Principal Components Analysis

(주성분 분석)

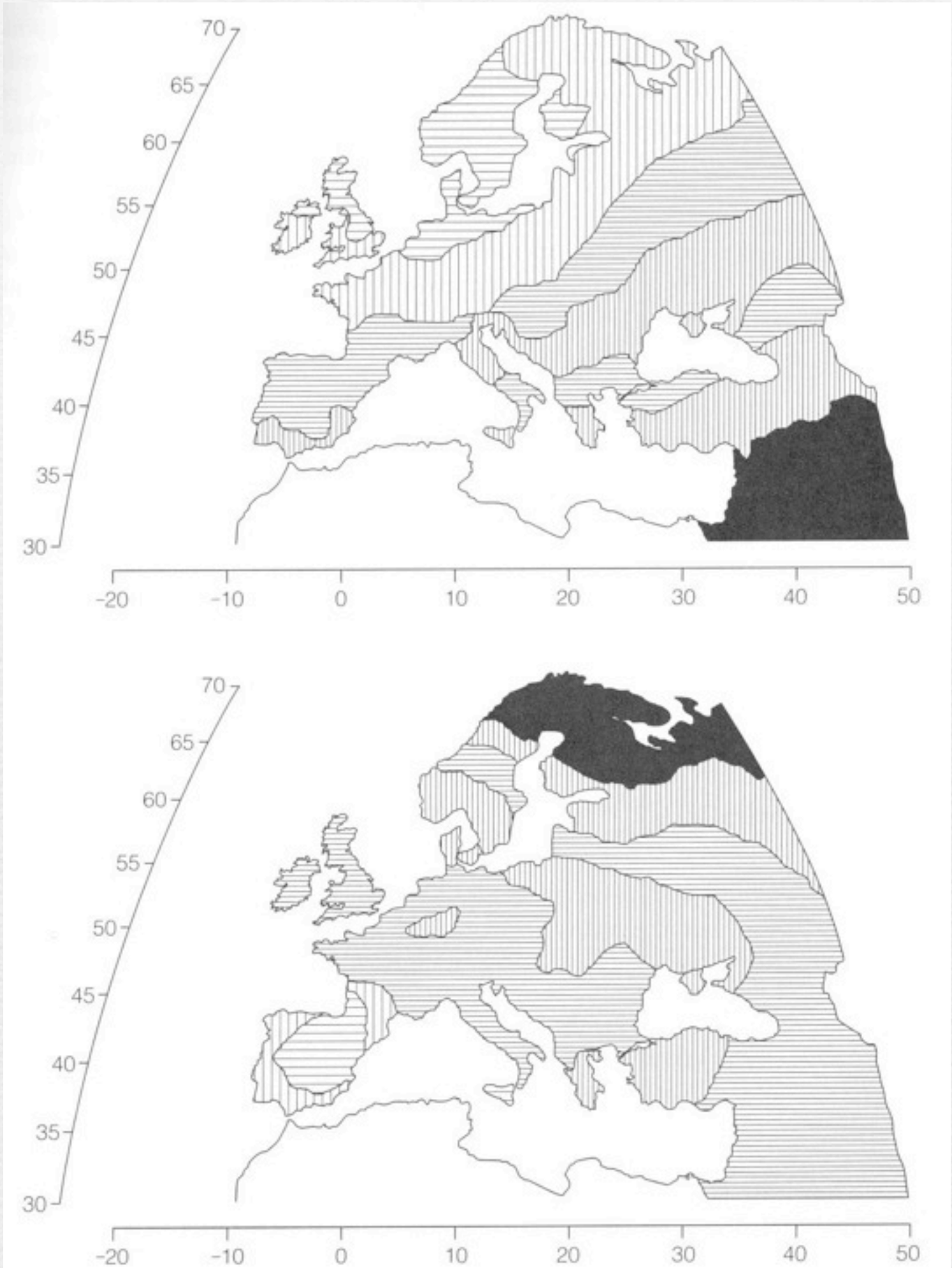
H. Park
HUFS

Motivation

- In this day, so many variables entering into analysis > _____
- often redundancy among dimensions > leading high level of _____ in regression
- need to _____ for subsequent analysis
- different from "_____"
  - ★ PCA = _____
  - ★ EFA = _____
- PC = uncorrelated linear combinations of the original variables explaining the variation the most.

Europe gene map
(Cavalli-Sforza,2000)

- 95 different genes throughout Europe & Middle East
- the first 3 PC explains more than 50%
- Instead looking at 95 maps, we can look at only 2-3 maps (save time & energy)
- darker area represents the higher score of PC


