Principal Components Analysis

Edps/Soc 584, Psych 594

Carolyn J. Anderson

Department of Educational Psychology
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Spring 2017

Overview

- History and overview
- Population Principal Components
  - Geometry
  - Algebra
- Principal Components obtained from Standardized Variables
- Sample Principal Components
- Graphing Principal Components
- Distinctions between PCA and factor analysis

Reading: Johnson & Wichern pages 430–459 & 466–470; good supplemental references Jolliffe (1986), Krzanowski (1988); Flury (1988).

History

- First introduced by Karl Pearson (1901) in *Philosophical Magazine* as a procedure for finding lines and planes which best fit a set of points in p-dimensional space. The focus was on geometric optimization.
- Harold Hotelling (1933) published a paper on PCA in *Journal of Educational Psychology*, which dealt with an algebraic optimization.
    - He re-invented it but from a different perspective. His motivation was to find a smaller "fundamental set of independent variables" that determines the values of the original set of p variables.
    - This is a "factor analytic" type idea, but PCA is not factor analysis (except in a very special and unrealistic case).
    - Hotelling choose components (linear combinations of p variables) so as to maximize their successive contribution to the total variance.

History continued

Not much was done with respect to applications until the early 1960's— the advent of the computer age.

- There was an explosion of applications and developments of the technique.
- Theory for sampling distributions (which lead to statistical inference) was developed.
- Lots of Extensions of PCA (e.g., PCA for sets of matrices... for SAS/IML macros (by me) and MATLAB (by Mark de Rooij) code see faculty.education.illinois.edu/cja/homepage/software_index.html — algorithm is based on work by Kiers (1990).

Basic Idea

Reduce the dimensionality of a data set in which there is a large number of inter-related variables while retaining as much as possible the variation in the original set of variables.

The reduction is achieved by transforming the original variables to a new set of variables, "principal components, that are uncorrelated and ordered such that the first few retains most of the variation present in the data.

Goals & Objectives

- Reduction and summary $\longrightarrow$ data reduction.
- Study the structure of $\mathbf{\Sigma}$ (or $\mathbf{S}$ or $\mathbf{R}$) $\longrightarrow$ Interpretation.

Applications

- Interpretation (study structure)
- Create a new set of variables (a smaller number that are uncorrelated). These can be used in other procedures (e.g., multiple regression).
- Select a sub-set of the original variables to be used in other multivariate procedures.
- Detect outliers or clusters of observations.
- Check multivariate normality assumption (before assuming multivariate normality and analyzing data using procedures that assume multivariate normality.

Population Principal Components

- All your observations (measurements) on made on the members of the "population".
  - European countries in one study could be considered the population and you have data for each of them (the variables are percents of people employed in different industries).
  - The psychological test data consist of measurements on 64 subjects. These subjects are a sample from some populations. If we repeated the study, we'd most likely have different individuals.
- In Population principal components, we can compute $\mathbf{\Sigma}$ and the principal components (PCs) are derived from $\mathbf{\Sigma}$.

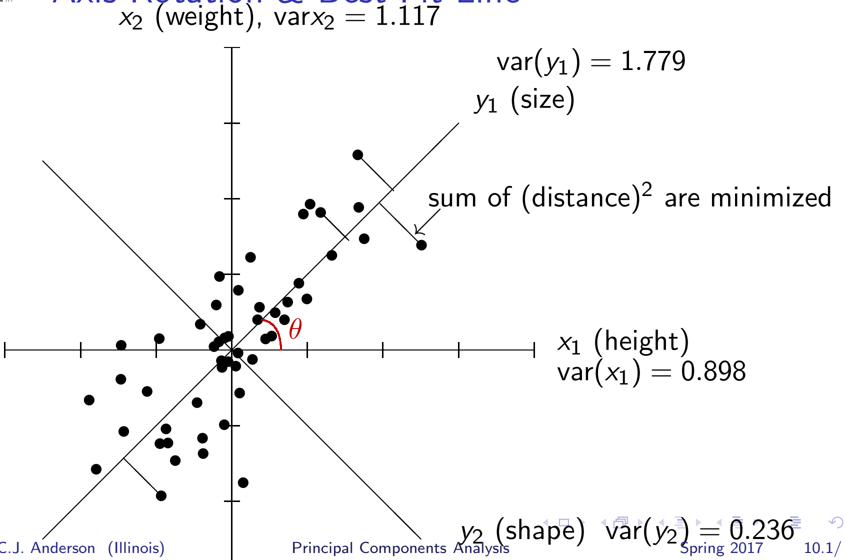
Two approaches

- Algebraically: PCs are linear combinations of p original variables $X_1, X_2, \ldots, X_p$ such that
  - The first PC has the largest variance as possible,
  - The second PC has the largest variance as possible and is orthogonal to the first
  - etc.
- Geometrically: (at least) 3 approaches
  - Rotation to a new coordinate system.
  - "Best" fit hyper-plane.
  - See appendix of the text for $n-$space interpretation

Geometry of PCA: p–space

- PCs represent a selection of a new coordinate system obtained by rotating the original axes to a set of new axes (to provide a simpler structure).
  - The first principal component represents the direction of maximum variability.
  - The second principal component represents the direction of maximum variability that is orthogonal to the first.
  - And so on, until the last PC which represents the direction of minimum variability & orthogonal to all of the others.
- "Best" fit is defined as minimizing the sum of squared distances between points that represent cases and space defined by principal components
  - The first principal component defines a line. The sum of squared distances (i.e., $\sum_{j=1}^{2} d_j^2$) between the points and this line are minimized.
  - The first Two principal components define a plane. The sum of squared distances between points and this plan are minimized.
  - etc.

Axis Rotation & Best Fit Line



Further Notes regarding PC

- They are "variance" preserving. For example,

$$\text{var}(x_1)+\text{var}(x_2)=0.898+1.117=2.015 = 1.779+0.236=\text{var}(y_1)+\text{var}(y_2)$$

- If you rotate PCs, you no longer have PCs.
- PCs only depend on $\mathbf{\Sigma}$ (or $\mathbf{R}$ if you're using standardized variables).
- PCs do not require any assumptions about distribution of the variables (e.g., multivariate normality).
- If variables do come from a multivariate normal populations, then
  - PCs can be interpreted in terms of constant density ellipsoids.
  - You can make inferences about the population from a sample.
- However, right now we're considering Population PC, so we don't have a sample and hence no inference is required.

The Algebra of Population PCA

We want to transform p variables to q orthogonal linear combinations (generally) where $q << p$.

$$\mathbf{X}'_{1\times p} = (X_1, X_2, \ldots, X_p) \quad \text{to} \quad \mathbf{Y}'_{1\times q} = (Y_1, Y_2, \ldots, Y_q)$$

There are p possible ones

$$\begin{aligned}
Y_1 &= \mathbf{a}'_1\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a}'_2\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
\vdots & \quad \vdots \\
Y_p &= \mathbf{a}'_p\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \\
\mathbf{Y} &= \mathbf{AX}
\end{aligned}$$

Given the covariance matrix $\mathbf{\Sigma}_X$ of the X's, we know

$$\text{var}(Y_i) = \mathbf{a}'_i\mathbf{\Sigma}_X\mathbf{a}_i \qquad \text{and} \qquad \text{cov}(Y_i, Y_k) = \mathbf{a}'_i\mathbf{\Sigma}_X\mathbf{a}_k$$

More Formal Definition of PCs

PCs are the uncorrelated linear combinations, $\text{cov}(Y_i, Y_k) = 0$ for all $i \neq k$, with variances as large as possible.

In particular,

$\text{var}(Y_1)$ is the maximum $\rightarrow$ find $\mathbf{a}_1 \supset \mathbf{a}_1'\mathbf{\Sigma}_X\mathbf{a}_1 = \max(\mathbf{a}'\mathbf{\Sigma}_x\mathbf{a})$

$\text{var}(Y_2)$ is the maximum and $\perp Y_1 \rightarrow$

find $\mathbf{a}_2 \supset \mathbf{a}_2'\mathbf{\Sigma}_X\mathbf{a}_2 = \max(\mathbf{a}'\mathbf{\Sigma}_x\mathbf{a})$ and $\mathbf{a}_1\mathbf{\Sigma}_X\mathbf{a}_2 = 0$

- At each step, select $\mathbf{a}_i$ such that $\mathbf{a}_i\mathbf{X}$ has maximum variance subject to being uncorrelated with all other linear combinations.
- Usually (but not always), we only use Y_1, $Y_2, \ldots,$ Y_q where q is much less than p (primary goal is data reduction).

More Formal Definition of PCs (continued)

The are p possible components, Y_1, Y_2,…, Y_p are needed to completely reproduce (represent) $\mathbf{\Sigma}_X$. So if $q < p$, we don't reproduce $\mathbf{\Sigma}_X$ exactly (unless the rank of $\mathbf{\Sigma}_X = q$).

Maximizing the Criteria

The criteria to be maximized is $\boxed{\max(\mathbf{a}'\boldsymbol{\Sigma}_X\mathbf{a})}$.

We can always multiply $Y_1 = \mathbf{a}'\mathbf{X}$ by a constant $|c| > 1$, which will increase the variance, $\text{var}cY_1 = \text{var}(c\mathbf{a}'\mathbf{X}) = c^2\text{var}(\mathbf{a}'\mathbf{X})$.

