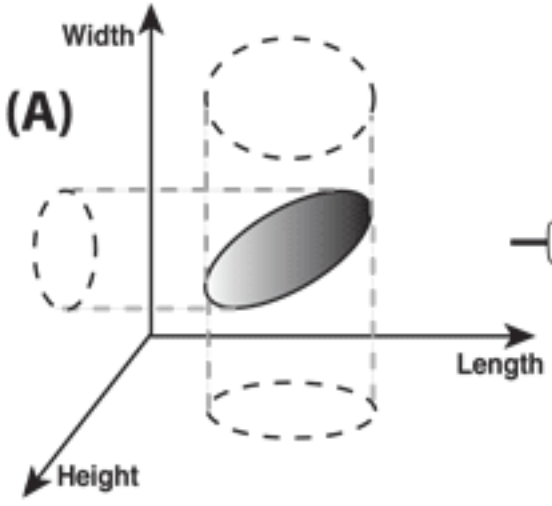
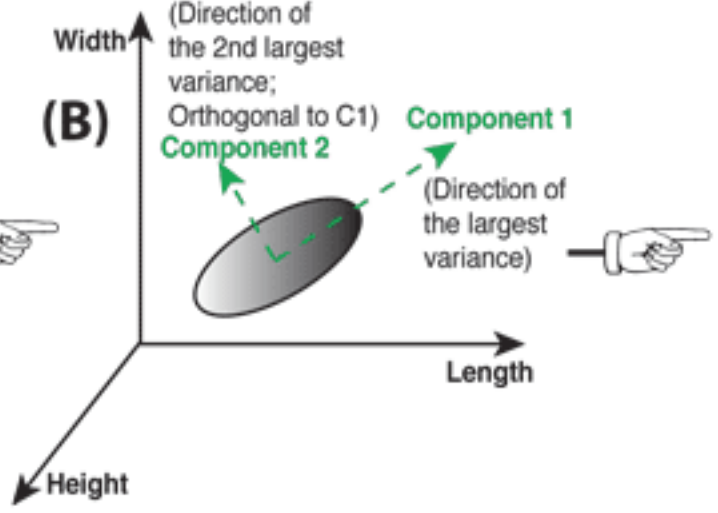
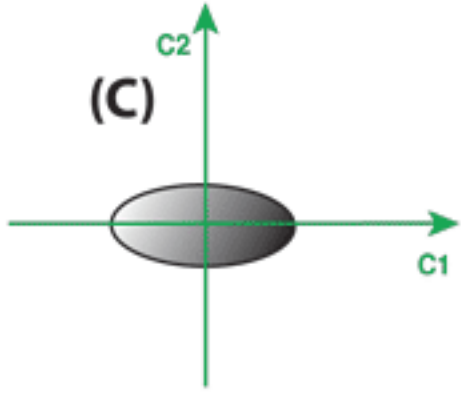
Principal component analysis is all about how to choose a good coordinate system







Data(i) = f(L(i), W(i), H(i))
[3-dimentional function]

Data(j) = g(C1(j), C2(j))
[2-dimential function]

Principal Component Analysis & Factor Analysis

Contents:

Principal Component Analysis

- Technique for transforming original variables into new ones which are uncorrelated and account for decreasing proportions of variance in the data.
- New variables are linear combinations of old ones.
- Transformation is a rotation/reflection of original points no essential statistical information is lost (or 'created').
- Can assess the importance of individual new components and assess 'how many are needed' (–scree plots etc).

- Can assess the importance of original variables (examination of loadings).

- *The objective of the PCA is to transform a set of interrelated variables into a set of unrelated linear combinations of variables. The set of linear combinations is chosen so that each of the linear combinations accounts for a decreasing proportion of the variance in the original variables.*

Objective

- Given the n variables $x_1, \ldots, x_n$ and $X = (x_1, \ldots, x_n)$, the objective is to find a linear transformation of $X \to Y = (y_1, \ldots, y_n)$ such that:

  The 1st component y_1 is the *most interesting*.
  The 2nd component y_2 is the *2nd most interesting*.
  The 3rd component y_3 is the *3rd most interesting*, etc.

- That is, we want to chose a new co-ordinate system so that the data, when refer to this new system, Y, are such that, the **1st** component contains *most information*, the 2nd component contains the next *most information*, etc.

- The hope is that the first *few* (2, 3 or 4 say) components contain *nearly all* the information in the data and the remaining 2,3,4 components contain relatively little information and can be *discarded*. I.e. the statistical analysis can be concentrated on just the first few components (much easier).

- A linear transformation $X \to Y$ is given by $Y = XQ$, where Q is an $n \times n$ non-singular matrix. If Q happens to be an orthogonal matrix, i.e. $Q^T Q = I_n$, then the transformation $X \to Y$ is an orthogonal transformation.

Specific

- The basic idea is to find a set of orthogonal coordinates such that the sample variances of the data with respect to these coordinates are in decreasing order of magnitude, i.e. the projection of the points onto the 1st principal component has maximal variance among all such linear projections, the projection onto the 2nd has maximal variance subject to orthogonality with the first, projection onto the 3rd has maximal variance subject to orthogonality with the first two, etc.

- most interesting – most information – *maximum variance*.

- This objective can be achieved by an eigenanalysis of the variance matrix S of the data, i.e. the matrix of all two-way covariances between the variables and variances along the diagonal. It can be shown that if we transform the original data to principal components then

  (a) the sample variances of the data on successive components are equal to the eigenvalues of the variance matrix S of the data;

(b) the total variation is exactly the same on the complete set of principal components as on the original variables, so no information is lost. It is just rearranged into order.

- Let S denote the $n \times n$ variance-covariance matrix of the n variables $x_1, \ldots, x_n$. E.g. let

$$X = (x_1 \;\; x_2 \;\; x_3) \quad \text{and} \quad \text{Var}(X) = S = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

- Since S is symmetric and positive definite matrix, it has real eigenvalues

$$\boxed{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0.}$$

- The spectral decomposition of S is given by:

$$Q^T S Q = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}.$$

E.g.

$$\begin{pmatrix} -0.38 & 0.92 & 0 \\ 0 & 0 & 1 \\ 0.92 & 0.38 & 0 \end{pmatrix} \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} -0.38 & 0 & 0.92 \\ 0.92 & 0 & 0.38 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 5.83 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0.17 \end{pmatrix}.$$

- Let y_i to denote a transformed variable, where $Y = (y_1, \ldots, y_n)$ and $Y = XQ$. That is,

$$(y_1 \ y_2 \ y_3) = (x_1 \ x_2 \ x_3) \begin{pmatrix} -0.38 & 0 & 0.92 \\ 0.92 & 0 & 0.38 \\ 0 & 1 & 0 \end{pmatrix}$$

or

$$\begin{aligned} y_1 &= -0.38x_1 + 0.92x_2 \\ y_2 &= x_3 \\ y_3 &= 0.92x_1 + 0.38x_2 \end{aligned}$$