70
65
60
55
50
45
40
35
30
-20
-10
0
10
20
30
40
50

1st PC

=_____

70
65
60
55
50
45
40
35
30
-20
-10
0
10
20
30
40
50

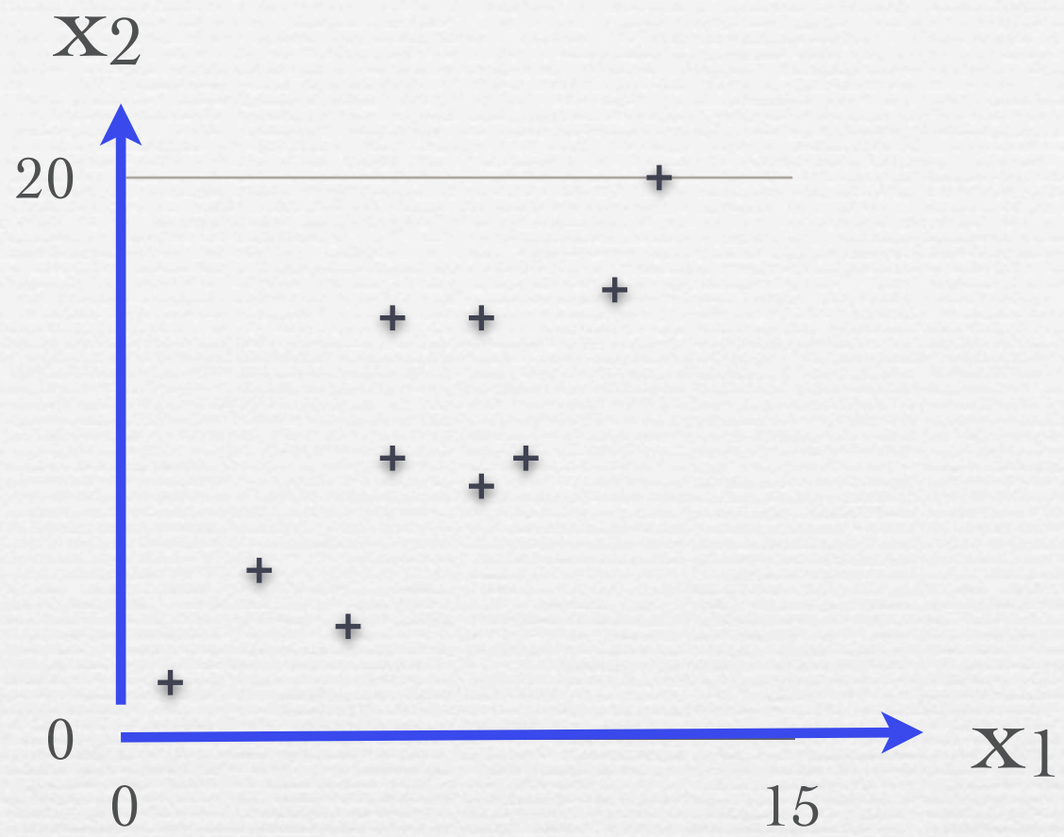
2nd PC

=_____

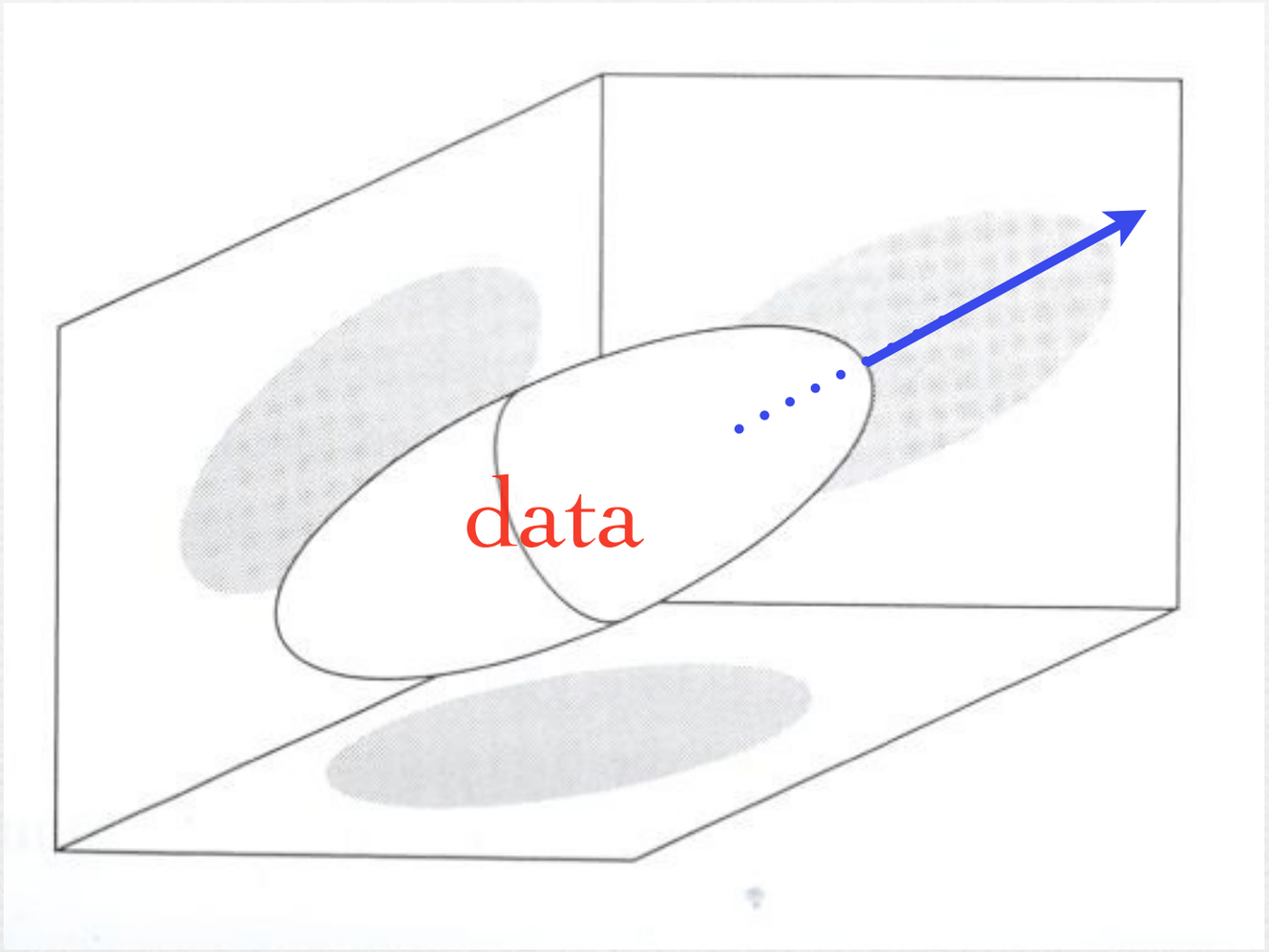
Principal Components Analysis (PCA)

- A vector space transform used to reduce multidimensional data to lower dimensions for analysis (Wikidepia)
- data 의 변동을 잘 설명해 주는 직교되는 벡터 =

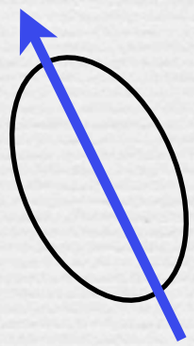
- Karl Pearson (1901) invented.
- Tool for EDA( )
- Used in _____
- Related to _____ Analysis



- 1st Principal Component ( ): data 의 변동을 제일 잘 나타내는 vector
- 2nd Principal Component ( ): 제1주성분에 의해 설명되지 않은 변동을 제일 잘 나타내면서 PC1과 직교되는 vector
- PC1 , PC2, PC3, ... are orthogonal each other
- PC 의 개수 =



서로 직각인 상자를 어떻게 움직이는가에 따라 그림자 크기가 달라진다

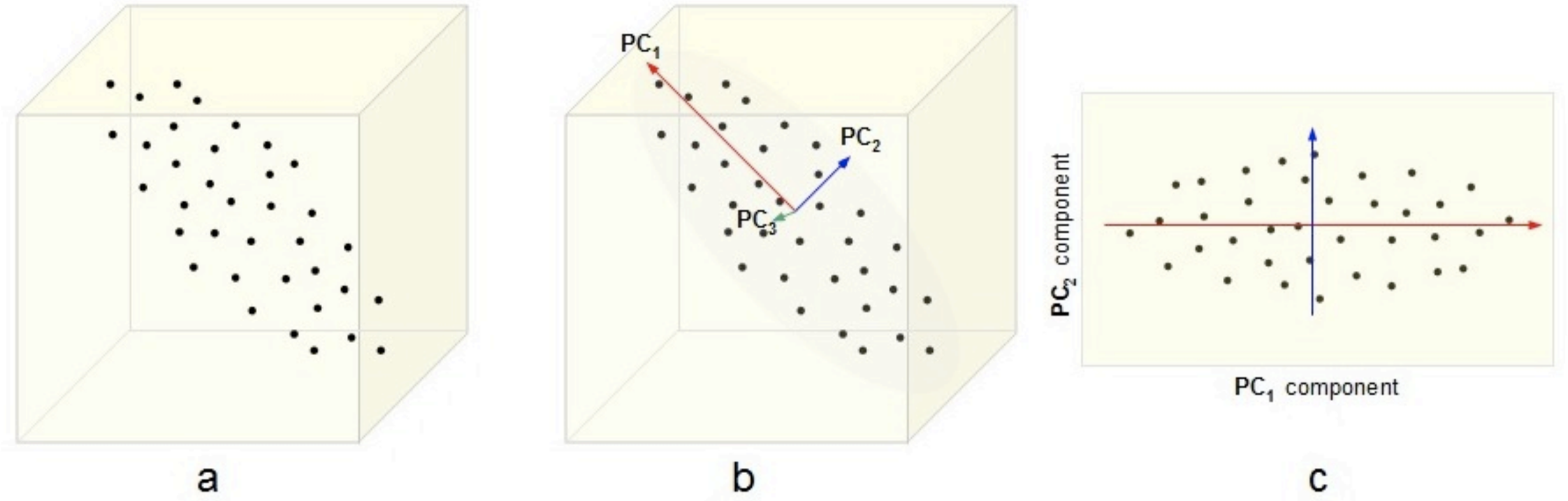


Data Reduction
(Dimension Reduction)

- 전체 자료의 변동을 몇 개의 PC로 설명가능할 때

x_1 , x_2 , … x_{10} →

→



PCA 의 활용

- x_1=math, x_2=eng., x_3 =science 의 종합성적?

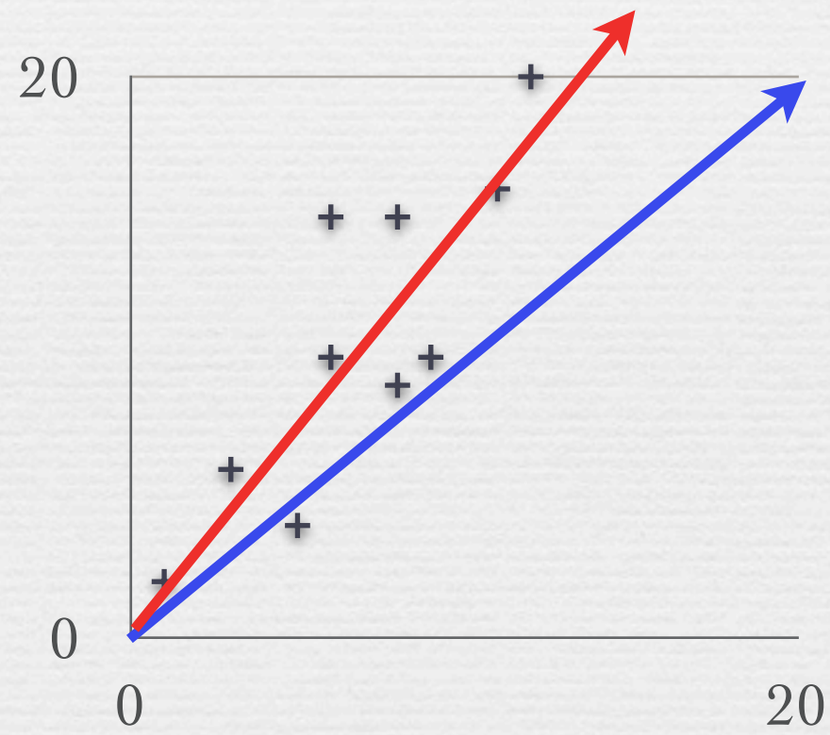
- 종합성적의 변별력을 높이려면?