Therefore, we normalize the combination vector

$$\mathbf{a}'\mathbf{a} = 1 = L^2_{\mathbf{a}} = L_{\mathbf{a}}$$

Our problem is to find $\mathbf{a}_1$ that maximizes variance subject to a constraint

$$\max_{\mathbf{a}} \left(\frac{\mathbf{a}'\boldsymbol{\Sigma}_X\mathbf{a}}{\mathbf{a}'\mathbf{a}}\right) = \text{var}(Y_1)$$

Use results on maximization in "more linear algebra" notes

$$\max_{\mathbf{a}} \left(\frac{\mathbf{a}'\boldsymbol{\Sigma}_X\mathbf{a}}{\mathbf{a}'\mathbf{a}}\right) = \lambda_1$$

Proof that this is Maximum

Showing is better than just believing...

$$\begin{aligned}
\frac{\mathbf{a}'\boldsymbol{\Sigma}_X\mathbf{a}}{\mathbf{a}'\mathbf{a}} &= \text{var}(Y_1) \\
\mathbf{a}'\boldsymbol{\Sigma}_X\mathbf{a} &= \text{var}(Y_1)\mathbf{a}'\mathbf{a} \\
\mathbf{a}'\boldsymbol{\Sigma}_X\mathbf{a} - \text{var}(Y_1)\mathbf{a}'\mathbf{a} &= 0 \\
\mathbf{a}'(\boldsymbol{\Sigma}_X\mathbf{a} - \text{var}(Y_1)\mathbf{a}) &= 0 \qquad\qquad (\text{since } \mathbf{a} \neq 0) \\
\boldsymbol{\Sigma}_X\mathbf{a} - \text{var}(Y_1)\mathbf{a} &= 0 \\
\underbrace{\boldsymbol{\Sigma}_X}_{p\times p}\underbrace{\mathbf{a}}_{p\times 1} &= \underbrace{\text{var}(Y_1)}_{scalar}\underbrace{\mathbf{a}}_{p\times 1}
\end{aligned}$$

which is just the equation what eigenvalues and eigenvectors solve.

So

$\boxed{Y_1 = \mathbf{e}_1'\mathbf{X} \text{ where } \mathbf{e}_1 \text{ is the } 1^{st} \text{ eigenvector of } \boldsymbol{\Sigma}_X \text{ and } \text{var}(Y_1) = \lambda_1.}$

Population PC: Result 1

Let $\mathbf{\Sigma}$ be the covariance matrix associated with the vector $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$. Let $\mathbf{\Sigma}$ have the eigenvector-eigenvalues pairs $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2)$,..., $(\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the i^{th} PC is given by

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p$$

for $i = 1, 2, \ldots, p$. Given this

$$\begin{aligned} \text{var}(Y_i) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i &= \mathbf{e}_i'(\lambda_i\mathbf{e}_i) \\ &= \lambda_i\mathbf{e}_i'\mathbf{e}_i = \lambda_i \end{aligned}$$

and for $i \neq k$ $\text{cov}(Y_i, Y_k) = \mathbf{e}_i\mathbf{\Sigma}\mathbf{e}_k = \mathbf{e}_i'(\lambda_k\mathbf{e}_k)$

If some of the λ_1 are equal, then the choice of the corresponding coefficient vectors $\mathbf{e}_i$ (and thus Y_i) are not unique.

Population PC continued

We can write all of this in terms of matrices:

$$\mathbf{Y} = \mathbf{P}'\mathbf{X} \Longrightarrow \text{cov}(\mathbf{Y}) = \mathbf{\Sigma}_Y = \mathbf{P}'\mathbf{\Sigma}_X\mathbf{P}$$

So,

$$\underbrace{\mathbf{\Sigma}_X}_{\text{cov}(\mathbf{X})} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}' \Longleftrightarrow \underbrace{\mathbf{\Sigma}_Y}_{\text{cov}(\mathbf{Y})} = \mathbf{\Lambda} = \mathbf{P}'\mathbf{\Sigma}_X\mathbf{P} = \text{diag}(\lambda_i)$$

More Population PC Results

Let $\mathbf{X}' = (X_1, X_2, \ldots, X_p)$ have covariance matrix $\mathbf{\Sigma}_X$ with eigenvalue and eigenvector pairs pairs $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2)$,..., $(\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = \mathbf{e}_1'\mathbf{X}$, $Y_2 = \mathbf{e}_2'\mathbf{X}$,... $Y_p = \mathbf{e}_p'\mathbf{X}$ be the PCs. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \sigma_{ii} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \lambda_i$$

The Total Population Variance is preserved by the transformation.

The Proportion of total variance due to the k^{th} PC is

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \qquad k = 1, \ldots, p$$

Proportion of Variance Accounted For

We often select q PCs such that the proportions for $k = 1, \ldots, q$ sum up as close to 1 (yet not too large of a value for q).

The Proportion of Variance accounted for by the first q PCs equals

$$\frac{\sum_{k=1}^{q} \lambda_k}{\text{trace}(\mathbf{\Sigma}_X)}$$

We try to balance the percent of variance (information) retained and the number of PCs (simplicity). We may want to replace $\mathbf{X}$ by $\mathbf{Y}$.

Often we're interested interpreting the new variables (i.e., the PCs), so we examine the elements of the $\mathbf{e}_i$'s

The size (magnitude) of the elements of $\mathbf{e}_i$ are an indicator of a variables "importance" to the i^{th} PC...

Correlation between Y_i and X_k

If $Y_1 = \mathbf{e}_i'\mathbf{X}$, $Y_2 = \mathbf{e}_2\mathbf{X}, \ldots Y_p = \mathbf{e}_p\mathbf{X}$ are the PCs obtained from $\mathbf{\Sigma}_X$, we can use ρ_{Y_i,X_k} to help interpret the contribution of an X_k to Y_i.

$$
\begin{aligned}
\rho_{Y_i,X_k} &= \frac{\text{cov}(Y_i, X_k)}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} \\
&= \frac{\text{cov}(\mathbf{e}_i'\mathbf{X}, \ell'\mathbf{X})}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} \text{ where } \ell'_{1\times p} = (0,\ldots, \underbrace{1}_{k^{th}\,position}, 0, \ldots, 0) \\
&= \frac{\ell'\mathbf{\Sigma}\mathbf{e}_i}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} \\
&= \frac{\ell'(\lambda_i\mathbf{e}_i)}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} \\
&= \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}
\end{aligned}
$$

Example: European Cars

The data are percentages of people employed in different industries in European countries during 1979 (cold war era). Data from *Euromonitor* (1979) "European Marketing Data and Statistics," London: Euromonitor Publications. . . I go it off of the web from http://www.cmu.edu/DASL.

- $N = 26$ countries
- There are 9 industries, but we'll start with just $p = 3$:
  - X_1 = percent in manufacturing.
  - X_2 = percent in services industry.
  - X_3 = percent in social and personal services.

$$\boldsymbol{\mu} = \begin{pmatrix} 27.008 \\ 12.958 \\ 20.023 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 49.109 & 6.535 & 7.379 \\ 6.535 & 20.933 & 17.879 \\ 7.379 & 17.879 & 46.643 \end{pmatrix}$$

$$\text{total variance} = \text{trace}(\boldsymbol{\Sigma}) = 116.68432$$

Example: Eigenvalues and Eigenvectors

i	var(Y_i) λ_i	Cumulative variance	Percent	Cumulative Percent
1	62.62	62.62	53.66	53.66
2	42.47	105.09	36.39	90.06
3	11.60	116.68	9.94	100.00

Eigenvectors, which give weights for principal components:

$$\begin{aligned} \mathbf{e}_1' &= (0.580, 0.396, 0.712) \\ \mathbf{e}_2' &= (0.811, -0.207, -0.546) \\ \mathbf{e}_3' &= (-0.069, 0.894, -0.442) \end{aligned}$$

So the Principal component are

$$\begin{aligned} Y_1 &= 0.580X_1 + 0.396X_2 + 0.712X_3 \\ Y_2 &= 0.811X_1 - 0.207X_2 - 0.546X_3 \\ Y_3 &= -0.069X_2 + 0.894X_2 - 0.442X_3 \end{aligned}$$

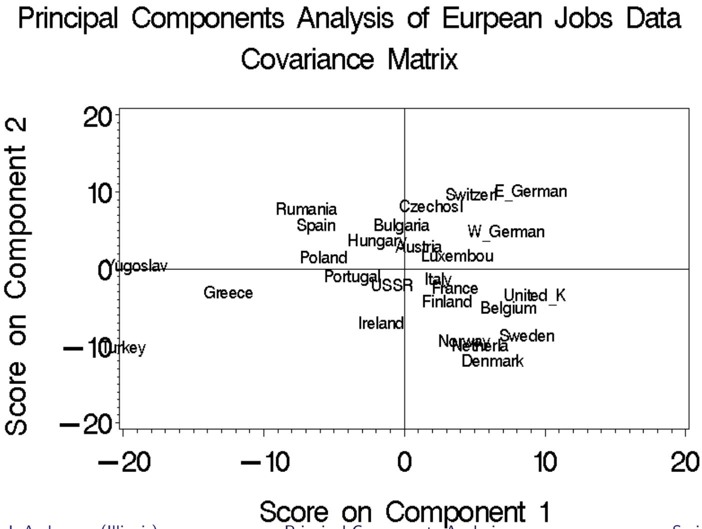
Example: Interpretation of Components

We'll look at correlations between Y_1 and Y_2 and each of the X_k's:

Original Variables		Principal Components Y_1	Y_2
Manufacturing	X_1	$\frac{\sqrt{62.62}}{\sqrt{49.109}}(0.580) = .66$	$\frac{\sqrt{42.47}}{\sqrt{49.109}}(0.811) = .75$
Service	X_2	$\frac{\sqrt{62.62}}{\sqrt{20.933}}(0.396) = .69$	$\frac{\sqrt{42.47}}{\sqrt{20.933}}(-0.207) = -.30$
Social & Personal	X_3	$\frac{\sqrt{62.62}}{\sqrt{46.643}}(0.712) = .82$	$\frac{\sqrt{42.47}}{\sqrt{46.643}}(-0.546) = -.52$

- Y_1: All variables are contributing to the first component; it's an "overall" percent employment in all industries.
- Y_2: This contrasts Manufacturing with Service and Social & Personal.