- The variance-covariance matrix of $\mathbf{Y} = (\mathbf{y_1}, \ldots, \mathbf{y_n})$ is given by:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(XQ) \\ &= Q^T \mathrm{Var}(X) Q \qquad (\text{where } \mathrm{Var}(X) = S) \\ &= Q^T S Q \\ &= \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_n \end{pmatrix} \end{aligned}$$

Verification

Note that

$$\mathrm{Var}(\alpha X + bY) = \alpha^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) + 2\alpha b \mathrm{Cov}(X,Y).$$

$$\begin{aligned}
\mathrm{Var}(y_1) &= \mathrm{Var}(-0.38x_1 + 0.92x_2) \\
&= (-0.38)^2 \mathrm{Var}(x_1) + (0.92)^2 \mathrm{Var}(x_2) \\
&\quad + 2(-0.38)(0.92)\mathrm{Cov}(x_1, x_2) \\
&= 0.15(1) + 0.85(5) - 0.7(-2) \\
&= 5.8 = \lambda_1 .
\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(y_1,y_2) &= \mathrm{Cov}(-0.38x_1+0.92x_2,\ x_3)\\ &= -0.38\mathrm{Cov}(x_1,x_3)+0.92\mathrm{Cov}(x_2,x_3)\\ &= -0.38(0)+0.92(0)\\ &= 0.\end{aligned}$$

- $y_1,\ldots,y_n$ are the principal components of $x_1,\ldots,x_n$.
- y_i has variance λ_i.
- y_i is uncorrelated with y_j $(i \neq j)$, since Λ, i.e. the covariance of Y, is diagonal.

- From $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ it follows that y_1 has the largest variance λ_1, y_2 has the second largest variance λ_2, and so on.

$$\lambda_1 = 5.83, \quad \lambda_2 = 2 \quad \lambda_3 = 0.17.$$

- The total variation in the original data is the sum of the variances of the original variables $x_1, \ldots, x_n$. That is,

$$S_{11} + S_{22} + \ldots S_{nn} = \textbf{trace}(S).$$

- Notice that:

$$\begin{aligned}\text{trace}(S) &= \text{trace}(Q^T \Lambda Q) \\ &= \text{trace}(\Lambda Q Q^T) \quad (\text{since } \text{trace}(AB) = \text{trace}(BA)) \\ &= \text{trace}(\Lambda) \quad (\text{since } QQ^T = I_n.) \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_n.\end{aligned}$$

which is the sum of variances of the n PCs $y_1, \ldots, y_n$.

$$S_{11} + S_{22} + \ldots S_{nn} = 8 = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

- Thus, the total variation is exactly the same on the complete set of principal components as on the original variables. So no information is lost – it is just rearranged into order.

- The sum of the variances $\sum_{i=1}^{n} \lambda_i$ of the n PCs $y_1, \ldots, y_n$ is the same as the sum of the variances $\sum_{i=1}^{n} S_{ii}$ of the original variables $x_1, \ldots, x_n$.

- The components with smaller variances could be ignored without significantly effective the total variance, thereby reducing the number of variables from n.

- The 'total' variation of the PCs is $\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} S_{ii} = \text{trace}(S)$. So we can interpret

$$\lambda_1 / \sum_{i=1}^{n} \lambda_i$$

as the proportion of the total variation 'explained' by the first principal component. In the example $\lambda_1 / \sum \lambda_i = 5.83/8 = 0.73$.

- The proportion of the total variation 'explained' by the first two PCs is given by $(\lambda_1 + \lambda_2) / \sum_{i=1}^{n} \lambda_i$.

$$(\lambda_1 + \lambda_2) / \sum \lambda_i = (5.83 + 2)/8 = 0.98.$$

- In general the proportion of the total variation 'explained' by the first k PCs is given by

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}.$$

- If the first 'few' PCs explain 'most' of the variation in the data, then the later PCs are redundant and 'little' information is lost if they are discarded (or ignored).

E.g. if $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} = \text{say } 80+\%,$

then the $(k+1)$th,...,nth components contain relatively little information and the dimensionality of the data can be reduced from n to k with little loss of information. Useful if $k = 1, 2, 3, 4?, 5???$

The figure of 80% is quite arbitrary and depends really on the type of data being analyzed – particularly on the applications area. Some areas might conventionally be satisfied if 40% of the variation can be explained in a few PCs, others might require 90%.

Scree-plot

- A figure (percentage) needs to be chosen as a trade-off between the convenience of a small value of k and a large value of the cumulative relative proportion of variance explained.

- If n is large an informal way of choosing a 'good' k is graphically with a *scree-plot*.

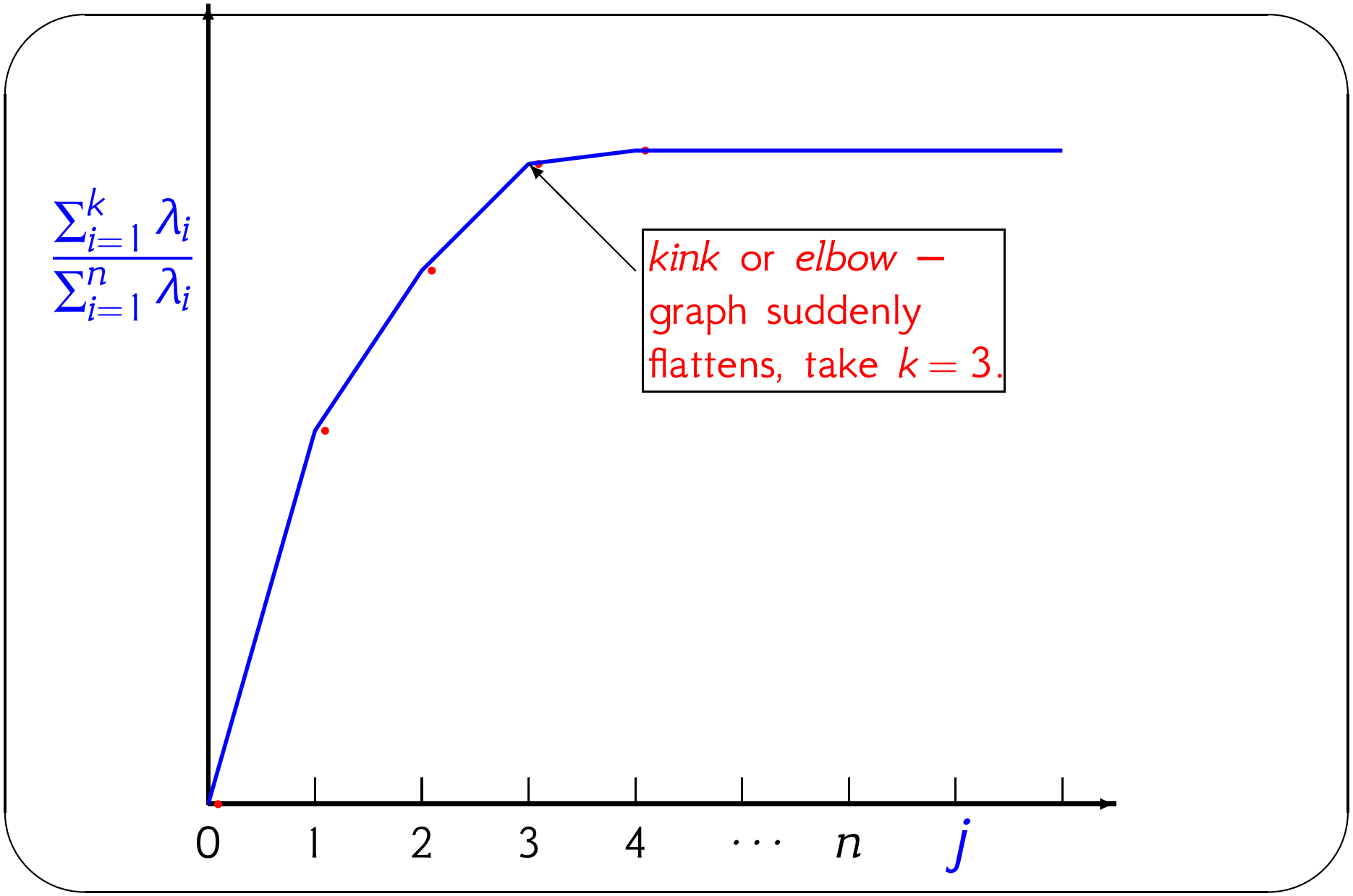
  That is, plot $\dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$ vs j.

