulative Proportion	0.2110499	0.4016581	0.5531409	0.6982782	0.8247412
	Comp.6	Comp.7			
Standard deviation	0.80580979	0.7599227			
Proportion of Variance	0.09276134	0.0824975			
Cumulative Proportion	0.91750250	1.0000000			

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
ISO	-0.0714726	0.6044901	0.4736265	-0.0812640	-0.1886824
光圈	0.6778339	-0.0083201	-0.0088570	-0.0707311	-0.0199902
快門	0.0633094	0.7067502	0.0454794	0.1313349	0.1457447
焦距	0.4731235	-0.1996835	0.4939576	-0.2235479	-0.4076155
相機型號	-0.0770545	-0.2603461	0.6730539	0.2121282	0.6414747
上傳時間	-0.5454875	-0.1516771	0.2716848	-0.1781902	-0.4716178
累計人氣	0.0639136	-0.0660834	0.0521810	0.9189236	-0.3777491

	Comp.6	Comp.7
ISO	-0.5171128	-0.3091128
光圈	0.3044372	-0.6650632
快門	0.5993112	0.3110681
焦距	0.0236462	0.5249793
相機型號	0.1026508	-0.0790929
上傳時間	0.5186529	-0.2871392
累計人氣	-0.0253233	-0.0326948

5 PCs

(RMD\_example 11.3)

# **Factor analysis**

# Factor analysis

- The purpose of factor analysis is to describe the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities call **factors**.
- Factor analysis is motivated by
  - variables can be grouped by their correlation
  - all variables within a particular group are highly correlated among themselves
  - variables have relatively small correlations with variables in a different group

# Orthogonal factor model

- Observed data:  $X = (X_1, X_2, \dots, X_p)$
- $X_1, X_2, \dots, X_p$  are linearly dependent on a few **unobservable** random variables  $F_1, F_2, \dots, F_m$  (common factor)
- $p > m$



$$X_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1$$

$$X_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2$$

$\vdots$

$$X_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_p + \varepsilon_p$$

(in a matrix fom:  $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$ )

$\mu_i$  = mean of variable  $i$

$\varepsilon_i$  =  $i$ th error

$F_j$  =  $j$ th common factor

$\ell_{ij}$  = loading of the  $i$ th variable on the  $j$ th factor

- Unobservable random vectors  $F_j$ 's and  $\varepsilon_i$ 's are mutually independent

# 分類熱門照片：factor analysis

Uniquenesses:

	ISO	光圈	快門	焦距	相機型號	上傳時間	累計人氣
	0.0912872	0.4054439	0.7589743	0.6987590	0.9552639	0.7812365	0.9982685

Loadings:

	Factor1	Factor2	Factor3
ISO	0.9531224	0.0108353	0.0123782
光圈	-0.0684252	0.7676951	-0.0227622
快門	0.3126406	0.0768856	-0.3706365
焦距	0.0137846	0.4001137	0.3754566
相機型號	-0.0113459	-0.0559101	0.2036606
上傳時間	0.0454631	-0.3968105	0.2433759
累計人氣	-0.0336443	0.0209563	-0.0131131