Plot of Component Scores



If Population is Multivariate Normal

We have an additional interpretation if $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Recall that the probability density contours (ellipsoids) are

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

The center is at $\boldsymbol{\mu}$ and the axes are at $\boldsymbol{\mu} \pm c\sqrt{\lambda_i}\,\mathbf{e}_i$, where λ_i and $\mathbf{e}_i$ are the i^{th} eigenvalue and vector of $\boldsymbol{\Sigma}$.

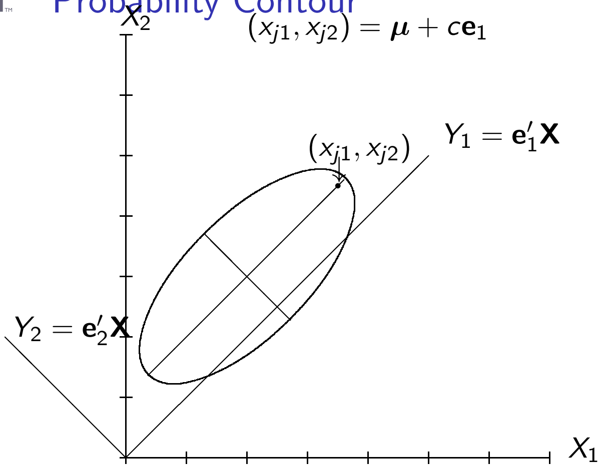
The principal components are

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} \\ Y_2 &= \mathbf{e}_2'\mathbf{X} \\ \vdots & \quad \vdots \\ Y_p &= \mathbf{e}_p'\mathbf{X} \end{aligned}$$

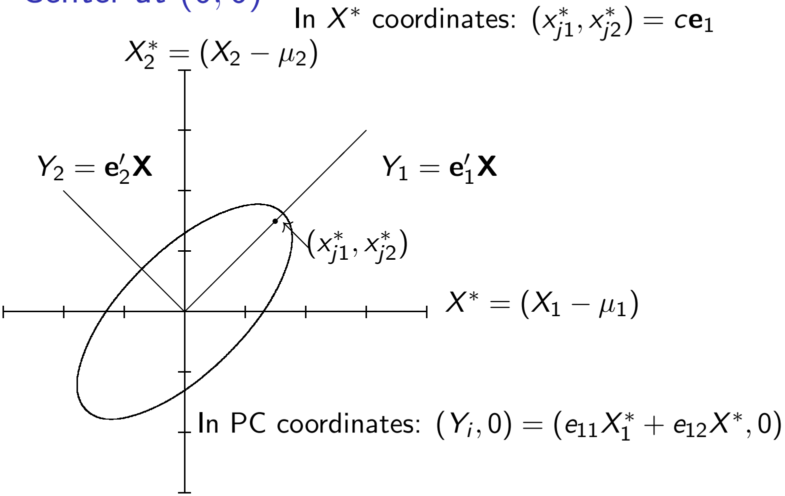
If Population is Multivariate Normal

The Principal components lie in the same directions as the axes of the probability contours (ellipsoids)

Probability Contour



Center at $(0, 0)$


In X^* coordinates: $(x^*_{j1}, x^*_{j2}) = c\mathbf{e}_1$
$X^*_2 = (X_2 - \mu_2)$
$Y_2 = \mathbf{e}_2'\mathbf{X}$
$Y_1 = \mathbf{e}_1'\mathbf{X}$
(x^*_{j1}, x^*_{j2})
$X^* = (X_1 - \mu_1)$
In PC coordinates: $(Y_i, 0) = (e_{11}X^*_1 + e_{12}X^*, 0)$

Summary example when $X \sim \mathcal{N}_p(\mu, \Sigma)$

Any point on the i^{th} axis of the ellipsoid has

- $\mathbf{X}$ coordinates $= \boldsymbol{\mu} + c\mathbf{e}_i$.
- $\mathbf{X}$ coordinates that are proportional to $\mathbf{e}_i' = (e_{i1}, e_{i2}, \ldots, e_{ip})$ in the coordinate system that has origin at $\boldsymbol{\mu}$ and axes parallel to the original $\mathbf{X}$ axes (i.e., the X^* coordinates).
- Subtracting mean doesn't change anything except move the origin to $(0, 0)$.
- In the coordinate system of the PC's the point has principal component $(Y_i, 0)$, because PC's are obtained by a rigid rotation of the original coordinate axes through an angle θ until they coincide with the axes of the ellipsoid.
- All of these results generalize to $p > 2$.

When Variances are Very Different

Principal Components obtained from Standardized Variables

If we use standardized variables ("z-scores)

$$
\begin{aligned}
Z_1 &= \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} \\
Z_2 &= \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} \\
\vdots &\qquad \vdots \\
Z_p &= \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}
\end{aligned}
$$

or in matrix notation

$$
\mathbf{Z} = \underbrace{\mathbf{V}^{-1/2}}_{\text{diag}(1/\sqrt{\sigma_{ii}})} \underbrace{(\mathbf{X} - \boldsymbol{\mu})}_{p\times 1} = \mathbf{V}^{-1/2}\mathbf{X} - \mathbf{V}^{-1/2}\boldsymbol{\mu}
$$

So $\mathbf{Z}$ is a linear combination of $\mathbf{X}$, which means. . .

PCs of Standardized Variables

We know that

$$E(\mathbf{Z}) = E(\mathbf{V}^{-1/2}(\mathbf{X}-\boldsymbol{\mu})) = \mathbf{V}^{-1/2}\underbrace{E(\mathbf{X})}_{\boldsymbol{\mu}} - \mathbf{V}^{-1/2}\boldsymbol{\mu} = \mathbf{0}$$

and

$$\boldsymbol{\Sigma}_Z = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}_X\mathbf{V}^{-1/2} = \mathcal{R}$$

which is the (population) correlation matrix of the X's.

The i^{th} PC of the standardized variables $\mathbf{Z}' = (Z_1, Z_2, \ldots, Z_p)$ with $\boldsymbol{\Sigma}_Z = \mathcal{R}$ is given by

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{e}}_i'\mathbf{Z} = \tilde{\mathbf{e}}_i'(\mathbf{V}^{-1/2}(\mathbf{X}-\boldsymbol{\mu})) \quad \text{for} \quad i = 1, 2, \ldots, p$$

where $\tilde{\mathbf{e}}_i$ is the i^{th} eigenvector and $\tilde{\lambda}_i$ is the i^{th} eigenvalue of $\mathcal{R}$.

Note that

$$\sum_{i=1}^{p}\text{var}(Z_i) = \sum_{i=1}^{p}\tilde{\lambda}_i = \sum_{i=1}^{p}\text{var}(\tilde{Y}_i) = \text{trace}(\mathcal{R}) = p$$

PCs of Standardized versus non-Std Variables

Almost always

$$\lambda_i \neq \tilde{\lambda}_i \qquad \text{and} \qquad \mathbf{e}_i \neq \tilde{\mathbf{e}}_i$$

That is

The PCs from $\boldsymbol{\Sigma}_X$ are not the same as PCs from $\mathcal{R}$

We'll look at a situation where standardization makes a difference

This will be the case when the scales of the X variables are (substantially or vastly) different and they are ont comparable.

Men's Track Data

From Johnson & Wichern: The data are from the Track and Field Statistics Handbook for the 1984 Los Angeles Olympics. These data are the national record times for men before the 1984 Olympics.
The record times for eight races (i.e., $p = 8$) are listed for 55 countries (i.e., $n = 55$).
The times are recorded for the following races:

- 100m: Record time for 100m race in seconds
- 200m: Record time for 200m race in seconds
- 400m: Record time for 400m race in seconds
- 800m: Record time for 800m race in minutes
- 1500m: Record time for 1500m race in minutes
- 5K: Record time for 5000m race in minutes
- 10K: Record time for 10000m race in minutes
- Marathon: Record time for the Marathon (approx. 26 miles) in minutes

Summary Statistics

Summary Statistics for each variable are given below:

	100m	200m	400m	800m	1500m	5K	10K	Marathon
$\bar{x}$	10.47	20.90	46.44	1.79	3.70	13.85	28.99	136.62
s	0.35	0.64	1.46	0.06	0.16	0.80	1.81	9.23

Covariance Matrix (truncated values)

	m100	m200	m400	m800	m1500	K5	K10	Mara.
m100	0.12	0.20	0.43	0.01	0.03	0.17	0.40	1.68
m200	0.20	0.41	0.79	0.03	0.07	0.35	0.81	3.54
m400	0.43	0.79	2.12	0.08	0.18	0.90	2.07	9.47
m800	0.01	0.03	0.08	0.004	0.00	0.04	0.10	0.47
m1500	0.03	0.07	0.18	0.01	0.02	0.11	0.26	1.24
K5	0.17	0.35	0.90	0.04	0.11	0.64	1.41	6.89
K10	0.40	0.81	2.07	0.10	0.26	1.41	3.26	15.732
Marathon	1.68	3.54	9.47	0.47	1.24	6.89	15.73	85.13

Eigenvalues of Σ

From the SAS/PRINCOMP Procedure:

Total Variance $=$ 91.738234815

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	89.914	88.500	0.980	0.9801
2	1.412	1.153	0.015	0.9955
3	0.260	0.150	0.003	0.9983
4	0.109	0.082	0.001	0.9995
5	0.027	0.015	0.000	0.9998
6	0.013	0.010	0.000	1.0000
7	0.002	0.002	0.000	1.0000
8	0.000		0.000	1.0000

Eigenvectors of Σ

	Principal Components		
Race	Prin1	Prin2	Prin3
m100	0.02	0.21	−.03
m200	0.04	0.36	−.02
m400	0.11	0.83	−.38
m800	0.01	0.02	0.01
m1500	0.01	0.04	0.05
K5	0.08	0.13	0.34
K10	0.18	0.30	0.85
Marathon	0.97	−.18	−.14

The 1st principal component is essentially the marathon, because it has by far the largest variance 85.13 compared to the next largest which is 3.26 (the 10K).