- The graph (scree-plot) will be necessarily monotonic and convex.

  Typically it will increase steeply for the first few values of j (i.e. the first few PCs) and then begin to level off.

  The point where it starts leveling off is the point where bringing in more PCs brings less returns in terms of variance explained.

- Some formal tests of $H : \lambda_i = 0$ are available (but not used much), similarly expressions for confidence intervals for λ_i.


$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$
kink or elbow – graph suddenly flattens, take $k=3$.
0
1
2
3
4
$\cdots$
n
j

Analytical Approach

Given n variables the objective of the principal components is to form n linear combinations:

$$\begin{aligned} y_1 &= w_{11}x_1 + w_{12}x_2 + \cdots + w_{1n}x_n \\ y_2 &= w_{21}x_1 + w_{22}x_2 + \cdots + w_{2n}x_n \\ &\vdots \\ y_n &= w_{n1}x_1 + w_{n2}x_2 + \cdots + w_{nn}x_n. \end{aligned}$$

Here

- $y_1, y_2, \ldots y_n$ are the n principal components.

- w_{ij} is the weight of the jth variable for the ith principal component.
- $\text{Var}(y_1) > \text{Var}(y_2) > \cdots > \text{Var}(y_n)$.
- $\sum_{k=1}^{n} w_{ik}^2 = w_{i1}^2 + \cdots + w_{in}^2 = 1$.
- $\sum_{k=1}^{n} w_{ik} w_{jk} = w_{i1} w_{j1} + \cdots + w_{in} w_{jn} = 0$.

  Example:

  $y_1 = 0.728 x_1 + 0.685 x_2$ and $y_2 = -0.685 x_1 + 0.728 x_2$.

  $0.728^2 + 0.685^2 = 1$ and

  $0.728 \times (-0.685) + 0.685 \times 0.728 = 0$.

Principal Components Scores and Loadings

- The principal component scores are the values (output) of the principal component variables.
- The *loading* is the simple correlation between the original and the new (principal component) variables.

  This is an indication of the extent to which the original variables are influential or important in forming the principal components. That is, the higher the loading the more influential the variable is in forming the principal component scores and vice-versa.

- The loadings can be obtained from the relationship:

$$l_{ij} = \frac{w_{i,j}}{\hat{\sigma}_j}\sqrt{\lambda_i} \ ,$$

where

1. l_{ij} is the loading of the jth variable to the ith principal component.
2. $w_{i,j}$ is the weight of jth variable to the ith principal component.
3. $\hat{\sigma}_j$ is the standard deviation of the jth variable.
4. λ_i is the eigenvalue (i.e. variance) of the ith principal component.

Example: Financial ratios X_1 and X_2

The table presents a small data set consisiting of 12 observations and 2 variables X_1 and X_2 (financial ratios).

The table also gives the mean corrected data (denoted by X_1^* and X_2^*), the **SSCP**, the covariance matrix **S** and the correlation matrix **R**.

The mean corrected variables are transformed (rotated) using the orthogonal matrix (rotation).

The rotation that took place is:

$$\begin{pmatrix} P_1 & P_2 \end{pmatrix} = \begin{pmatrix} X_1^* & X_2^* \end{pmatrix} \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

E.g. if $\theta = 43.261$, then $\cos(\theta) = 0.728$ and $\sin(\theta) = 0.685$.

For example,

$$\begin{pmatrix} 9.25 & -1.84 \end{pmatrix} = \begin{pmatrix} 8 & 5 \end{pmatrix} \begin{pmatrix} 0.728 & -0.685 \\ 0.685 & 0.728 \end{pmatrix}.$$

	Original		Mean Corrected		Rotated by 43.26^o	
	X_1	X_2	X_1^*	X_2^*	P_1	P_2
	16	8	8	5	9.25	-1.84
	12	10	4	7	7.71	2.36
	13	6	5	3	5.70	-1.24
	11	2	3	-1	1.50	-2.78
	10	8	2	5	4.88	2.27
	9	-1	1	-4	-2.01	-3.60
	8	4	0	1	0.69	0.73
	7	6	-1	3	1.33	2.87
	5	-3	-3	-6	-6.29	-2.31
	3	-1	-5	-4	-6.38	0.51
	2	-3	-6	-6	-8.48	-0.26
	0	0	-8	-3	-7.88	3.30
Mean:	8	3	0	0	0	0
Var:	23.09	21.09	23.09	21.09	38.58	5.60

Original Variables:

$$\mathbf{SSCP} = \begin{pmatrix} 254 & 181 \\ 181 & 232 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 23.09 & 16.46 \\ 16.46 & 21.09 \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{pmatrix}$$

New Variables:

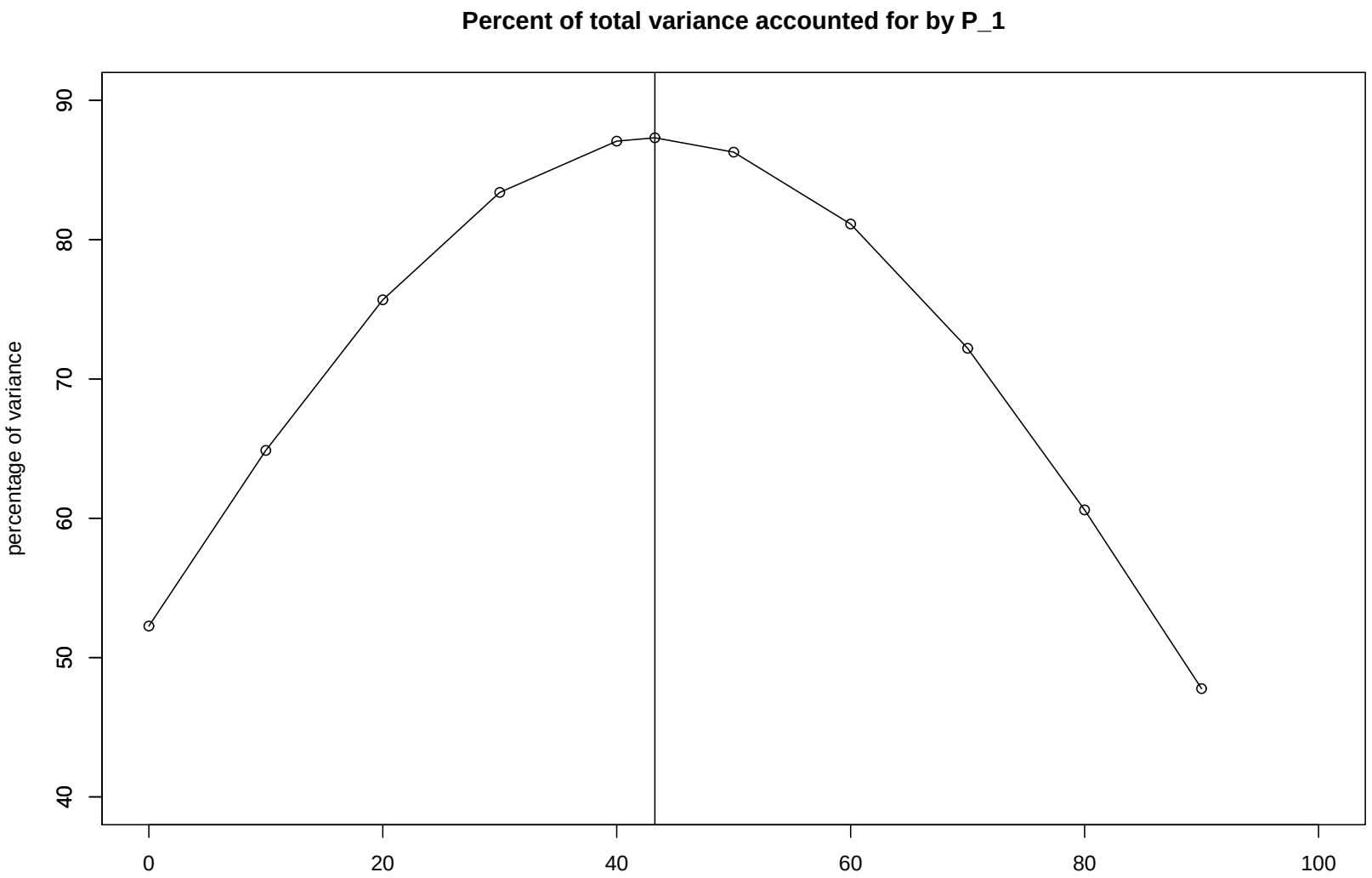
$$\mathbf{SSCP} = \begin{pmatrix} 424.33 & 0.0 \\ 0.0 & 61.67 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 38.58 & 0.0 \\ 0.0 & 5.60 \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$

The variance accounted for by the new variables P_1 and P_2 for various angles (rotations).

Rotate by Angle θ	Total Variance	Variance for P_1	Percent %
0	44.12	23.09	52.26
10	44.12	28.66	64.87
20	44.12	33.43	75.68
30	44.12	36.84	83.39
40	44.12	38.47	87.07
43.261	44.12	38.58	87.31
50	44.12	38.12	86.28
60	44.12	35.84	81.12
70	44.12	31.90	72.20
80	44.12	26.78	60.60
90	44.12	21.09	47.77


Percent of total variance accounted for by P_1
percentage of variance
40
50
60
70
80
90
0
20
40
60
80
100
Angle theta for rotation

Computer Output: Financial ratios X_1 and X_2

Simple Statistics

	X_1	X_2
Mean	8.00	3.00
St. Dev	4.81	4.59

Covariances

	X_1	X_2
X_1	23.09	16.45
X_2	16.45	21.09

Total Variance $= 44.18$

Importance of Components:

	Comp.1	Comp.2
Standard deviation	6.21	2.37
Proportion of Variance	0.87	0.13
Cumulative Proportion	0.87	1.00

Eigenvectors:

	Comp.1	Comp.2
X_1	0.728	0.685
X_2	0.685	-0.728

Correlation Coefficients:

	X_1	X_2	Comp.1	Comp.2
X_1	1.00	0.75	0.94	0.34
X_2	0.75	1.00	0.93	-0.38
Comp.1	0.94	0.93	1.00	-9.9e-17
Comp.2	0.34	-0.38	-9.9e-17	1.00

Scores:

X_1	X_2	Comp.1	Comp.2
16	8	9.25	1.84
12	10	7.71	-2.36
13	6	5.70	1.24
11	2	1.50	2.78
10	8	4.88	-2.27
9	-1	-2.01	3.60
8	4	0.69	-0.73
7	6	1.33	-2.87
5	-3	-6.30	2.31
3	-1	-6.38	-0.51
2	-3	-8.48	0.26
0	0	-7.88	-3.30

Simple Statistics:

	Mean	St.Dev
X_1	8.00	4.81
X_2	3.00	4.59
Comp.1	7.4e-17	6.21
Comp.2	-1.5e-16	2.36

Summary:

- The total variances of X_1 and X_2 is 44.18 (i.e. $23.09 + 21.09$).
- The variables X_1 and X_2 have correlation coeeficient 0.746.
- The percentage of the total variance accounted for by X_1 and X_2 are given respectively, by 52.26% and 47.74%.
- Each of the new variables (i.e. principal cmpoinents P_1 and P_2) are linear combinations of the original variables and remain mean corrected. That is their means are zero.

- The total SS (Sum of Squares) for P_1 and P_2 is the same as the total SS for the original variables ($424.33 + 61.67 = 486$).
- The variances of P_1 and P_2 are, respectively, 38.58 and 5.61. The total variance of the principal components is 44.18 and is the same as the total of the variance of the original variables X_1 and X_2.
- The percentage of the total variance accounted for by P_1 and P_2 are, respectively, 87.31% ($= 38.58/44.18$) and 12.69% ($= 5.61/44.18$).

- The variance accounted for by the first principal component P_1 is greater than the variance accounted for by any one of the original variables.
- The second principal component P_2 accounts for variance that has not been accounted for by P_1. The two principal components accounts for all the variance in the data.
- The correlation between the principal components is zero, that is, P_1 and P_2 are uncorrelated.