Factor  
loadings

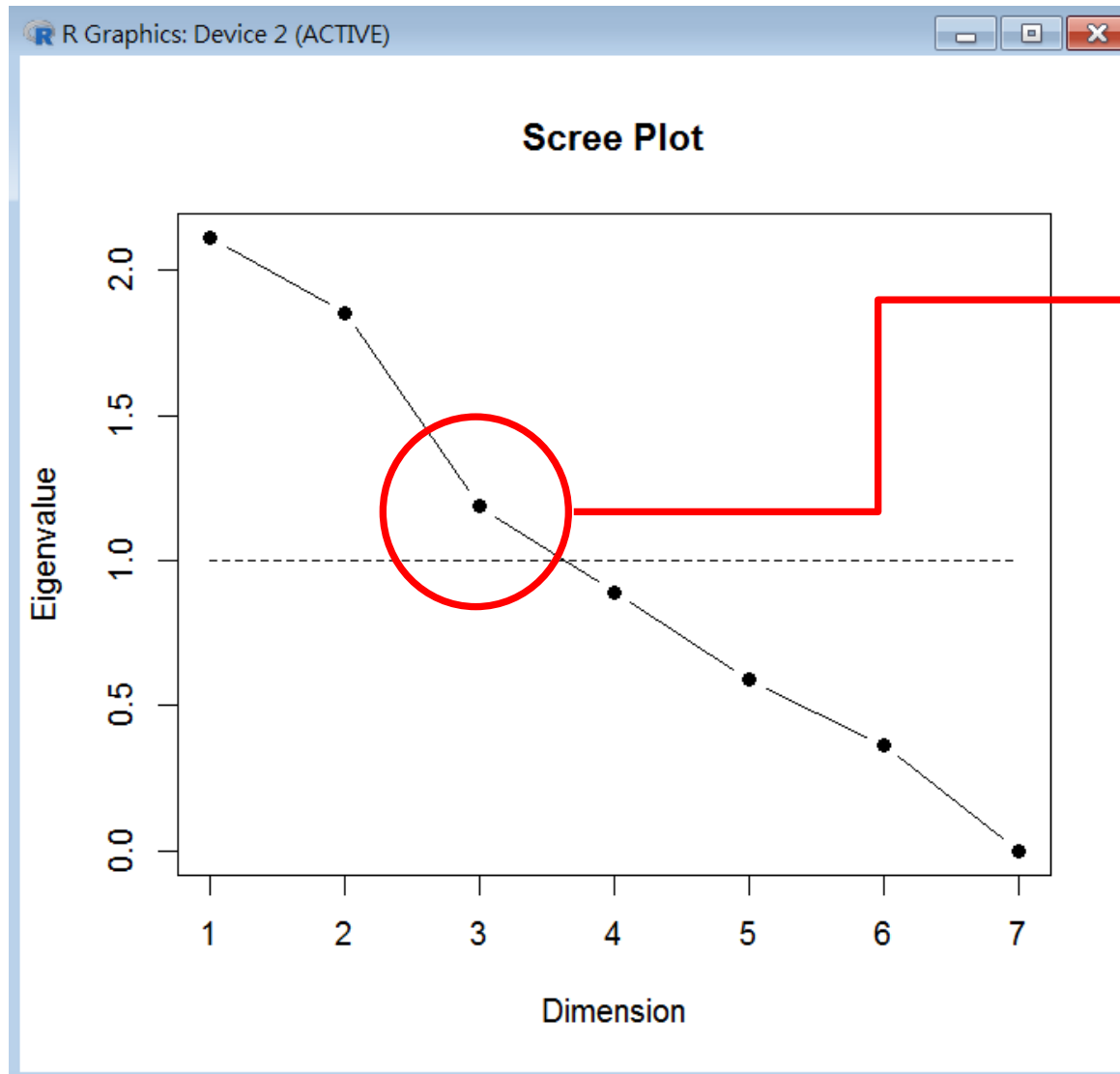
	Factor1	Factor2	Factor3
SS loadings	1.0143859	0.9164991	0.3798918
Proportion Var	0.1449123	0.1309284	0.0542703
Cumulative Var	0.1449123	0.2758407	0.3301110

(RMD\_example 11.4)

# The number of factors

- Use criteria similar to those used in PCA
  - The amount of total variance explained until a suitable proportion
  - The relative size of eigenvalues (variances): scree plot of **the covariance matrix**
  - Set  $m$  equal to **the number of eigenvalues of the correlation matrix greater than one**
  - Subject-matter interpretation of the factors

# 分類熱門照片：# of factors



The number of eigenvalues of the correlation matrix greater than one = 3

# Factor rotation

- Factor loadings  $L$  are determined only up to an orthogonal matrix  $T$  ( $TT' = T'T = I$ ):

$$X - \mu = LF + \varepsilon = L(TT')F + \varepsilon = L^*F^* + \varepsilon$$

- $L^* = LT$  (rotated loadings)  
 $F^* = T'F$  (rotated factors)

# Why factor rotation?

- Original loadings may not be readily interpretable, usual rotate them until a **simpler** structure is achieved.
- Ideally, we like to see a pattern of loadings such that each variable loads highly on a single factor and has small loadings on the remaining factors:

$$X_1 - \mu_1 = \underbrace{\ell_{11}}_{\text{large}} F_1 + \underbrace{\ell_{12}}_{\text{small}} F_2 + \cdots + \underbrace{\ell_{1m}}_{\text{small}} F_m + \varepsilon_1$$

$$X_2 - \mu_2 = \underbrace{\ell_{21}}_{\text{small}} F_1 + \underbrace{\ell_{22}}_{\text{large}} F_2 + \cdots + \underbrace{\ell_{2m}}_{\text{small}} F_m + \varepsilon_2$$