The variance on the 1st component is 89.914. . .

The Correlation Matrix

Values are Truncated

	m100	m200	m400	m800	m1500	K5	K10	Mara.
m100	1.00	.92	.84	.75	.70	.61	.63	.51
m200	.92	1.00	.85	.80	.77	.69	.69	.59
m400	.84	.85	1.00	.87	.83	.77	.78	.70
m800	.75	.80	.87	1.00	.91	.86	.86	.80
m1500	.70	.77	.83	.91	1.00	.92	.93	.86
K5	.61	.69	.77	.86	.92	1.00	.97	.93
K10	.63	.69	.78	.86	.93	.97	1.00	.94
Marathon	.51	.59	.70	.80	.86	.93	.94	1.00

The Eigenvalues of the Correlation Matrix

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	6.62	5.745	0.828	0.828
2	0.87	0.718	0.110	0.938
3	0.15	0.035	0.020	0.957
4	0.12	0.044	0.025	0.973
5	0.08	0.012	0.010	0.983
6	0.06	0.022	0.018	0.991
7	0.04	0.024	0.015	0.997
8	0.02		0.002	1.000

Total variance = 8.

The first 2 principal components account for 93.8% of the total variance.

The Eigenvectors of the Correlation Matrix

The First Two Eigenvectors

Race	Component "Loadings" 1	2
100m	.318	.567
200m	.337	.462
400m	.356	.248
800m	.369	.012
1500m	.373	−.140
5K	.364	−.312
10k	.367	−.307
Marathon	.342	−.439

The Eigenvectors of the Correlation Matrix

Interpretation?

- First component: An overall measure — High values on this component indicate slower runners.
- Second component: Contrast long and short races —
  - Small values indicate faster on short races than long ones.
  - Large values indicate slower on short races than long ones.
  - Value near zero means that tend to be similar on short and long races (could be slow, fast, or somewhere in between on all races).

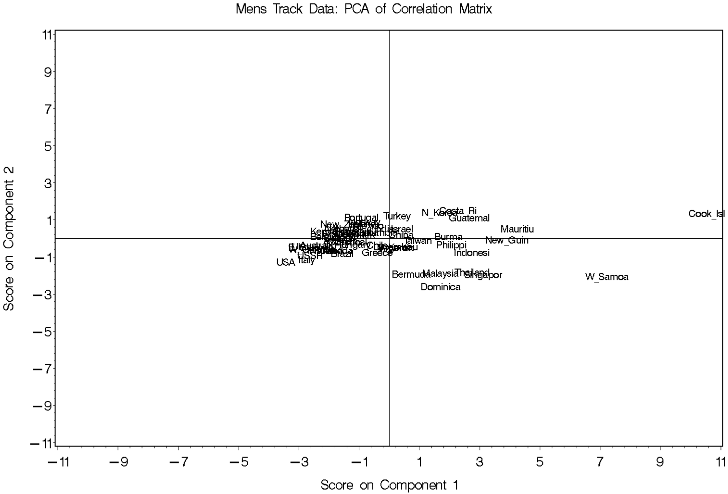
Correlations(Variables, Components)

The correlations between the standardized variables and values on the principal components equal

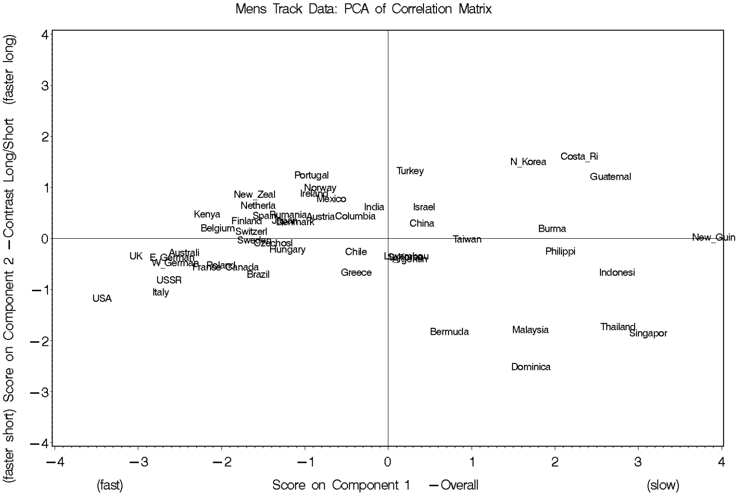
$$r_{Z_k,Y_i} = \sqrt{\lambda_i} e_{ki} \qquad \text{(e.g., } \sqrt{6.622}(.318) = .82)$$

	Components	
Race	1	2
100m	.82	.53
200m	.87	.43
400m	.92	.23
800m	.95	.01
1500m	.96	−.13
5K	.94	−.29
10k	.94	−.29
Marathon	.88	−.41

Graph of Countries Component Scores



Graph of Countries Component Scores



Sample Principal Components

Used to summarize the sample variation by PCs.

The Algebra is the same as in population principal components.

- $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are n independent observations from a population with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.
- $\bar{\mathbf{x}}_{p\times 1}$ = sample mean vector.
- $\mathbf{S}_{p\times p} = \{s_{ik}\}$ = sample covariance matrix.
- $\mathbf{S}$ has eigenvalue/vector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$.
- The $\hat{}$ indicates these are estimates of population values.
- The i^{th} sample principal component is given by

$$\hat{\mathbf{y}}_i = \hat{\mathbf{e}}_i \mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p$$

- The i^{th} PC sample variance $= \text{var}(\hat{y}_i) = \hat{\lambda}_i$ for $i = 1, \ldots, p$.
- The PC sample covariances $= \text{cov}(\hat{y}_i, \hat{y}_k) = 0$ for all $i \neq k$.

Algebra of Sample PC continued

- Total sample variance

$$\text{trace}(\mathbf{S}) = \text{tr}(\mathbf{S}) = \sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i$$

- Proportion of total sample variance accounted for by the i^{th} PC

$$\frac{\hat{\lambda}_i}{\sum_{k=1}^{p} \hat{\lambda}_k}$$

- Correlations between $\hat{y}_i$ and x_k

$$r_{\hat{y}_i, x_k} = \frac{\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}} \hat{e}_{ik}$$

Note if you use standardized x's, then $r_{\hat{y}_i, z_k} = \sqrt{\hat{\hat{\lambda}}_i} \hat{\hat{e}}_{ik}$

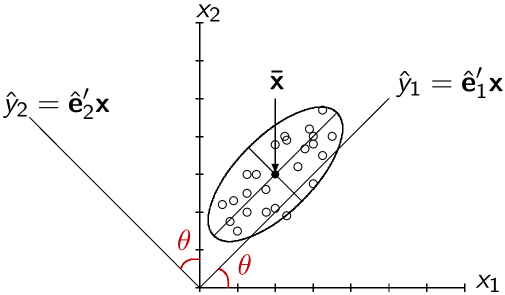
Algebra of Sample PC continued

- The sample PCs based on **S** are not the same as those based on **R**. (I'll use $\tilde{}$ to denote those based on **R**).
- Use **S** when observations are not in the same unit or when the variances s_{ii} are not vastly different.

Geometry of Sample PC

- PCs based on a sample of n p-dimensional observations are new variables specified by a rigid rotation of the original axes to a new orientation such that the directions of the axes in the new orientation have maximum variances in the sample.
  - The rotation must be rigid since the new variables must be $\perp$.
  - Directions of the new axes are based on $\mathbf{S}$ (or $\mathbf{R}$)

Geometry of Sample PC



Geometry of Sample continued

The PCs are projections of observations onto the principal axes of the ellipsoids.

We can re-center the x's, which also centers the $\hat{y}$'s; that is

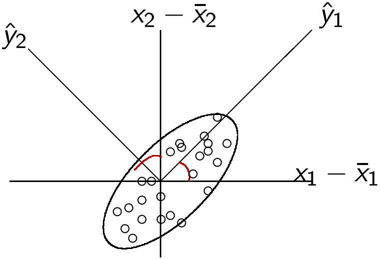
$$(\mathbf{x}_i - \bar{\mathbf{x}}) = 0 \longrightarrow \hat{y}_i \quad \text{has mean } 0$$

Subtraction of $\bar{\mathbf{x}}$ only effects the mean and does not effect variances and covariance.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \underset{\text{shift location}}{\longrightarrow} \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{pmatrix} \underset{\text{rigid rotatation}}{\longrightarrow} \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix}$$

Geometry of Sample continued

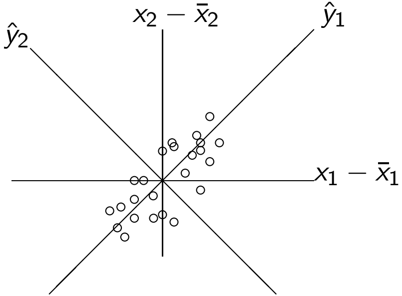
The PCs are projections of observations onto the principal axes of the ellipsoids.