Effect of type of data on PCA

Consider the data below which shows the *Estimated Retail Prices by Cities, March 1973, U.S. Department of Labour, Boureau of Labor Statistics, pp1-8.*

City	Average Price (in cents per pound) Bread	Burger	Milk	Oranges	Tomatoes
Atlanta	24.5	94.5	73.9	80.1	41.6
Baltimore	26.5	91.0	67.5	74.6	53.3
Boston	29.7	100.8	61.4	104.0	59.6
Buffalo	22.8	86.6	65.3	118.4	51.2
Chicago	26.7	86.7	62.7	105.9	51.2
Cincinnati	25.3	102.5	63.3	99.3	45.6
Cleveland	22.8	88.8	52.4	110.9	46.8
Dallas	23.3	85.5	62.5	117.9	41.8
Detroit	24.1	93.7	51.5	109.7	52.4
Honolulu	29.3	105.9	80.2	133.2	61.7
Houston	22.3	83.6	67.8	108.6	42.4
Kansas City	26.1	88.9	65.4	100.9	43.2
Los Angeles	26.9	89.3	56.2	82.7	38.4
Milwaukee	20.3	89.6	53.8	111.8	53.9
Minneapolis	24.6	92.2	51.9	106.0	50.7
. . . . .					

City	Average Price (in cents per pound) Bread	Burger	Milk	Oranges	Tomatoes
Atlanta	24.5	94.5	73.9	80.1	41.6
.....					
Minneapolis	24.6	92.2	51.9	106.0	50.7
New York	30.8	110.7	66.0	107.3	62.6
Philadelphia	24.5	92.3	66.7	98.0	61.7
Pittsburgh	26.2	95.4	60.2	117.1	49.3
St. Louis	26.5	92.4	60.8	115.1	46.2
San Diego	25.5	83.7	57.0	92.8	35.4
San Francisco	26.3	87.1	58.3	101.8	41.5
Seattle	22.5	77.7	62.0	91.1	44.9
Washington DC	24.2	93.8	66.0	81.6	46.2
Mean	25.3	91.9	62.3	103.0	48.8
Variance	6.3	57.1	48.3	202.8	57.8
% of total Variance:	1.7	15.3	13.0	54.5	15.5

Total Variance: 372.22

The objective is to form a measure of the Consumer Price Index (CPI). That is, we would like to form a weighted sum of the various food prices that would summarize how expensive or cheap are a given city's food items. PCA would be an appropriate technique for developping such an index.

Principal components analysis can be either done on mean-corrected or standardized data. Each data set could givce a different solution depending upon the extend to which the variances of the variables differ. That is, the variances of the variables could have an effect on PCA.

PCA:

Simple Statistics

	Bread	Burger	Milk	Oranges	Tomatoes
Mean	25.29	91.86	62.30	102.99	48.77
St. Dev	2.51	7.55	6.95	14.24	7.60

Covariance Matrix

	Bread	Burger	Milk	Oranges	Tomatoes
Bread	6.284	12.91	5.7191	1.3104	7.285
Burger	12.911	57.08	17.5075	22.6919	36.295
Milk	5.719	17.51	48.3059	-0.2750	13.443
Oranges	1.310	22.69	-0.2750	202.7563	38.762
Tomatoes	7.285	36.29	13.4435	38.7624	57.801

Importance of Components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	14.799	9.577	6.137	4.5619	1.74047
Proportion of Variance	0.588	0.246	0.101	0.0559	0.00814
Cumulative Proportion	0.588	0.835	0.936	0.9919	1.00000

Eigenvectors:

	PC1	PC2	PC3	PC4	PC5
Bread	0.028	0.165	-0.021	0.190	-0.967
Burger	0.200	0.632	-0.254	0.659	0.249
Milk	0.042	0.442	0.889	-0.108	0.036
Oranges	0.939	-0.314	0.121	0.069	-0.015
Tomatoes	0.276	0.528	-0.361	-0.717	-0.034

Correlation Coefficients:

	Bread	Burger	Milk	Oranges	Tomatoes
PC1	0.168	0.39	0.089	0.976	0.54
PC2	0.632	0.80	0.609	-0.211	0.67
PC2	-0.052	-0.21	0.785	0.052	-0.29

Scores:

	PC1	PC2	PC3	PC4	PC5
Baltimore	-25.33	13.3	-0.27	-6.106	-0.92
Los Angeles	-22.63	-3.1	-3.52	5.307	-1.75
Atlanta	-22.48	10.1	9.47	3.897	2.44
Washington DC	-20.28	8.1	1.15	1.036	2.09
Seattle	-15.15	-7.8	3.35	-7.872	-0.52
⋮	⋮	⋮	⋮	⋮	⋮
Dallas	10.76	-12.6	6.16	1.436	0.36
New York	11.94	20.4	-6.09	3.437	-1.05
Pittsburgh	14.04	-2.7	-1.26	3.323	-0.31
Buffalo	14.14	-6.0	5.05	-4.940	0.89
Honolulu	35.60	14.8	11.25	0.896	-0.64

The first principal component PC1 is given by:

$$\begin{aligned}\text{PC1} =& 0.028\,\text{Bread} + 0.2\,\text{Burger} + 0.042\,\text{Milk}\\ &+ 0.939\,\text{Oranges} + 0.276\,\text{Tomatoes}.\end{aligned}$$

The variance of PC1 is 218.99 and acounts for 58.8% of the total variance of the original data. The PC1 is the sum of all food prices and is very much affected by the price of oranges.

Since all the weights of PC1 are positive a high score will imply that the food prices are high and vice-versa. Thus, from the score (values) of PC1 suggests that *Honolulu* is the most expensive city and *Bαltimore* is the

least expensive city.

The main reason the price of oranges dominates the formation of PC1 is that there exist a wide variation in the price of oranges across the cities. That is, the variance of the price for oranges is very high compared to the variances of the priuces of other food items.

In general, the weight assigned to a variable is affected by the relative variance of the variable. If we do not want the relative variance to affect the weights, then the data should be standartized so that the variance of each variable is the same (i.e. one).