- 소비자 물가지수 = a1(기름)+a2(학원)+a3(쌀)

- 중회귀분석 & 다중공선성

(Q) How to decide PC's ?

$$
\begin{aligned}
PC1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q = \boldsymbol{a}_1^T\boldsymbol{x} \\
PC2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q = \boldsymbol{a}_2^T\boldsymbol{x} \\
&\vdots \\
PCq &= a_{q1}x_1 + a_{q2}x_2 + \cdots + a_{qq}x_q = \boldsymbol{a}_q^T\boldsymbol{x}
\end{aligned}
$$



- In general,

How to get PC's

- centered $x_1, x_2, \ldots, x_q$ >> S_{qxq}
- eigenvalues of S_{qxq} =
- eigenvectors of S_{qxq} =
- i-th PC =

- (주의) $x_1, x_2, \ldots x_q$ 단위가 다른 경우,

$$PC_1 = 0.9x_1 + 0.001x_2 + 0.0029x_3 + \cdots - 0.0017x_q$$

Properties of PC's

- Variance :
- j-th PC 는 전체 변동의 % 설명
- $PC_1 + PC_2$ 는 전체의 % 설명
- 주성분점수 (PC Score)

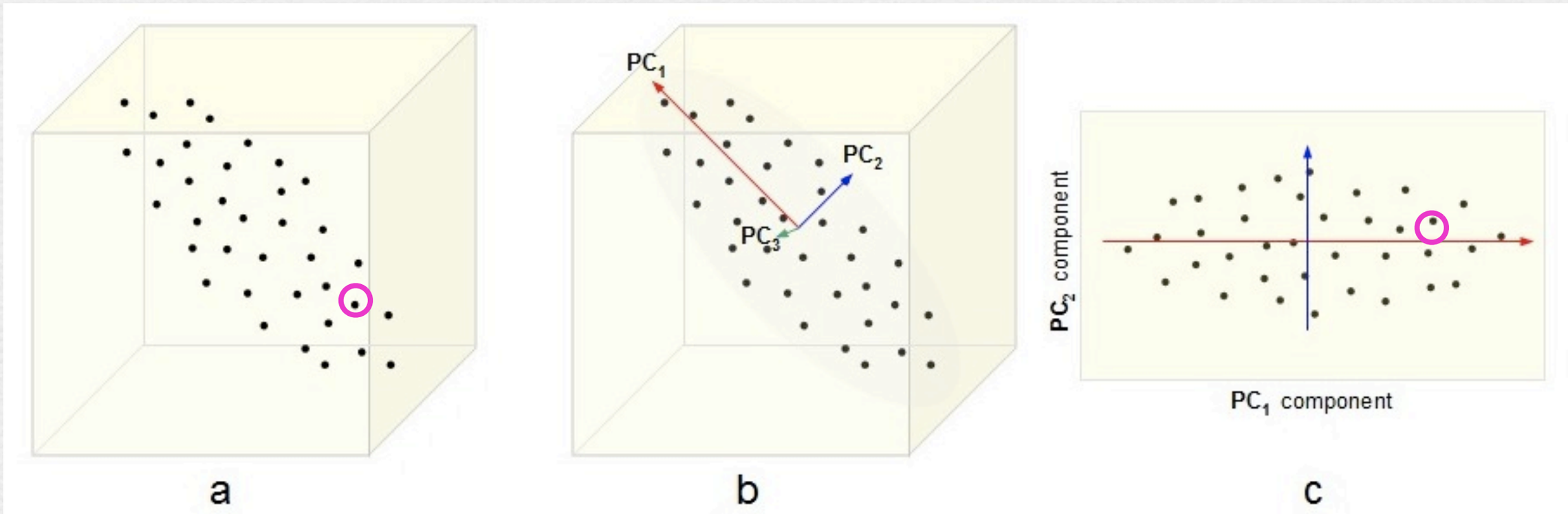
$$X_1,\ X_2,\ \cdots,\ X_q \longrightarrow Y_1,\ Y_2,\ \cdots,\ Y_m,\ (m \leq q)$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ \vdots & & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{iq} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \qquad \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ \vdots & & & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{im} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix}$$

PC Score (주성분점수)

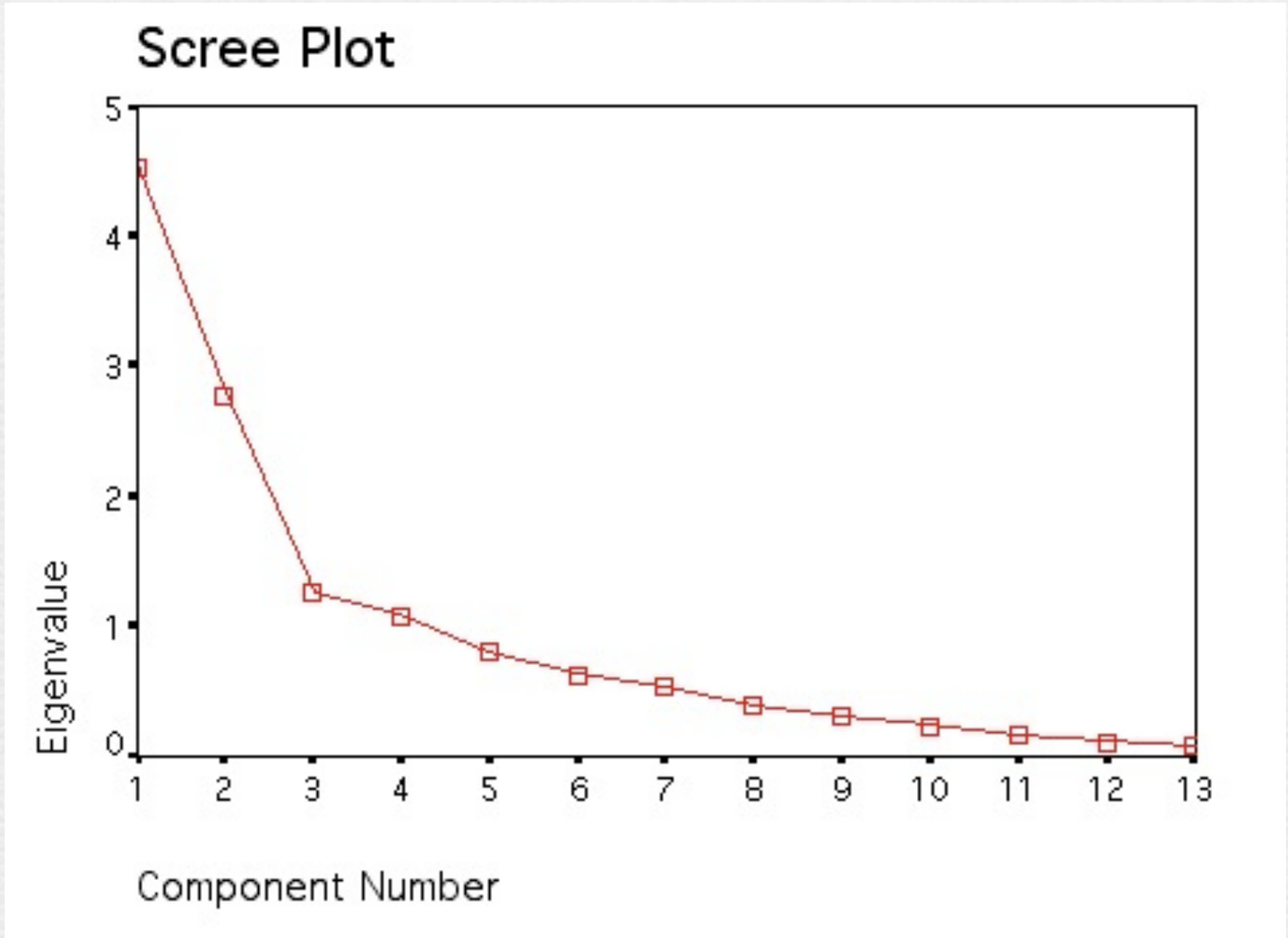
for i-th subject

$$
\begin{aligned}
y_{i1} &= \boldsymbol{a}_1^T \boldsymbol{x}_i \\
y_{i2} &= \boldsymbol{a}_2^T \boldsymbol{x}_i \\
&\vdots \\
y_{im} &= \boldsymbol{a}_1^T \boldsymbol{x}_i
\end{aligned}
$$