2nd Geometric Interpretation

The 1st PC $\hat{y}_1$ minimizes the sum of squared deviations (distances) of the points to a line (least squared best fit).

When you approximate p-dimensional data by $r << p$ PCs, the t PCs minimize the sum of squared distances of points in p-space onto the r dimensional sub-space.



Swiss Bank Notes

These data are from Flurry & Riedwyl (1988) *Multivariate Statistics: A practical approach*.

The data consist of $p = 6$ measurements in millimeters on $n = 100$ genuine Swiss Bank notes (old ones)... picture on next slide

- x_1: Length of the bank note,
- x_2: Height of the bank note, measured on the left,
- x_3: Height of the bank note, measured on the right,
- x_4: Distance of inner frame to the lower border,
- x_5: Distance of inner frame to the upper border,
- x_6: Length of the diagonal.

Picture of Bank Note

Swiss Bank Notes: sample statistics

Sample Means:

$$\bar{\mathbf{x}}' = (214.969, 129.943, 129.720, 8.305, 10.168, 141.517)$$

The sample covariance matrix **S**:

	Length X_1	Left X_2	Right X_3	Bottom X_4	Top X_5	Diagonal X_6
X_1	.1502	.0580	.0573	.0571	.0145	.0055
X_2	.0580	.1326	.0859	.0567	.0491	−.0431
X_3	.0573	.0959	.1263	.0582	.0306	−.0238
X_4	.0571	.0567	.0582	.4132	−.2635	−.0002
X_5	.0145	.0491	.0306	−.2635	.4212	−.0753
X_6	.0055	−.0431	−.0238	−.0002	−.0753	.1998

Eigenvalues of S

The variances of the principal components (i.e., the eigenvalues of $\mathbf{S}$):

PC	$\hat{\lambda}_i$	Proportion of of variance	Cummulative Proportion
1	.6891	.4774	.4774
2	.3598	.2490	.7264
3	.1856	.1286	.8550
4	.0872	.0604	.9154
5	.0802	.0555	.9709
6	.0420	.0291	1.000

Eigenvectors of Genuine Bank notes

The principal components (eigenvectors of $\mathbf{S}$):

		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
Length	X_1	.0613	.3784	.4715	−.7863	.1114	.0114
Left	X_2	.0127	.5066	.1013	.2441	−.3584	−.7381
Right	X_3	.0374	.4543	.1963	.2807	−.4812	.6659
Bottom	X_4	.6970	.3577	−.1075	.2421	.5599	.0510
Top	X_5	−.7055	.3648	.0738	.2434	.5483	.0626
Diagonal	X_6	.1060	−.3643	.8438	.3543	.1161	−.0716

Correlation between measures and PCs

The correlations between the original variables and the principal components (i.e., $r_{x_k,y_i} = \hat{e}_{ki}\sqrt{\hat{\lambda}_i}/\sqrt{s_{kk}}$):

		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
Length	X_1	.1313	.5853	.5241	−.5990	.0814	−0.006
Left	X_2	.0290	.8340	.1198	.19793	−.2787	.4152
Right	X_3	.0874	.7662	.2379	.2332	−.3837	−.3838
Bottom	X_4	.9000	.3336	−.0721	.1112	.2466	−.0163
Top	X_5	−.9023	.3369	.0490	.1108	.2392	−.0198
Diagonal	X_6	.1969	−.4885	.8132	.2341	.0735	.0328

- Y_1 is a contrast between Bottom & Top.
- Y_2 is overall size, except for Diagonal.
- Y_3 & Y_4 — nothing obvious.
- Y_5 is something like "image".
- Y_6 measurement error or "slant of cut."

The Latter PCs

We've focused on the first PCs, but the last ones can also be informative.

Small values for the smallest eigenvalues from either **S** or **R** indicate:

- Undetected linear dependencies in the data.
- One (or more) of the variables is redundant with others and could be deleted.
- Such PCs can be substantively just an important as PCs associated with the largest eigenvalues.
- The latter ones could be due to pure error variability (measurement error).

The Latter PCs

Swiss Bank Note example: The last PC is basically $X_2 - X_3 =$ (Right)$-$(Left). Typically, $X_2 - X_3 > 0$. So this last PC could

- Reflect the "slant" of the cut.
- If X_2 and X_3 are measuring the same thing (quantity), the only reason that $\hat{\lambda}_6 > 0$ is due to measurement error (error variability).

Sampling Theory

Asymptotic & complex

If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is a sample from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the sample principal components

$$\hat{Y}_i = \hat{\mathbf{e}}_i'(\mathbf{X} - \bar{\mathbf{X}})$$

are observations ("realizations") of the population principal components

$$Y_i = \mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu})$$

and since $\hat{y}_i$ is a linear combination of $\mathbf{x}$ which come from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\hat{\mathbf{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_p \end{pmatrix} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda})$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$.

Sampling Theory continued

Assume $\mathbf{X}_j \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ i.i.d. for $j = 1, 2, \ldots n$.

$\boldsymbol{\Sigma}$, which is unknown, has eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ (assumption) with associate eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$.

For n "very large" n

- $\hat{\lambda}_i$ is independent of its corresponding $\hat{e}_i$.
- $\sqrt{n}(\hat{\lambda} - \lambda_i) \approx \mathcal{N}_p(\mathbf{0}, 2\boldsymbol{\Lambda}^2)$ or that

$$\hat{\lambda} \approx \mathcal{N}_p(\lambda, \frac{2}{n}\boldsymbol{\Lambda}^2)$$

where $\hat{\lambda}$ are eigenvalues of $\mathbf{S}$, and λ are eigenvalues of $\boldsymbol{\Sigma}$. So

$$\hat{\lambda}_i \approx \mathcal{N}_1(\lambda_i, \frac{2}{n}\lambda_i^2) \quad \text{for} \quad i = 1, \ldots, p$$

Sampling Theory continued

And

$\sqrt{n}(\hat{\mathbf{e}}_i - \mathbf{e}_i) \approx \mathcal{N}_p(\mathbf{0}, \mathbf{E}_i)$ where

$$\mathbf{E}_i = \lambda_i \sum_{k \neq i} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{e}_k \mathbf{e}_k'$$

Note: $\mathbf{E}_i$ is not diagonal, and the Eigenvectors are not independent.

Using Distribution of $\hat{\lambda}$'s

Since the $\hat{\lambda}_i$'s are asymptomatically (very large n) independent and normal with mean λ_i and variance $(2/n)\lambda_i^2$, a $(1-\alpha)100\%$ confidence interval for λ_i is

$$\frac{\hat{\lambda}_i}{(1 + z_{\alpha/2}\sqrt{2/n})} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{(1 - z_{\alpha/2}\sqrt{2/n})}$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)^{th}$ percentile of $\mathcal{N}(0,1)$.

or If we can do a Bonferroni-type procedure and use $z_{\alpha/(2m)}$ where $m =$ number of intervals you plan to constructs.

Using Distribution of $\hat{\lambda}$'s

Swiss Bank Note Example: The 95% confidence interval for λ_1 is :

$$\frac{.6891}{1 + 1.96\sqrt{\frac{2}{100}}} \leq \lambda_1 \leq \frac{.6891}{1 - 1.96\sqrt{\frac{2}{100}}} \longrightarrow (.5395, .9534)$$

and the rest are on the next slide...

Swiss Bank note: CI's for λ's

PC	$\hat{\lambda}_i$	Proportion of of variance	Cumulative Proportion	95% Confidence Intervals Lower	Upper
1	.6891	.4774	.4774	.5395	.9534
2	.3598	.2490	.7264	.2817	.4978
3	.1856	.1286	.8550	.1453	.2568
4	.0872	.0604	.9154	.0683	.1206
5	.0802	.0555	.9709	.0628	.1110
6	.0420	.0291	1.0000	.0329	.0581

Using the Distribution of $\hat{e}_i$

- The $\hat{\mathbf{e}}_i$'s are approximately normal with mean $\mathbf{e}_i$.
- The elements of each $\hat{\mathbf{e}}_i$ are correlated and these correlations depend on the ratios
$$\frac{\lambda_k}{(\lambda_k - \lambda_i)^2}$$
That is, how far λ_k is from λ_i.
- It can be useful to look at the diagonal elements of $\sqrt{(1/n)}\hat{\mathbf{E}}_i$. These are the standard errors of $\hat{e}_{ki}$'s.
- Recall
$$\hat{\mathbf{E}}_i = \hat{\lambda}_i \sum_{k \neq i} \frac{\hat{\lambda}_i}{(\hat{\lambda}_k - \hat{\lambda}_i)^2} \hat{\mathbf{e}}_k \hat{\mathbf{e}}_k'$$
- Notes:
  - The variances of $\hat{\lambda}$ increases as λ increases, so large λ's can have very wide confidence intervals.
  - These sampling results do not apply to **R**—they only apply to **S**.

Testing $H_o : \lambda_i = \lambda$ for $i = (r+1), \ldots, p)$

Bartlett (1947) developed a test for the hypothesis that $(p - r)$ smaller eigenvalues of $\mathbf{\Sigma}$ are equal for $0 < r < p - 1$.

If data support this hypothesis, then there probably will be little interest in using more than r components.