PCA after standartizing the data:

Correlation Matrix

	Bread	Burger	Milk	Oranges	Tomatoes
Bread	1.000	0.68	0.3282	0.0367	0.38
Burger	0.682	1.00	0.3334	0.2109	0.63
Milk	0.328	0.33	1.0000	-0.0028	0.25
Oranges	0.037	0.21	-0.0028	1.0000	0.36
Tomatoes	0.382	0.63	0.2544	0.3581	1.00

Importance of Components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.556	1.051	0.859	0.7026	0.4907
Proportion of Variance	0.484	0.221	0.148	0.0987	0.0481
Cumulative Proportion	0.484	0.705	0.853	0.9518	1.0000

Eigenvectors:

	PC1	PC2	PC3	PC4	PC5
Bread	0.496	-0.3086	-0.3864	0.5093	0.49990
Burger	0.576	-0.0438	-0.2625	-0.0281	-0.77264
Milk	0.340	-0.4308	0.8346	0.0491	-0.00788
Oranges	0.225	0.7968	0.2916	0.4790	0.00597
Tomatoes	0.506	0.2870	-0.0123	-0.7127	0.39120

Correlation Coefficients:

	Bread	Burger	Milk	Oranges	Tomatoes
PC1	0.772	0.896	0.529	0.350	0.788
PC2	-0.324	-0.046	-0.453	0.837	0.302

Scores:

	PC1	PC2	PC3	PC4	PC5
Seattle	-2.14	-0.376	0.6639	-0.567	0.703
San Diego	-1.93	-0.741	-0.5847	0.967	0.194
Houston	-1.32	0.152	1.5685	0.253	-0.085
Cleveland	-1.24	1.336	-0.5450	-0.117	-0.277
Los Angeles	-1.21	-1.362	-1.3190	0.595	0.048
⋮	⋮	⋮	⋮	⋮	⋮
Pittsburgh	0.62	0.825	-0.2319	0.594	-0.149
Philadelphia	0.89	0.032	0.5239	-1.546	0.466
Boston	2.30	-0.075	-1.1192	-0.129	0.535
New York	3.78	-0.259	-1.0153	-0.079	-0.122
Honolulu	4.17	0.505	1.6790	0.708	0.022

Since the data are standartized, the variance of each variable is one and each variable accounts for 20% of the total variance.

The first principal component accounts for 48.84% ($= 1.556^2/5$) of the total variance and is given by:

$$\begin{aligned} \text{PC1} =& 0.496\,\text{Bread} + 0.576\,\text{Burger} + 0.340\,\text{Milk} \\ & + 0.225\,\text{Oranges} + 0.506\,\text{Tomatoes}. \end{aligned}$$

The second principal component accounts for 22.1% ($= 1.051^2/5$) of the total variance and is given by:

$$\begin{aligned} \text{PC2} = &-0.309\,\text{Bread} - 0.044\,\text{Burger} - 0.431\,\text{Milk} \\ &+ 0.797\,\text{Oranges} + 0.287\,\text{Tomatoes}. \end{aligned}$$

The PC1 is a weighted sum of all the food prices and no one food item dominates the formation of the score.

The value of PC1 suggests that *Honolulu* is the most expensive city and *Seαttle* is now the least expensive city, as compared to *Bαltimore* when the data were not standartized.

Therefore, the weights that are used to form the CPI are affected by the relative variances of the variables.

The decision of how many principal components should retain is dependent on how much information (i.e. uncounted variance) one is willing to sacrifice, which is a judgement question.

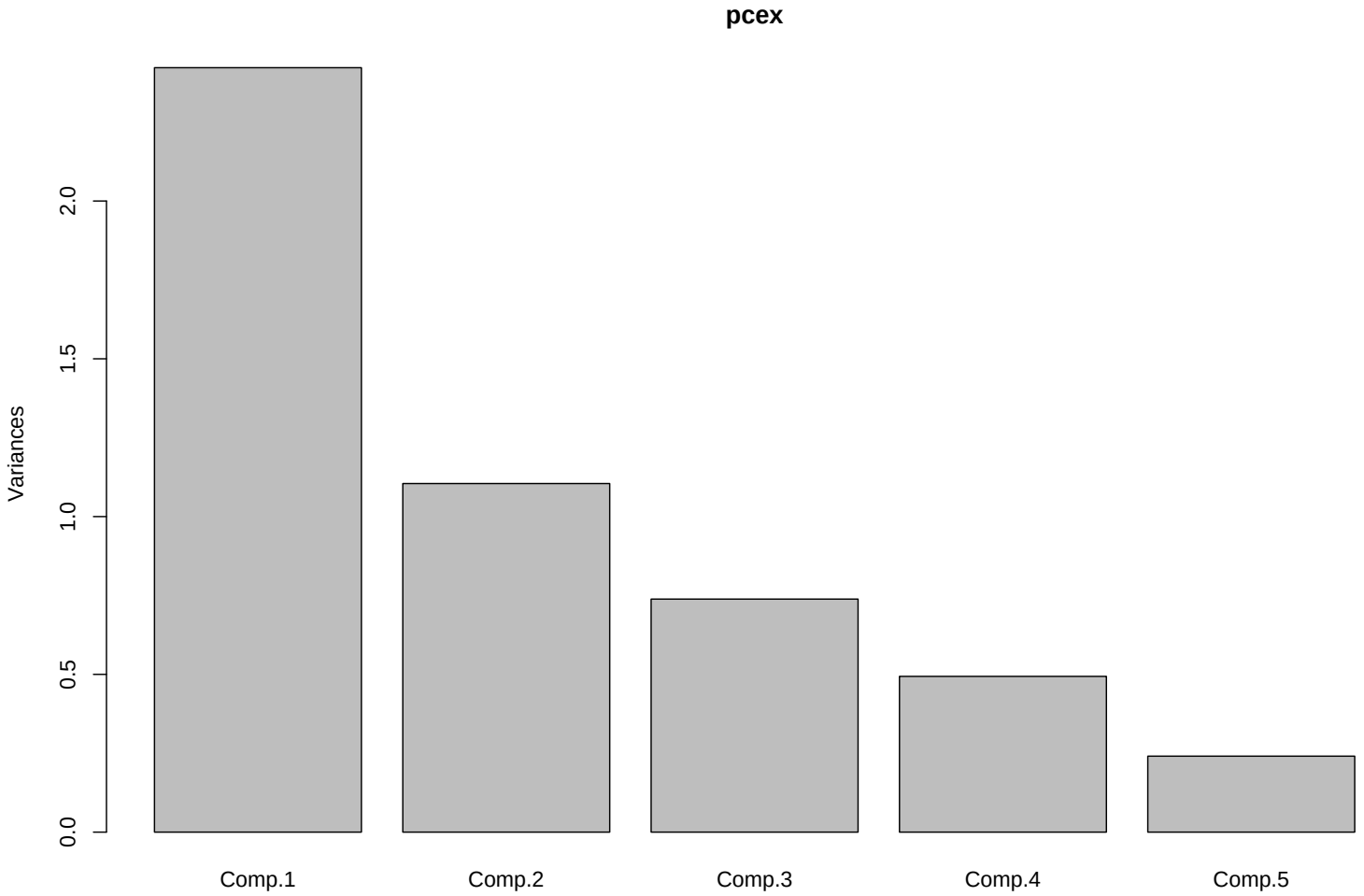
Two alternatives are:

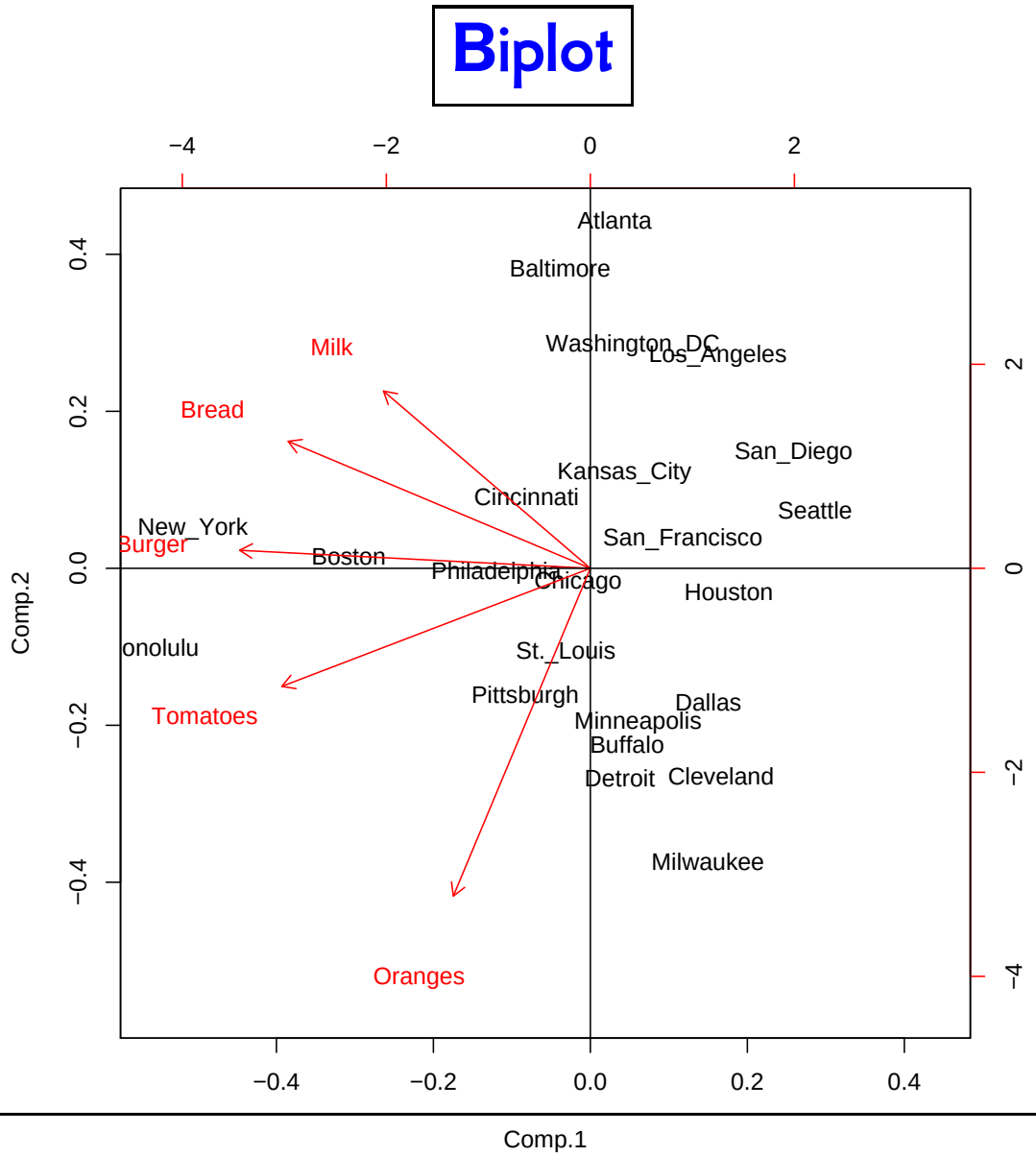
- Use the scree plot and look for an elbow. The rule can be used for both mean-corrected and standartized data.

- In the case of standartized data retain only those components whose eigenvalues (variance) are greater than one. this is referred to as the *eigenvalue-greater-than-one* rule.

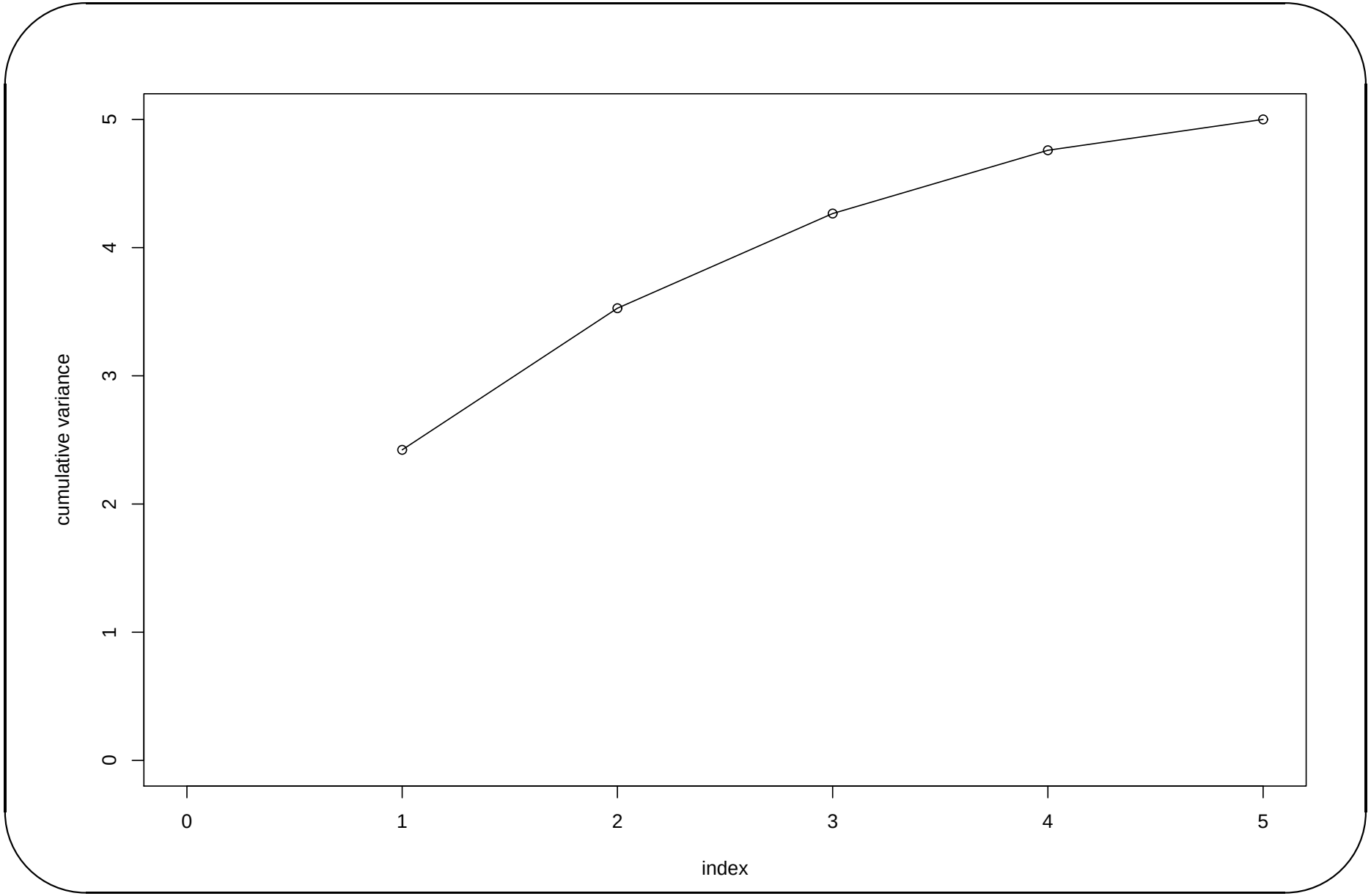
  This rule is the default option in most of the statistical packages (SAA and SPSS). The rationale for this rule is that for standartize data the amount of variance extracted by each component should at minimum be equal to the variance of at least one variable.