Scree Plot

- 주성분 개수를 결정하는 그래프



Example: PCA

(주성분 분석 예제)

(예1) 변수선택(Variable Selection)

여러 개의 변수 ($x_1, x_2, \ldots x_q$) 를 대표하는 몇 개(x_1, x_3, x_8) 만을 선택하는 방법

(1) ($x_1, x_2, \ldots x_q$) 의 분산행렬(혹은 상관행렬)을 사용하여 P.C. 를 구한다. (_____)

(2) P.C. 가운데 _____인 PC 만을 선택한다

(3) 선택된 P.C. 에 대하여 _____를 각각 선택한다.

example

$x_1,\ x_2,\ x_3,\ x_4 \longrightarrow\ ?$

$$
\begin{aligned}
PC1 &= 0.5x_1 + 0.6x_2 + 0.8x_3 - 0.2x_4, \quad \lambda_1 = 4.3 && > 0.7 \\
PC2 &= 0.8x_1 + 0.2x_2 - 0.5x_3 + 0.4x_4, \quad \lambda_2 = 1.4 && > 0.7 \\
PC3 &= 0.1x_1 - 0.6x_2 + 0.2x_3 + 0.02x_4, \quad \lambda_3 = 0.65 \\
PC4 &= 0.2x_1 - 0.2x_2 + 0.7x_3 + 0.5x_5, \quad \lambda_4 = 0.4
\end{aligned}
$$

선택

(예2) 육상기록 분석(track data)

- Men's National Track Records in 1984
- 55 countries
- 100m, 200m, 400m, 800m, 1500m,
- 5000m, 10000m, marathon

x1	x2	x4	x8	x15	x50	x100	mraton	country
10.39	20.81	46.84	1.81	3.70	14.04	29.36	137.72	argentin
10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30	australi
10.44	20.81	46.82	1.79	3.60	13.26	27.72	135.90	austria
10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95	belgium
10.28	20.58	45.91	1.80	3.75	14.68	30.55	146.62	bermuda
10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13	brazil
10.64	21.52	48.30	1.80	3.85	14.45	30.28	139.95	burma
10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15	canada
10.34	20.80	46.20	1.79	3.71	13.61	29.30	134.03	chile
. . .								

PCA in SAS

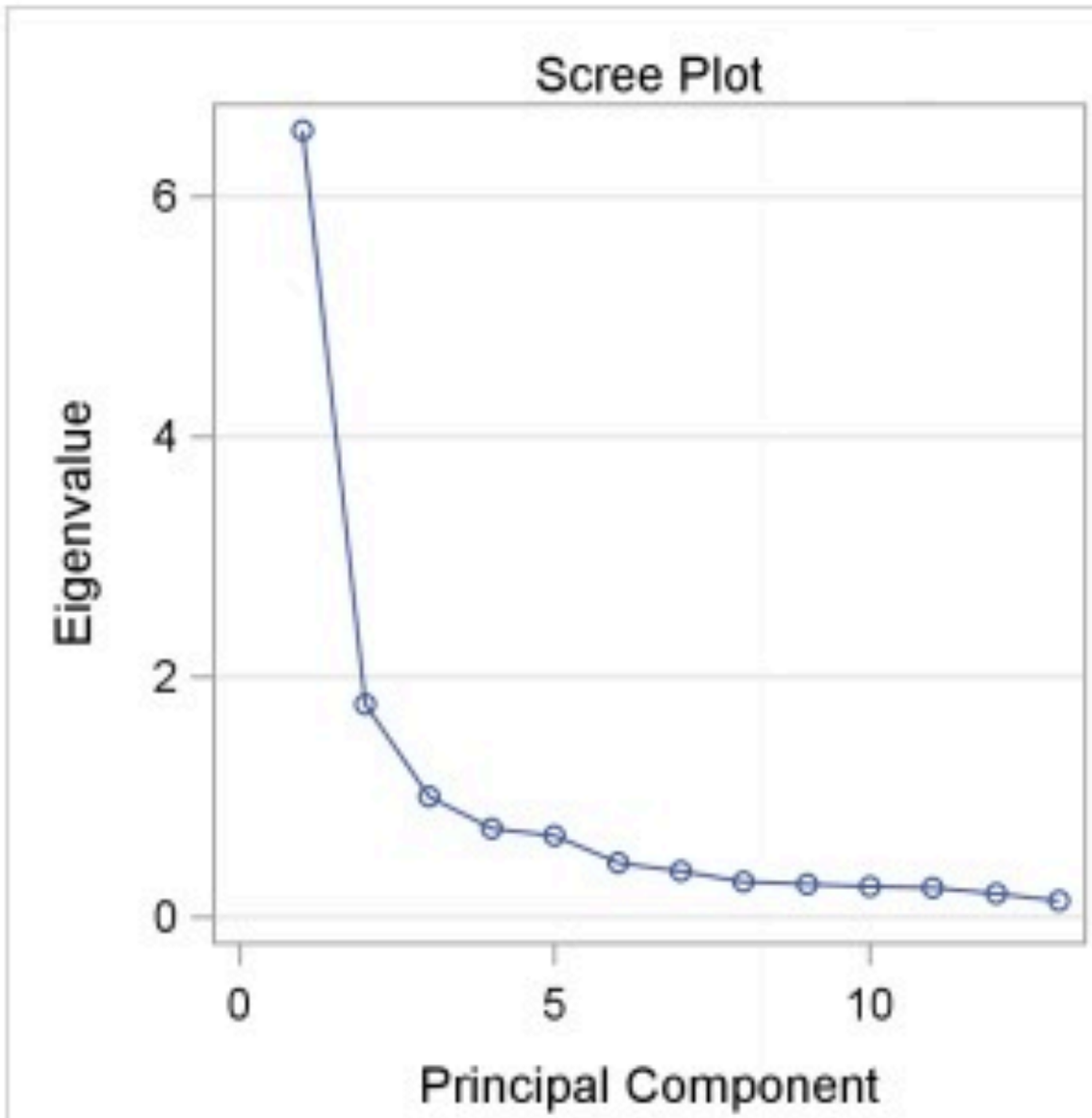
```
proc princomp N=2 Cov outstat=lambda
out=pca_score;

   var x1 x4 x15 x100 ;

run;

proc gplot data=pca_score;
      plot  prin1*prin2 ;
run;
```