Bartlett's approximate χ^2 statistics has the following form

$$M\left[-\ln(\det(\mathbf{S})) + \sum_{i=1}^{r} \ln(\lambda_i) + (p-r)\ln(\lambda)\right]$$

where

$$M = n - r - \frac{1}{6}\left(2(p-r) + 1 + \frac{2}{(p-r)}\right)$$

$$\lambda = \frac{1}{(p-r)}\left(\text{tr}(\mathbf{S}) - \sum_{i=1}^{r} \lambda_i\right)$$

$$df = \frac{1}{2}(p-r-1)(p-r+2)$$

Bartlett's Test continued

- Lawley (1956) gave a modification to Bartlett's test.
- Anderson (1963) discusses related test; that is, the hypothesis that some k intermediate eigenvalues are equal (i.e.,

$$H_o : \lambda_1, \lambda_2, \ldots, \lambda_q, \underbrace{\lambda_{q+1}, \ldots \lambda_{q+k}}_{\text{all equal}}, \lambda_{q+k+1} \ldots, \lambda_p$$

- Bartlett's test — Swiss Bank note example:

$p = 6$, $n = 100$, $r = 3$ and $H_o : \lambda_4 = \lambda_5 = \lambda_6$.

$$\begin{aligned} M &= n - r - \frac{1}{6}\left(2(p-r) + 1 + \frac{2}{(p-r)}\right) \\ &= 100 - 3 - \frac{1}{6}\left(2(6-3) + 1 + \frac{2}{(6-3)}\right) \\ &= 100 - 3 - 1.2777 = 95.722 \end{aligned}$$

$$\lambda = \frac{1}{p-r}\left(\text{tr}(\mathbf{S}) - \sum_{i=1}^{r} \lambda_i\right) = \frac{1}{3}(1.4433 - 1.2340) = .0697$$

Swiss Bank Note Example

$$\det(\mathbf{S}) = .0000135$$

$$\sum_{i=1}^{r} \ln(\lambda_i) = -3.080228$$

$$\begin{aligned} \text{Test Statistic} &= M\left(-\ln(\det(\mathbf{S})) + \sum_{i=1}^{r} \ln(\lambda_i) + (n-r)\ln(\lambda)\right) \\ &= 95.722(-(-11.21447) + (-3.080228) + (100-3)(\ln(\lambda))) \\ &= 14.090177 \end{aligned}$$

$$df = \frac{1}{2}(p-r-1)(p-r+2) = \frac{1}{2}(2)(5) = 5$$

Comparing 14.0902 to a chi-square distribution with $df = 5$, we find p–value$= .02$

How about another value for r? (SAS module).

Graphing Principal Components

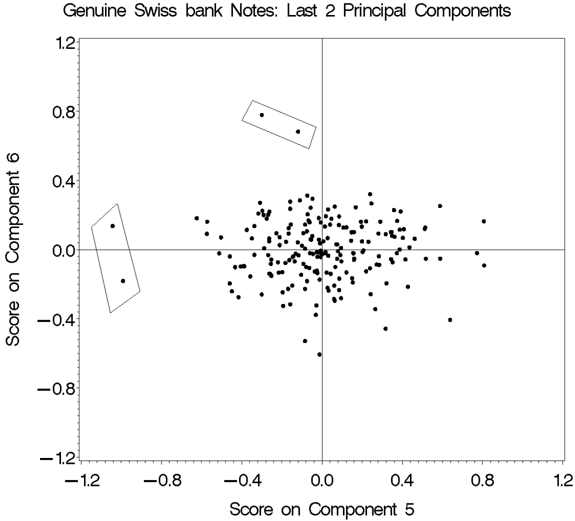
Compute $y_i = \mathbf{e}_i'\mathbf{x}$ and plot these.

- Reveal suspect observations (outliers, influential observations).
- Check multivariate normality assumptions.
- Look for clusters.
- Provide insight into structure in the data.

Suspect Observations

- The first PCs can help reveal influential observations: those that contribute more to variances than other observations such that if we removed them the results change quite a bit.
- The last PCs can help to reveal outliers: those observations that are a typical of the data set; they're inconsistent with the rest of the data (could be miss-coded).

Swiss Bank Notes: Outliers?



Why Look at Last to find Outliers?

Multivariate outliers may not be extreme on any of the original variables. They can still be an outlier in multivariate space because they do not conform with the correlational structure of the rest of the data.

Mathematical explanation: Recall that $\hat{\mathbf{Y}}_{p\times 1} = \hat{\mathbf{P}}_{p\times p}\mathbf{X}_{p\times 1}$ where $\mathbf{P} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p)$.

So since $\mathbf{P}\mathbf{P}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$, $\mathbf{X} = \hat{\mathbf{P}}'\hat{\mathbf{Y}} \Longrightarrow$ The $\mathbf{X}$'s are a linear combination of the principal components (i.e., the $\hat{\mathbf{Y}}$'s).

Consider an observation $\mathbf{x}_j$,

$$
\begin{aligned}
\mathbf{x}_j &= \hat{\mathbf{P}}'\hat{\mathbf{y}}_j \\
&= \hat{y}_{1j}\hat{\mathbf{e}}_1 + \hat{y}_{2j}\hat{\mathbf{e}}_2 + \cdots + \hat{y}_{pj}\hat{\mathbf{e}}_p \\
&= (\hat{y}_{1j}\hat{\mathbf{e}}_1 + \cdots + \hat{y}_{q-1,j}\hat{\mathbf{e}}_{q-1}) + (\hat{y}_{qj}\hat{\mathbf{e}}_q + \cdots + \hat{y}_{pj}\hat{\mathbf{e}}_p)
\end{aligned}
$$

Outliers & Influential Observations

The size (magnitude) of the last PCs determine how well the first few PCs fit observations; that is,

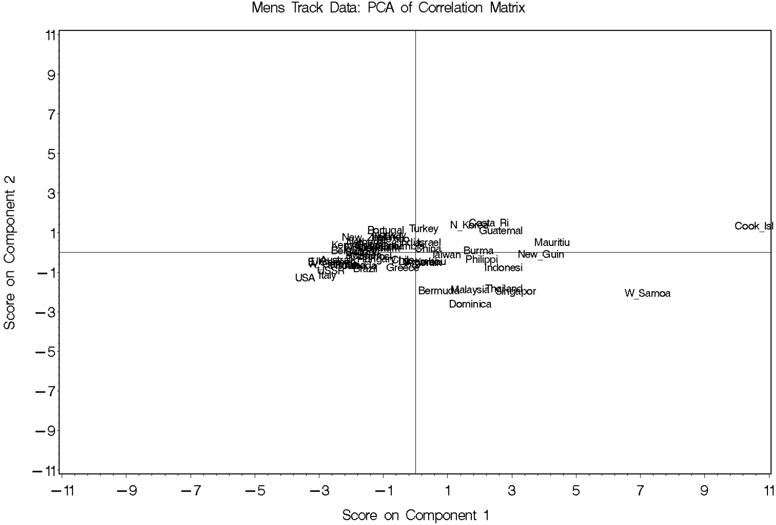
$$(\hat{y}_{1j}\hat{\mathbf{e}}_1 + \cdots + \hat{y}_{q-1,j}\hat{\mathbf{e}}_{q-1}) \text{ differs from } \mathbf{x}_j \text{ by } (\hat{y}_{qj}\hat{\mathbf{e}}_q + \cdots + \hat{y}_{pj}\hat{\mathbf{e}}_p)$$

The suspect observations are the ones where at least one of the coordinates $\hat{y}_{qj}, \ldots, \hat{y}_{pj}$ is large.

The influential observations are also based on the fact that $x_j = \mathbf{P}'\mathbf{y}_j$. Again consider

$$\mathbf{x}_j = \underbrace{(y_{1j}\mathbf{e}_1 + \cdots + y_{q-1,j}\mathbf{e}_{q-1})}_{\text{large } y \text{ values here}} + (y_{qj}\mathbf{e}_q + \cdots + y_{pj}\mathbf{e}_p)$$

Potential Influential Observations in Men's Track



Men's Track Data: Influential Observations?

Western Somoa and the Cook Islands are "off" the scale when we did principal component analysis of the Men's track data.

When we removed these two countries. . .

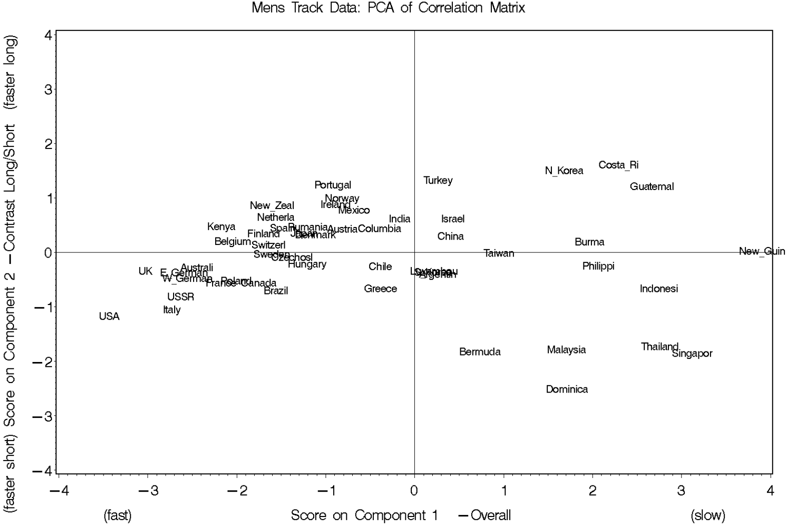
All The Data			Without The Two		
Eigenval.	Prop.	Cum.	Eigenval.	Prop.	Cum.
6.62	0.828	0.828	5.99	0.748	0.748
0.87	0.110	0.938	1.27	0.159	0.907
0.15	0.020	0.957	0.27	0.033	0.941
0.12	0.025	0.973	0.16	0.019	0.960
0.08	0.010	0.983	0.14	0.017	0.977
0.06	0.018	0.991	0.08	0.010	0.987
0.04	0.015	0.997	0.06	0.008	0.996
0.02	0.002	1.000	0.04	0.004	1.000

Change in Component Weights?