⋮

$$X_p - \mu_p = \underbrace{\ell_{p1}}_{\text{small}} F_1 + \underbrace{\ell_{p2}}_{\text{small}} F_2 + \cdots + \underbrace{\ell_{pm}}_{\text{large}} F_p + \varepsilon_p$$

# Methods of factor rotation

- **Varimax rotations**
  - Select the orthogonal transformation  $T$  that maximizes the **spreading out** of the squares of the loadings in each factor.
  - We hope to find groups of large and negligible coefficients in any column of  $L^*$

- **Oblique (non-orthogonal) rotations**
  - Many investigators in social sciences consider the rotated factors are not necessarily to be independent.
  - Orthogonal rotations- a rigid rotation of the coordinate axes (factors) such that the rotated axes pass as closely the clusters of variables as possible.
  - Oblique rotation- a nonrigid rotation such that the rotated axes (no longer perpendicular) pass through the clusters of variables.



# 分類熱門照片：

## rotated vs. unrotated loadings

Loadings:

	Factor1	Factor2	Factor3
ISO	0.9325038	0.0598971	0.1885787
光圈	-0.1482474	0.7303608	0.1978678
快門	0.2281710	-0.0083560	0.4346206
焦距	0.0437581	0.4850338	-0.2531325
相機型號	0.0327628	-0.0007951	-0.2089456
上傳時間	0.1299696	-0.3160405	-0.3193487
累計人氣	-0.0374397	0.0151079	0.0106340

Rotated  
(varimax)

Loadings:

	Factor1	Factor2	Factor3
ISO	0.9531224	0.0108353	0.0123782
光圈	-0.0684252	0.7676951	-0.0227622
快門	0.3126406	0.0768856	-0.3706365
焦距	0.0137846	0.4001137	0.3754566
相機型號	-0.0113459	-0.0559101	0.2036606
上傳時間	0.0454631	-0.3968105	0.2433759
累計人氣	-0.0336443	0.0209563	-0.0131131

Unrotated

(RMD\_example 11.4)

# Factor scores

- Estimated values of the common factors are called **factor scores**.
- Factor scores can be used for diagnostic purposes and inputs to a subsequent analysis.
- Factor scores are not estimates of unknown parameters. Factor scores are estimates of unobserved random factor vectors  $F_1, F_2, \dots, F_m$ .

# 分類熱門照片：factor scores

```
> factfit$scores
```

	Factor1	Factor2	Factor3
[1,]	-0.744478512	-0.4390968203	-0.2421767848
[2,]	-0.572821108	-0.5262584971	-0.4302286198
[3,]	-0.843173384	0.2953278622	0.5238103500
[4,]	0.514612877	-0.2920302012	0.0908727086
[5,]	-0.839981043	0.2847687444	0.5027908527
[6,]	-0.847735204	0.2585411098	0.6602863806
[7,]	-0.845381209	0.2545780777	0.6522816904
[8,]	-0.832918679	0.2673538481	0.5658902695
[9,]	-0.821629459	0.2905749951	0.4021470393
[10,]	1.266750889	1.9190923884	-0.7707859163
[11,]	-0.795406138	0.2283314115	0.4833648686
[12,]	-0.793919006	0.2040492470	0.6437493174
[13,]	-0.514118004	0.2288582020	0.6643793857
[14,]	-0.769122005	0.2446826331	0.3046530819
[15,]	1.096261264	0.3261509954	1.0949048259
[16,]	0.680574299	1.7287376488	-1.5034657744
[17,]	-0.458116823	0.2322163316	0.6653574986
[18,]	1.100411881	0.3207546911	1.0836600315
[19,]	1.099298921	0.3235670407	1.0895351702
[20,]	-0.526639685	0.2218586079	0.6519107506
[21,]	-0.066986803	0.2568968261	0.6733340165
[22,]	0.072574391	-0.3867257670	-0.0505240060
[23,]	-0.003387440	0.2635979134	0.6801051329

Every  
observation  
(picture) has  
scores on all 3  
factors!

# Exploratory factor analysis (EFA)

- EFA is a technique within factor analysis whose goal is to identify the underlying relationships between measured variables (i.e., identifying **common factors**).
- It should be used when the researcher has no a priori hypothesis about factors or patterns of measured variables.