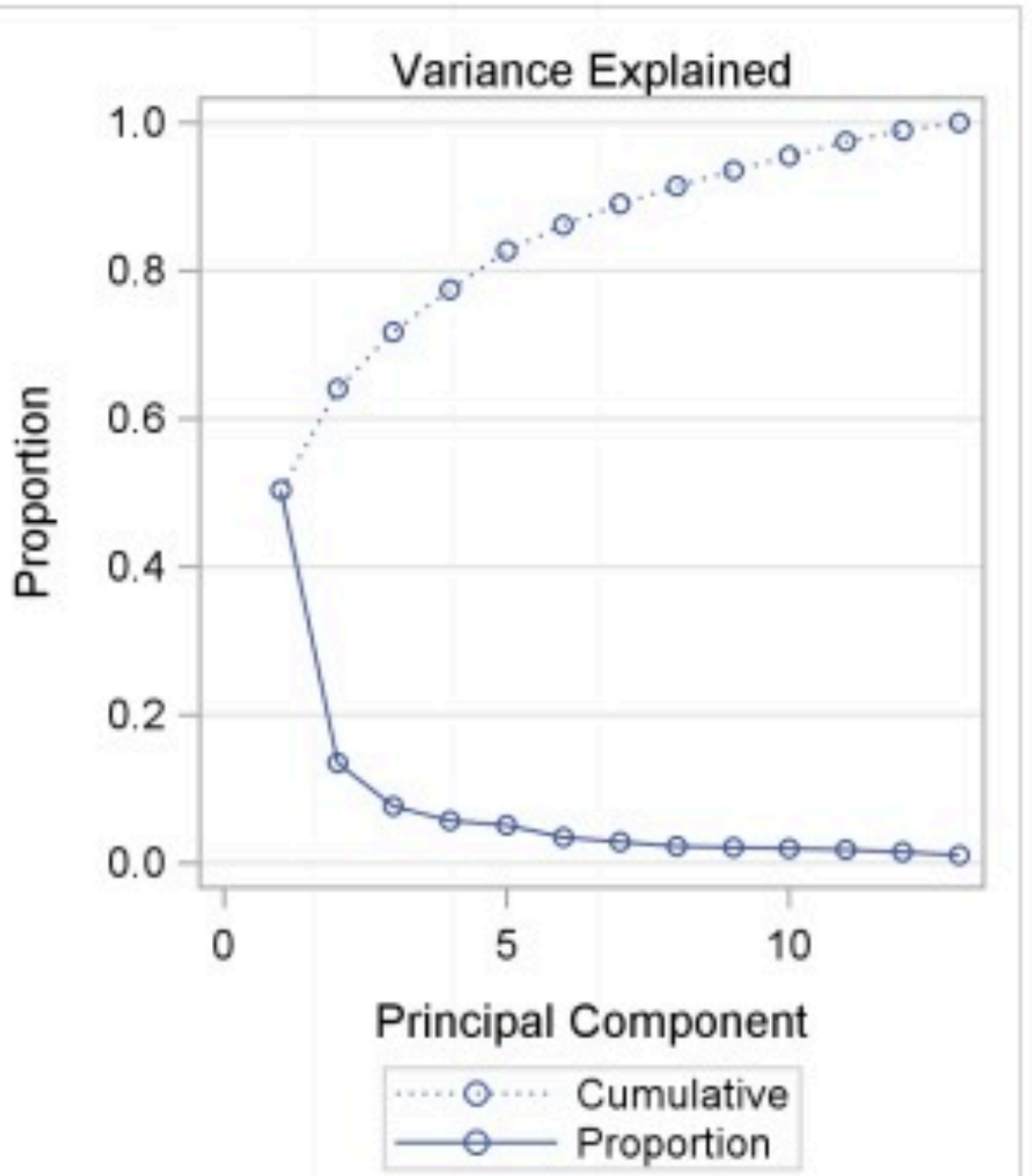
- N=2 >> _____
- Cov >> use S (otherwise ___ will be used)
- outstat >> _____ will be stored

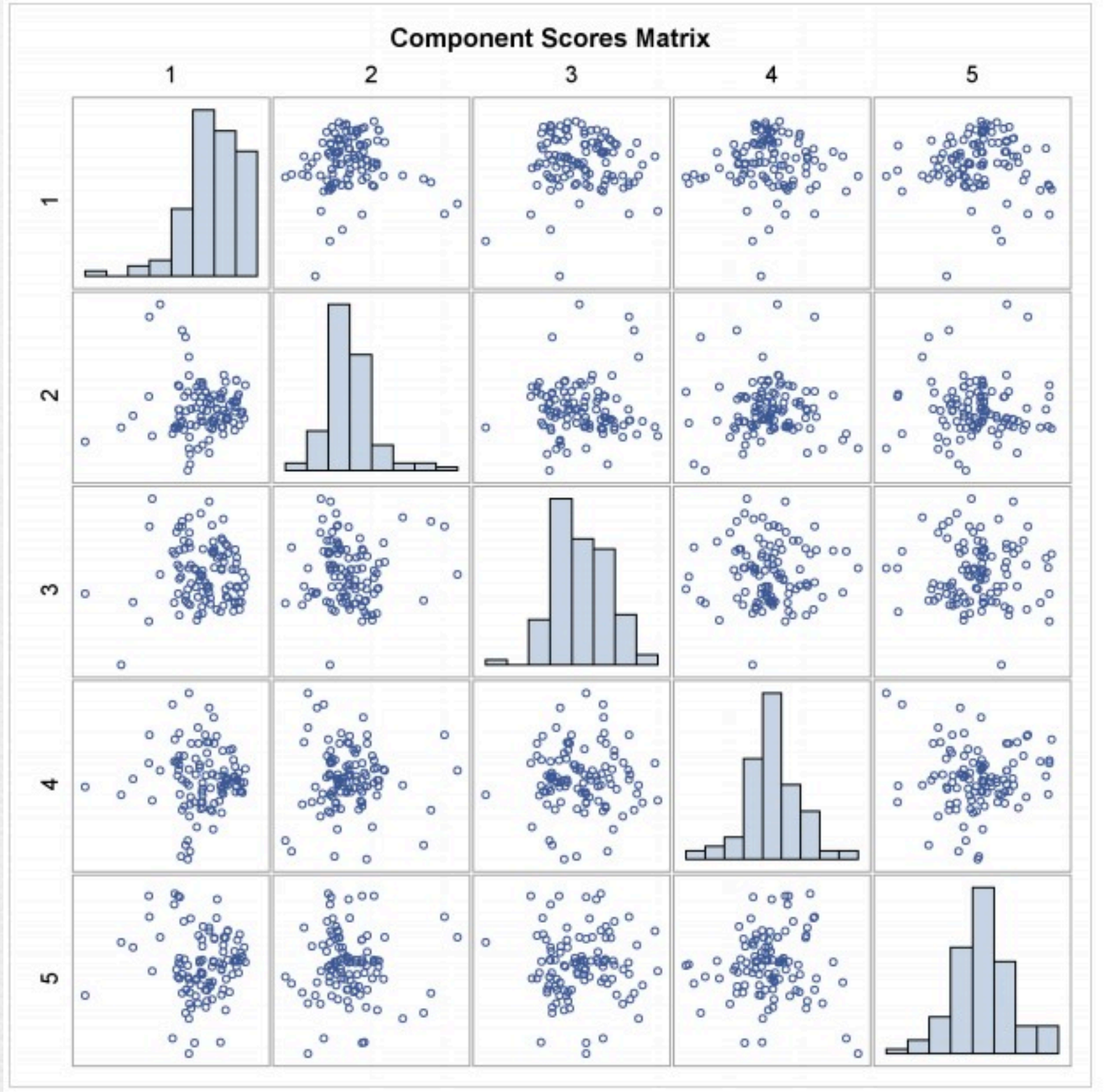
PCA in SAS (2)