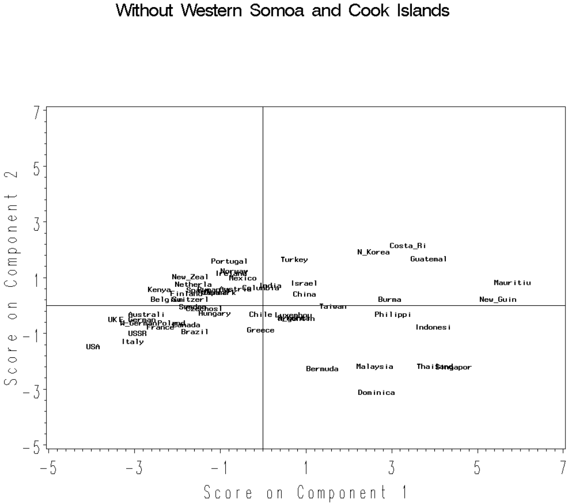
When we removed these two countries...

	All The Data		Without Them	
Race	1	2	1	2
100m	.318	.567	.293	.569
200m	.337	.462	.339	.440
400m	.356	.248	.354	.234
800m	.369	.012	.376	.087
1500m	.373	−.140	.384	−.109
5K	.364	−.312	.366	−.335
10k	.367	−.307	.371	−.324
Marathon	.342	−.439	.337	−.436

Components from All Data



Components without The Two



Checking for Multivariate Normality

If $\mathbf{X}$ is multivariate normal, then $\mathbf{Y}_i = \mathbf{e}_i\mathbf{X}$ should be normal. So we can study the distributions of $\mathbf{Y}_i$'s and also look at pairs of them.

- Scatter Plots.
- Q-Q plots
- Test for distribution

Example using the Swiss Bank note data and SAS/Interactive data analysis.

But since this went away after v9.2, let's use PROC UNIVARIATE for Q-Q plots and test for distribution.

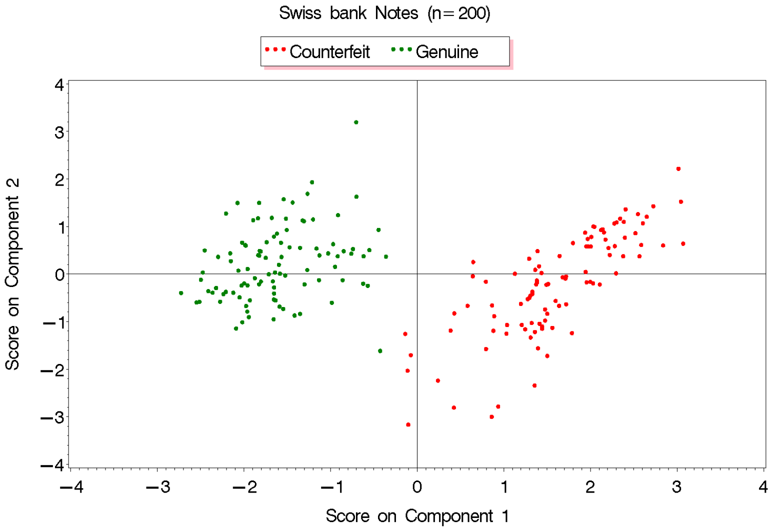
Looking for Patterns and Clusters

Since PCs given an approximations (projections) of higher p–dimensional space, examining plots of the first few PCs may reveal patterns of clusters of observations that we can't otherwise see (e.g., by just plotting distributions and pairs of X variables).

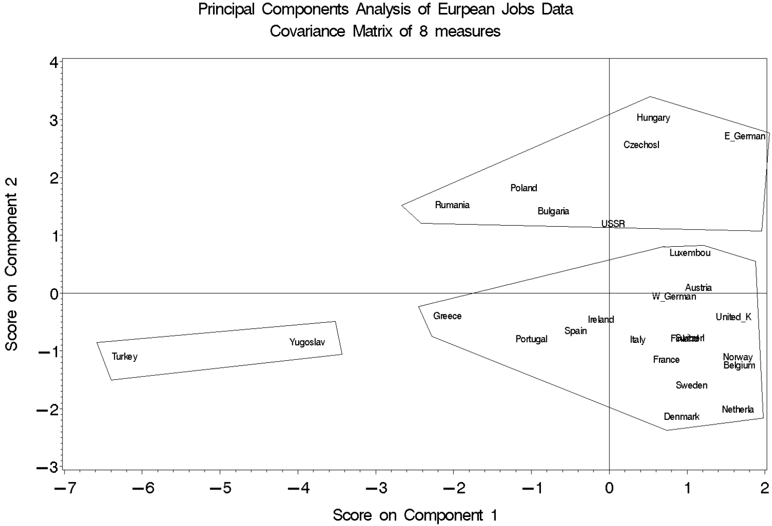
We'll look at 3 examples:

- Swiss Bank Notes.
- European countries.
- Four Psychological Tests.

Swiss Bank Notes: Definite Clusters



European Countries: Clusters



European Countries Variances

The PRINCOMP Procedure

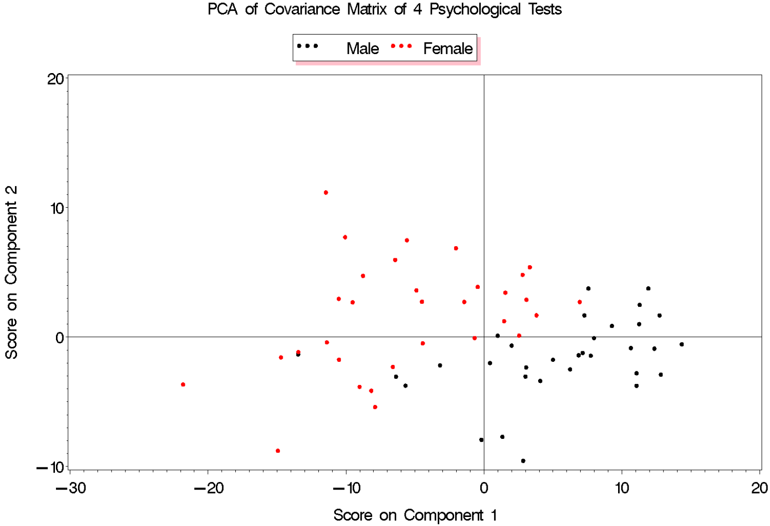
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.48715127	1.35697813	0.3875	0.3875
2	2.13017314	1.03121553	0.2367	0.6241
3	1.09895761	0.10447463	0.1221	0.7463
4	0.99448298	0.45126525	0.1105	0.8568
5	0.54321773	0.15979006	0.0604	0.9171
6	0.38342767	0.15767361	0.0426	0.9597
7	0.22575406	0.08896413	0.0251	0.9848
8	0.13678993	0.13674430	0.0152	1.0000
9	0.00004563		0.0000	1.0000

European Countries Component Weights

	Prin1	Prin2
Percent:		
agriculture	-.523791	0.053594
mining	-.001323	0.617807
manufacturing	0.347495	0.355054
power supply industries	0.255716	0.261096
construction	0.325179	0.051288
service industries	0.378920	-.350172
finance	0.074374	-.453698
social and personal services	0.387409	-.221521
transport and communications	0.366823	0.202592

Psychological Test Data and Pattern



Psychological Test Data and PCA Results

Total Variance 106.23685516

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	72.7174121	56.6068705	0.6845	0.6845
2	16.1105416	2.9961988	0.1516	0.8361
3	13.1143428	8.8197842	0.1234	0.9596
4	4.2945586		0.0404	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
Test1	0.274379	-.001983	0.326835	0.904373
Test2	0.284175	0.184968	0.854066	-.394465
Test3	0.856017	-.408886	-.271343	-.162543
Test4	0.333460	0.893642	-.300205	0.009282

PCA as a Preliminary to Other Analysis

PCA is often used in conjunction with other data and statistical procedures, including

- Multiple regression to overcome problems of multicollinearity (use PCs as independent/predictor variables) or to select a sub-set of the original variables.
- MANOVA
- Discriminant analysis: get a lower-dimensional "look" at structure in data.
- Cluster analysis: Scaling (i.e., PCA) and clustering are often both used when concern is with finding groups of similar objects in a space.

How Many Components to Retain

when want to summarize.

There is no universally accepted method. The decision is largely judgmental and a matter of taste.

Here are some commonly used ones that range from rules-of-thumb to significance tests to heuristic graphicalarguments.

For PCA of Covariance Matrices

- When using sampling distribution results, only retain those that are significantly difference from zero.

  Note: Even with moderate sample sizes, many of the components will typically be statistically significant, even thought these smaller PCs only account for a small percentage of the variance.

- Percent of variance criterion (ad hoc). Use the cumulative variance accounted for as a criterion

$$\frac{\sum_{i=1}^{q}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \quad \text{for} \quad q < p$$

Components to Retain for PCA of R

These are factor analytic-like.

- "Root greater than one" (originally suggested by Kaiser).

  Idea: retain only those PCs with $\lambda_i > 1$, because a PC should account for more variance than any single variables in standardized score space.