# Confirmatory factor analysis (CFA)

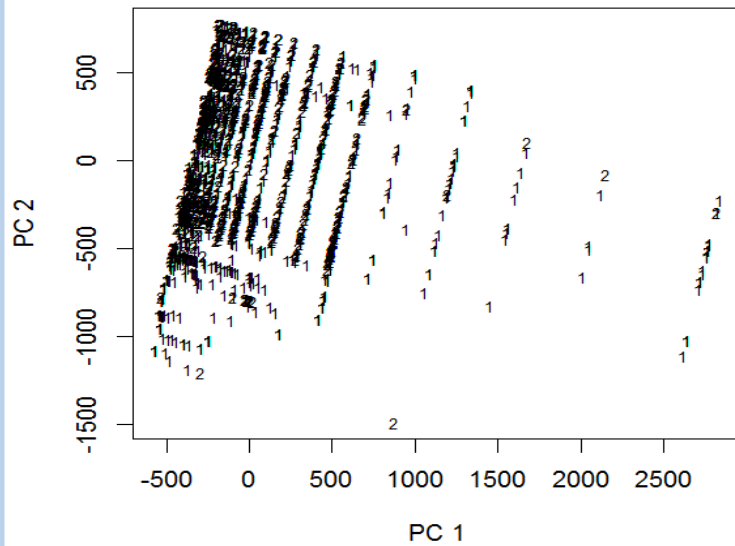
- CFA is a special form of factor analysis, it is used to test whether measures of a construct are consistent with a researcher's understanding of the nature of that construct (or factor).

1. In CFA, first develop a hypothesis about what factors he believes are underlying the measures he has used.
2. Impose constraints (e.g., zero loadings) on the model based on these a priori hypotheses. By imposing these constraints, the researcher is forcing the model to be consistent with his theory.
3. **Model fit measures** could then be obtained to assess how well the proposed model captured the covariance between all the measures in the model.
  - If the constraints the researcher has imposed on the model are inconsistent with the sample data, then the results of statistical tests of model fit will indicate a poor fit, and the model will be rejected.

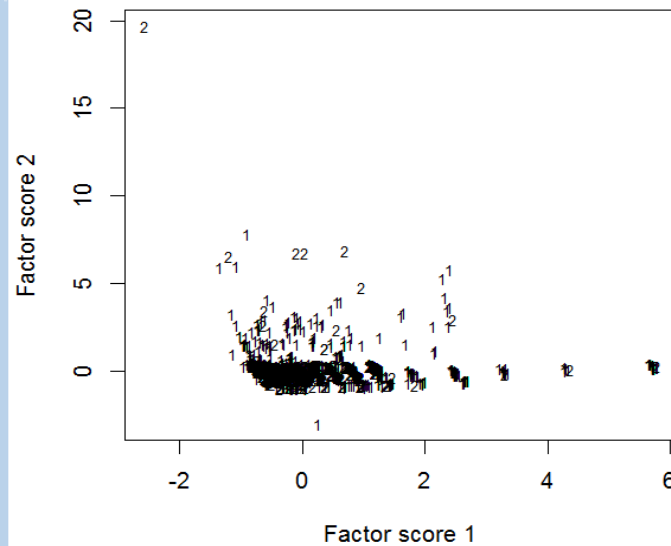
# Visualization of multivariate data

- Visualising and processing high-dimensional datasets, while still retaining as much of the variances in the data as possible.
- Can plot scores on the first two principal components or factors

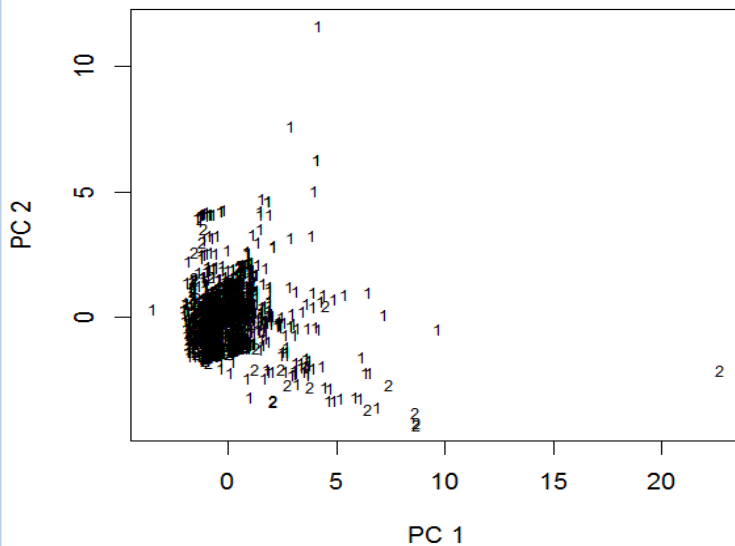
PCA with Cov



Factor analysis



PCA with Cor



分類熱門照片：  
**visualization**