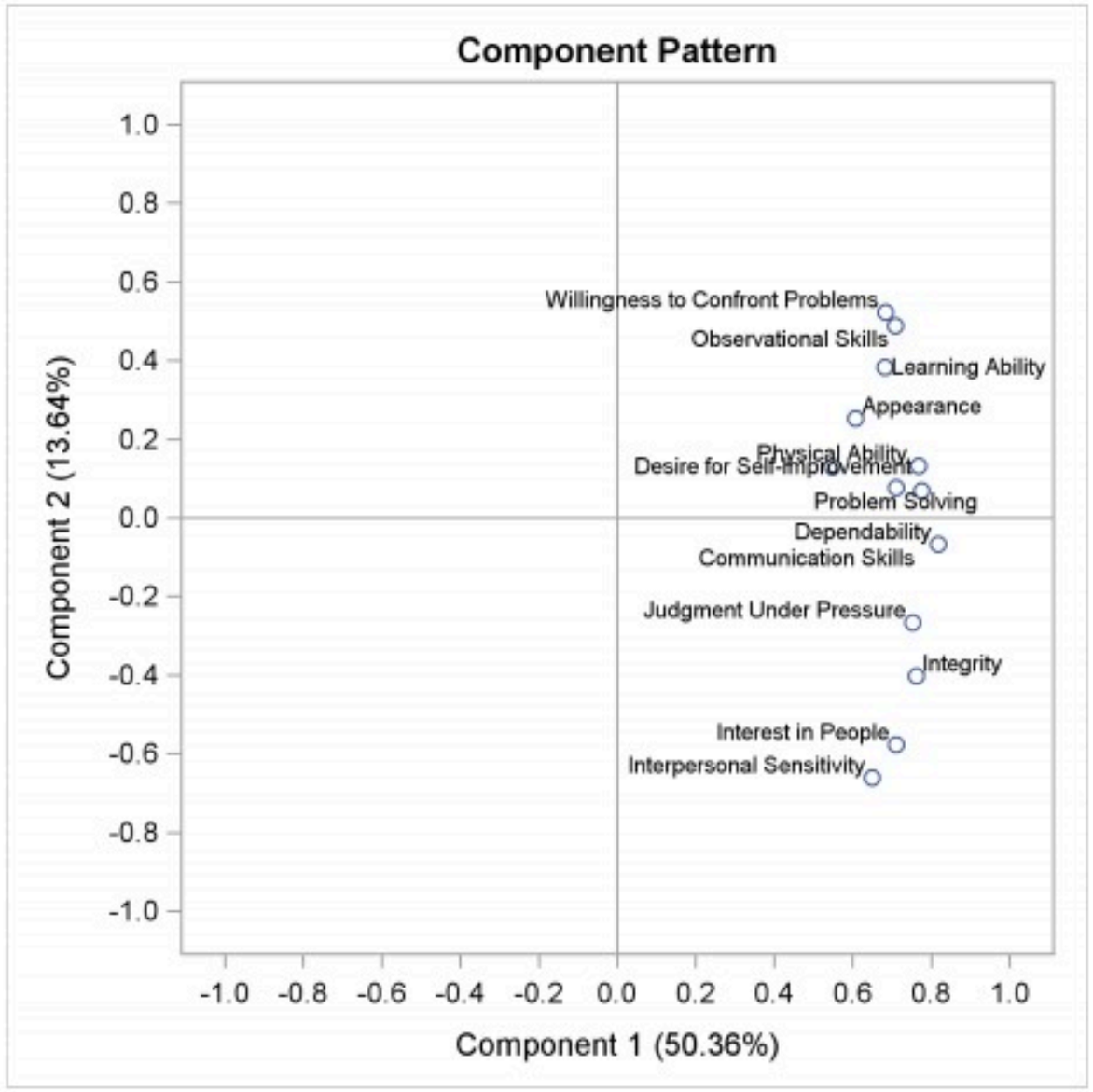
```
proc princomp N=5 Cov outstat=lambda

     plots(ncomp=3)=all ;

   var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13;

run;
```

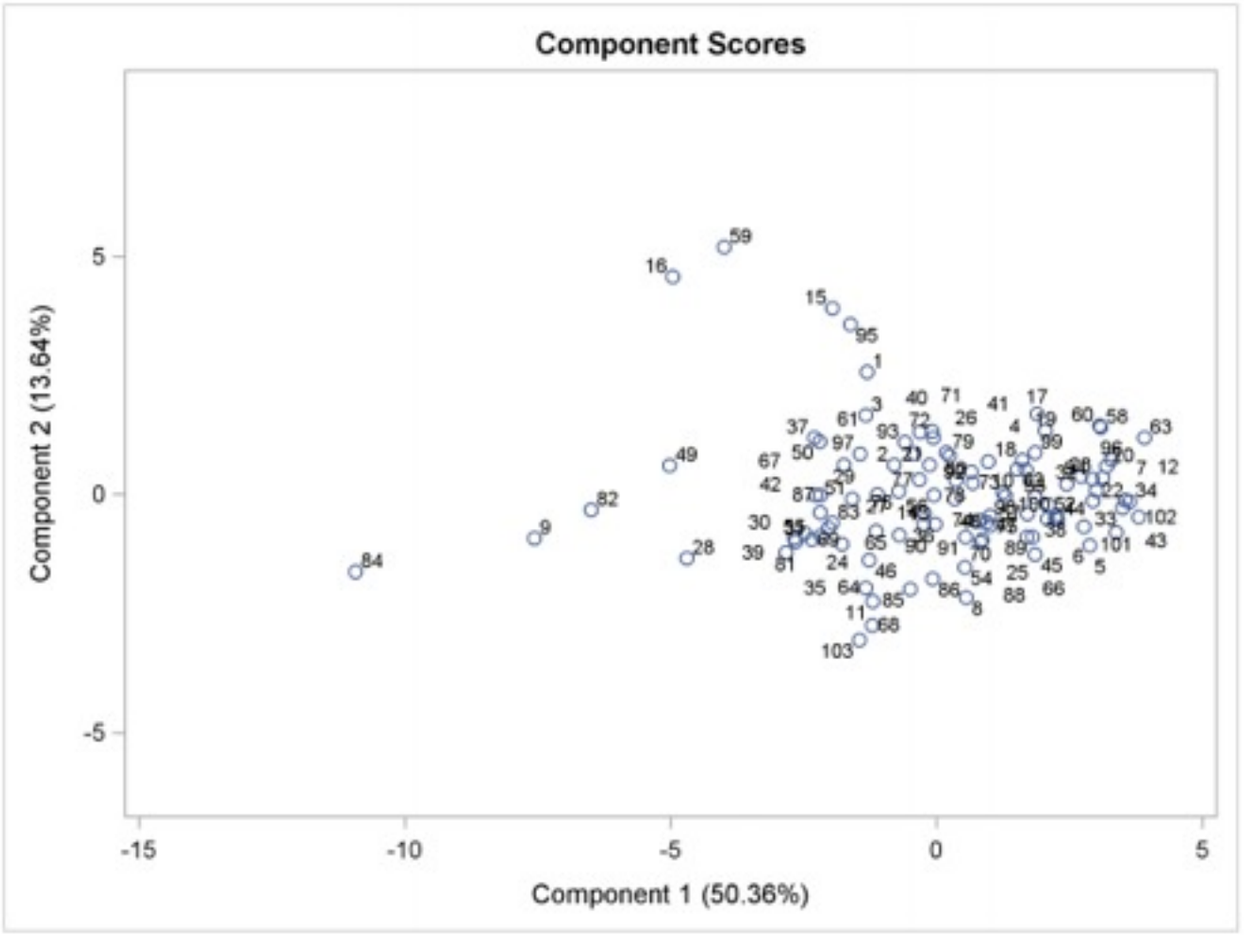
```
proc princomp N=5 Cov outstat=lambda

     plot=scree;

   var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13;

run;
```