	Mens	European	Swiss Bank Notes		
i	track	countries	Genuine	Counterfeit	Both
1	6.62	3.49	2.20	1.94	2.95
2	0.87	2.13	1.70	1.76	1.28
3	0.15	1.10	0.97	0.99	0.87
4	0.12	0.99	0.58	0.78	0.45
5	0.08	0.54	0.33	0.32	0.27
6	0.06	0.38	0.22	0.21	0.19
7	0.04	0.23			
8	0.02	0.14			
9		0.00			

Components to Retain for PCA of R

"Scree Test" (proposed by Cattell)

"Scree" is the rubble at the bottom of a cliff.

Plot eigenvalues of each component in successive order and then identify an "elbow" in the curve (apply a straight edge to the bottom part).

The number of components to retain is given by the point at which the components curve is above the straight edge.

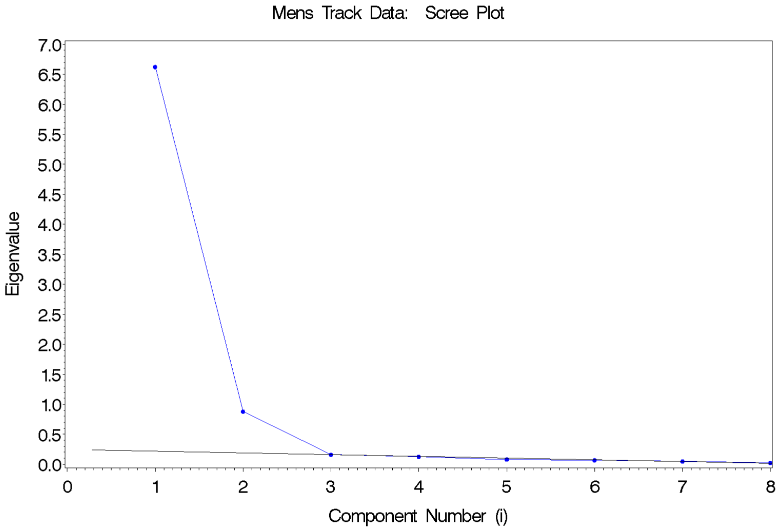
This is like what we did with PCA of covariance matrix when testing $H_o : \lambda_q = \lambda_{q+1} = \cdots + \lambda_p$.

Potential Problems with this:

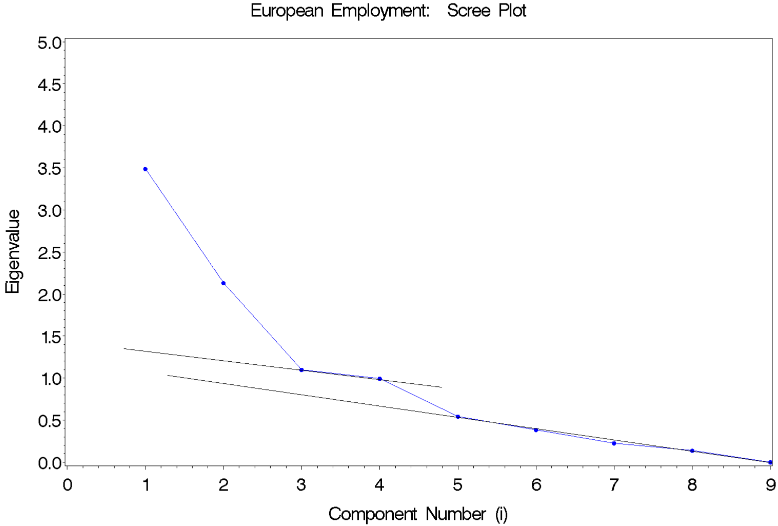
- There may be on obvious break.
- There may be several breaks.

Examples Follow

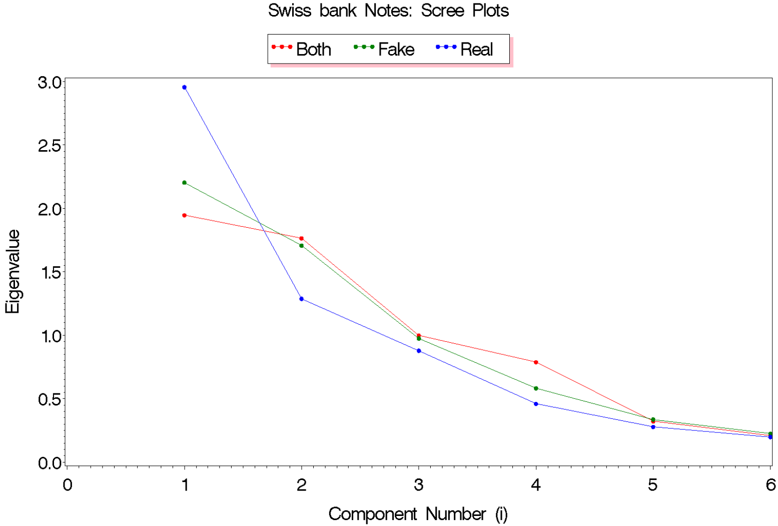
Scree “Test” for Men’s Track Data



Scree “Test” for European Employment Data



Scree "Test" for Swiss Bank Notes



SAS/PROC Princomp

```
title 'Mens Track Data: principal components analysis of R';
proc princomp data=MensTrack out=comscor;
var m100 m200 m400 m800 m1500 K5 K10 Marathon;
```

If you want to use $\mathbf{\Sigma}$ or $\mathbf{S}$, add the "cov" option to the proc princomp statement: `proc princomp out=comscor cov;`

If you want to have text labels as points, you need to create an annotate data set. For example,

```
data coor;
set comscor;
x = prin1;
y = prin2;
xsys = '2';
ysys = '2';
text = Country ;
size = 1;
label x = 'Score on Component 1'
   y = 'Score on Component 2';
keep x y text xsys ysys size;
run;
```

PCA in SAS continued

Now for plotting of components:

```
/* Plot of first two component scores */
goptions reset=(axis, legend, pattern, symbol, title, footnote)
         norotate hpos=0 vpos=0 htext=2.25 ftext=swiss
         ctext= target= gaccess= gsfmode=;
goptions device=win ;
axis2 label=(angle=90 'Score on Component 2')
          order=-5.0 to 7.0 by 2.0;
axis1 label=('Score on Component 1')
           order=-5.0 to 7.0 by 2.0;
proc gplot data=coor;
symbol1 v=none;
plot y*x=1 / annotate=coor frame haxis=axis1
           vaxis=axis2 href=0 vref=0;
title 'Mens Track Data';
run;
```

Distinctions Between PCA & FA

Both Factor Analysis (FA) and PCA are concerned with identification of structure within a set of observed variables. They both establish dimensions within data and both serve as data reduction techniques.

Purposes of FA & PCA

- Reduce the number of variables for further analysis while retaining as much of the original information as possible.
- When the number of variables is so large that it's beyond comprehension, both search the data for qualitative and quantitative distinctions.
- Test hypotheses about qualitative and quantitative distinctions in the data (when appropriate or possible)

FA and PCA are not the same! (PCA is a special case of FA.)

Major Difference Between FA & PCA

There is an underlying latent variable (psychometric) model in factor analysis but there is no such model in PCA.

The model:

$$
\begin{aligned}
(X_1-\mu_1) &= l_{11}F_1 + l_{l2}F_2 + \cdots + l_{1m}F_m + \epsilon_1 \\
(X_2-\mu_2) &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \\
\vdots \quad &\ \vdots \qquad \vdots \\
(X_p-\mu_p) &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \epsilon_p \\
\mathbf{X}_{c,p\times 1} &= \mathbf{L}_{p\times m\, m\times 1} + _{p\times 1}
\end{aligned}
$$

- X_i's are observed variables.
- F's are unobserved random variables.
- ϵ's are unobserved random variables (factors unique to an X).
- The l's are regression coefficients or “factor loadings”; they

Factor Analysis versus PCA

Factor Analysis: Each observed variable is composed of two parts

$$X_i = (\text{common part}) + (\text{unique part})$$

- The common part accounts for the observed relationships between the X_i's.
- The unique part is specific to each observed variable.
- In FA, the new set of variables which are sought express that which is in common among the original variables.
- There is an emphasis on correlations (covariances) between variables. These are due to the common factors and according to the FA model the correlations should be fit "perfectly."

PCA: Defines basic dimensions of the data and makes not assumptions about "common factors" and "unique" parts. The data are taken as given and we attempt to determine the dimensions that account for the total variance; the emphasis is on variance.

More Specific Differences between FA & PCA

- The PCA, we can have 1 variable (i.e., X) define a component (e.g., Men's track data using $\mathbf{S}$ and $X_9 \approx Y_1 =$ marathon). In FA you must have at least 2 variables to define a factor.
- Suppose we initially decide to use 2 factors/components but then decide to add one more:
  - In PCA, the 1st two components are unaltered.
  - In FA, the 1st two factors could be different.

More Specific Differences between FA & PCA

In FA, after an initial solution is found, they are often "rotated" (orthogonal or oblique) to find a more substantively interpretable pattern of loadings (i.e., the l_{ik}'s).

In PCA, you do not rotate!. PCs are defined as those linear combinations of the observed variable that maximize variance. PCA is a rotation. If you rotate, they aren't PCs anymore. Also, the components won't necessarily have substantively meaningful interpretations.

PCA can be calculated exactly from $\mathbf{X}$, but FA cannot. PCs are linear combinations (functions) of $\mathbf{X}$, but the factors are not. There is a non-exact relationship between factors and $\mathbf{X}$ because of the present of uniqueness. Factor scores must be estimated (and there are a number of ways of doing this.