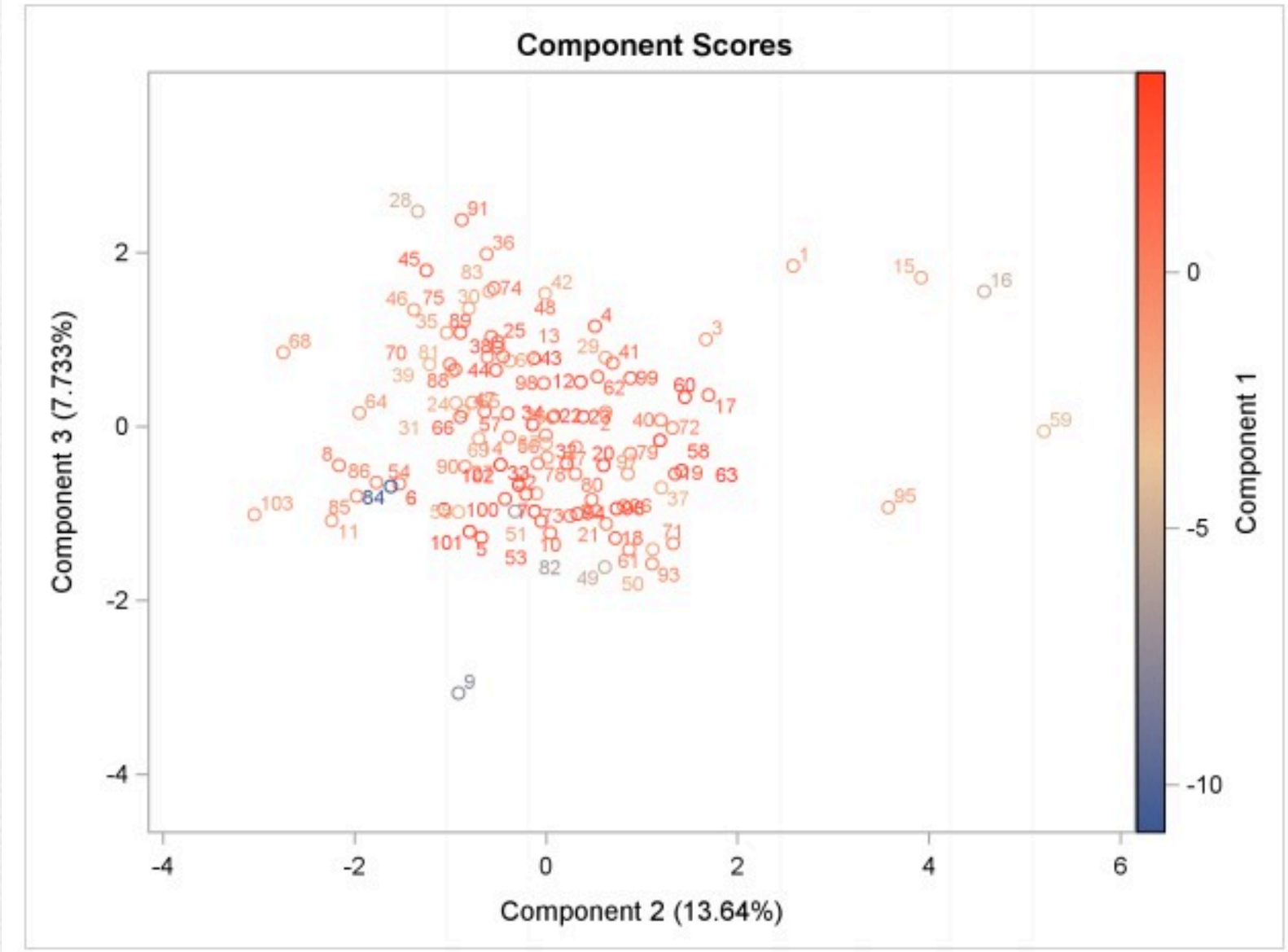

Scree Plot
Eigenvalue
6
4
2
0
0
5
10
Principal Component


Variance Explained
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
Principal Component
Cumulative
Proportion


Component Scores Matrix
1
2
3
4
5
1
2
3
4
5


Component Pattern
Component 2 (13.64%)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Willingness to Confront Problems
Observational Skills
Learning Ability
Appearance
Physical Ability
Desire for Self-improvement
Problem Solving
Dependability
Communication Skills
Judgment Under Pressure
Integrity
Interest in People
Interpersonal Sensitivity
-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
Component 1 (50.36%)


Component Scores
Component 2 (13.64%)
5
0
-5
-15
-10
-5
0
5
Component 1 (50.36%)


Component Scores
Component 3 (7.733%)
2
0
-2
-4
-4
-2
0
2
4
6
Component 2 (13.64%)
0
-5
-10
Component 1

results...

eigenvalues : 3.37 0.449 0.121 0.060

1st PC + 2nd PC :

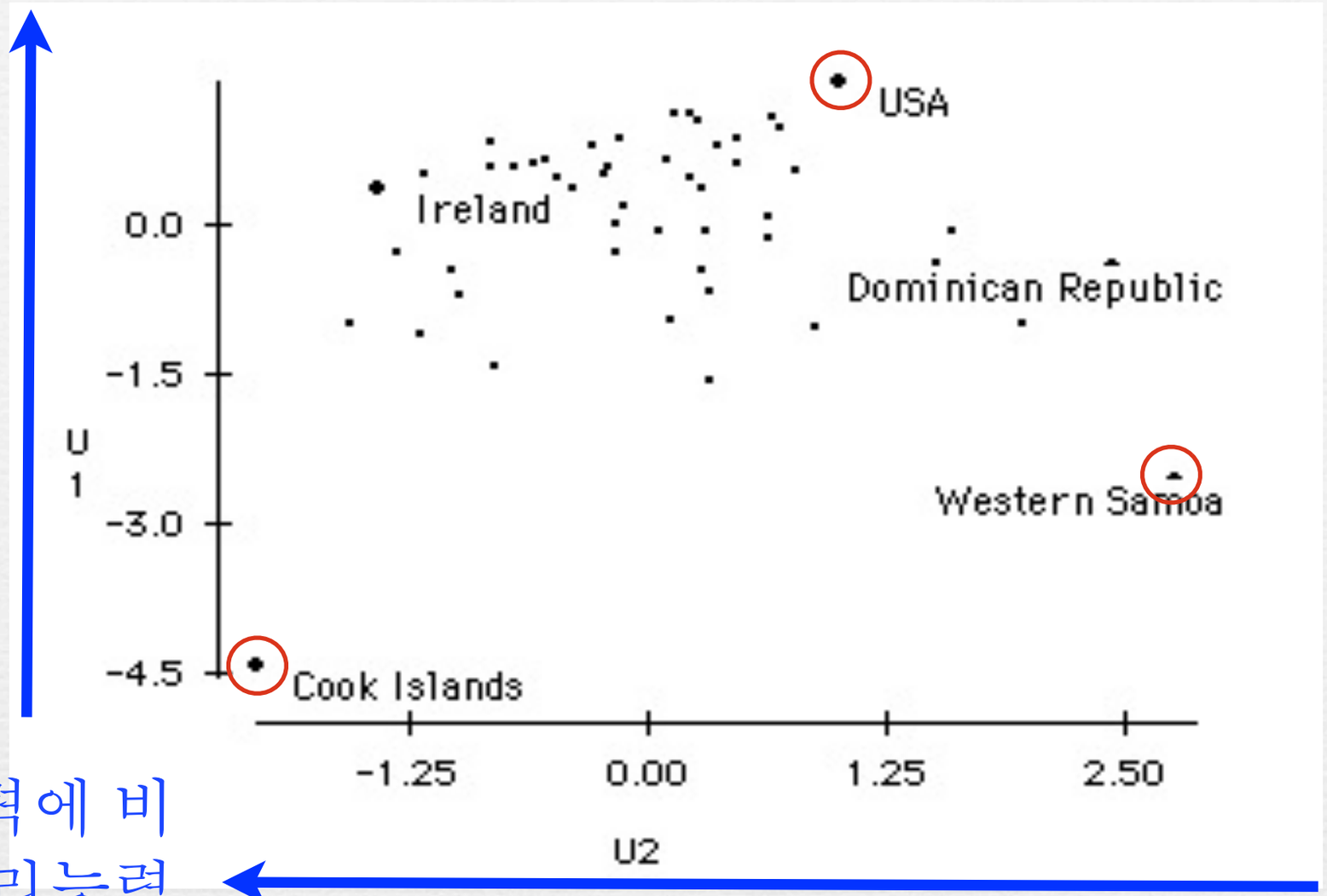
$$PC_1 = -0.468x1 - 0.514x4 - 0.517x15 - 0.499x100$$
$$PC_2 = -0.712x1 - 0.251x4 + 0.378x15 + 0.535x100$$

- PC1 : - (x1, x4, x15, x100 기록의 평균. Cov 사용시 x는 centered data)

  >> overall performance (클수록 좋은 기록)

- PC2 : 장거리-단거리 (값 클수록 장거리에 비해 단거리 우수)

Overall Performance



단거리능력에 비교한 장거리능력

- USA : 전체적 육상기록 우수
- Western Samoa, Cook Islands : 전체적 육상기록 저조
- Western Samoa : 장거리보다 단거리 우세 
- Cook Islands : 단거리보다 장거리 우세 

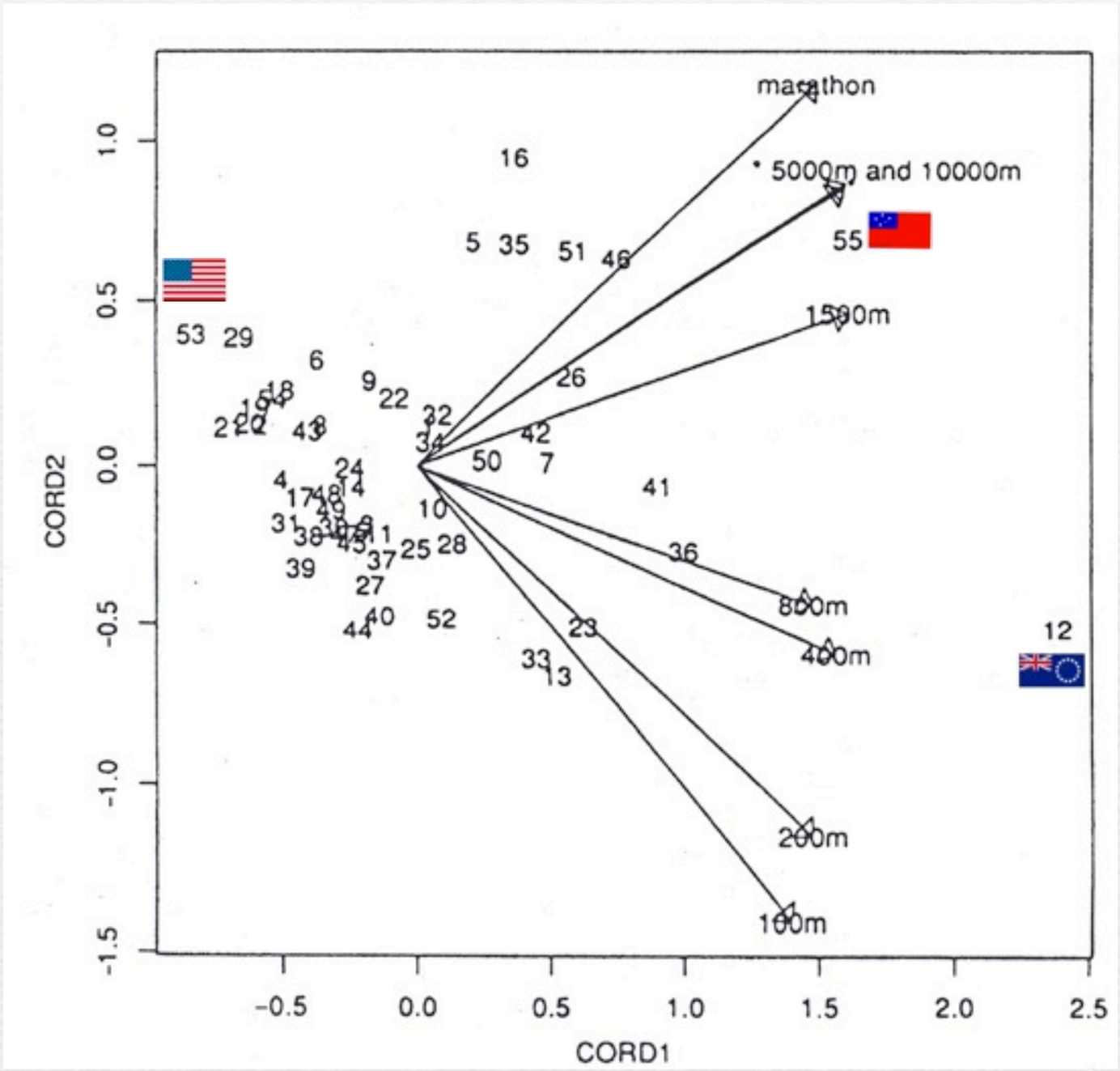
biplot(행렬도) -Gabriel, 1981-

$X_{n\times p}$ 에 대하여...

- n개 개체(subject)에 대한 상대적 위치 정보
- p개 변수(variables) 간 분산/공분산 정보

} biplot

- 행렬도의 수평축 = 제1주성분,
- 행렬도의 수직축 = 제2주성분
- 길이 = variance(x_j)
- 두 직선 간 각도가 작으면= 두 변수 간 상관계수 높다
- 어느 두 개체의 위치가 가까우면 = similarity



- 장거리 경기
- 단거리 경기
- 제1주성분 = 전체적인 기록
- 제2주성분 = 단거리 vs. 장거리
- 53 = USA
- 12 = Cook Islands
- 55= Western Samoa
- 비슷한 나라들 ...

In SAS,

%biplot.sas 라는 macro 프로그램 download

http://www.datavis.ca/books/sssg/biplot.html

%biplot(var=x1-x5, id=country)

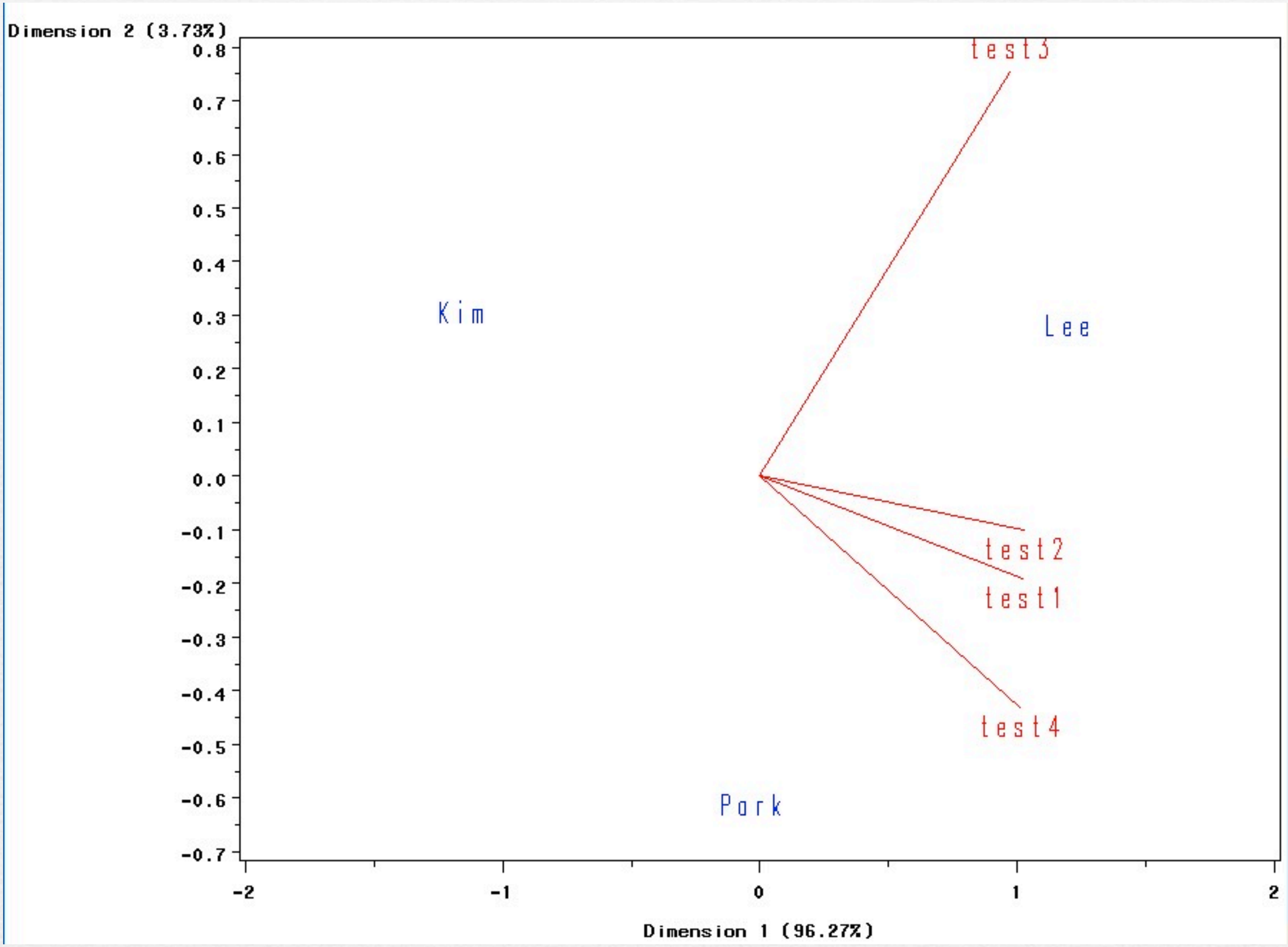
In R,

>athlete<-princomp(X)
>biplot(athlete, expand=0.7)

Editor - Untitled1 *

```
    delete _range_ _zero_;

    run; quit;



%done:
%if &abort %then %put ERROR: The BIPLOT macro ended abnormally.;

%mend BIPLOT;
```

```
data a; input name $  test1 test2 test3 test4; cards;
Kim 2.43 3.12 3.68 4.04
Park 3.41 3.91 4.07 5.10
Lee 4.21 4.65 5.87 5.69
;
%biplot(var=test1-test4, id=name);
```


Dimension 2 (3.73%)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-2
-1
0
1
2
Dimension 1 (96.27%)
Kim
Lee
Park
test3
test2
test1
test4