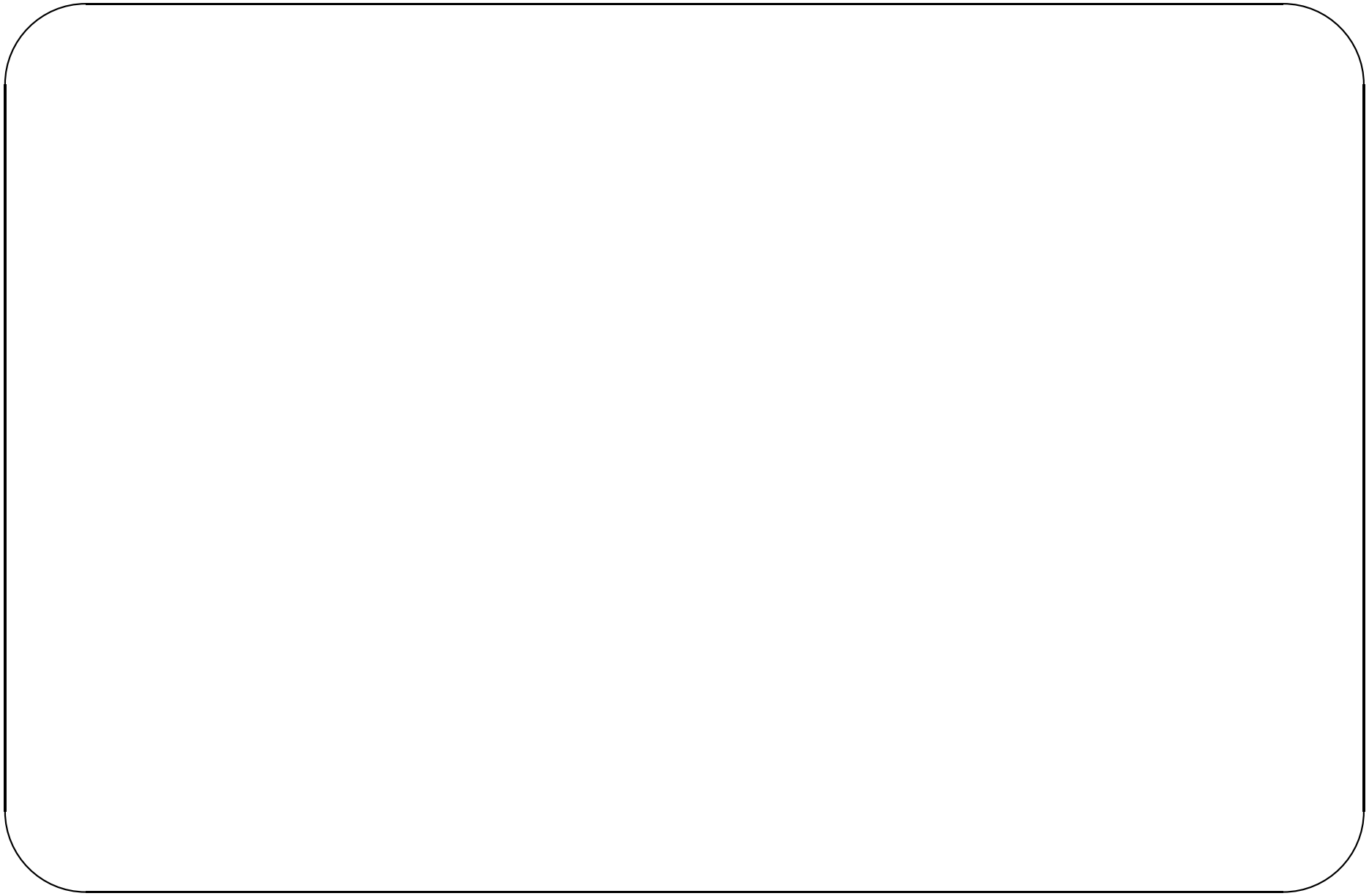
Variance plot




Biplot
Atlanta
Baltimore
Washington_DC
Los_Angeles
Milk
Bread
San_Diego
Kansas_City
Cincinnati
Seattle
New_York
Burger
San_Francisco
Boston
Philadelphia
Chicago
Houston
onolulu
St._Louis
Pittsburgh
Dallas
Tomatoes
Minneapolis
Buffalo
Detroit
Cleveland
Milwaukee
Oranges
Comp.1
Comp.2


pcex
Scree Plots
Variances
2.5
2.0
1.5
1.0
0.5
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5


cumulative variance
0
1
2
3
4
5
index
0
1
2
3
4
5



Swiss Bank Notes

Six variables measured on 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes. The data is from "Multivariate Statistics A practical approach", by Bernhard Flury and Hans Riedwyl, Chapman and Hall, 1988, Tables 1.1 and 1.2 pp. 5-8. The columns correspond to the following 6 variables:

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.

Observations 1-100 are the genuine bank notes and the other 100 observations are the counterfeit bank notes.

Length x_1	Height (left) x_2	Height (right) x_3	Inner Frame (lower) x_4	Inner Frame (upper) x_5	Diagonal x_6
214.8	131.0	131.1	9.0	9.7	141.0
214.6	129.7	129.7	8.1	9.5	141.7
214.8	129.7	129.7	8.7	9.6	142.2
214.8	129.7	129.6	7.5	10.4	142.0
⋮	⋮	⋮	⋮	⋮	⋮
214.9	130.3	130.5	11.6	10.6	139.8
215.0	130.4	130.3	9.9	12.1	139.6
215.1	130.3	129.9	10.3	11.5	139.7
214.8	130.3	130.4	10.6	11.1	140.0
214.7	130.7	130.8	11.2	11.2	139.4
214.3	129.9	129.9	10.2	11.5	139.6

The variances and covariances of the variables are:

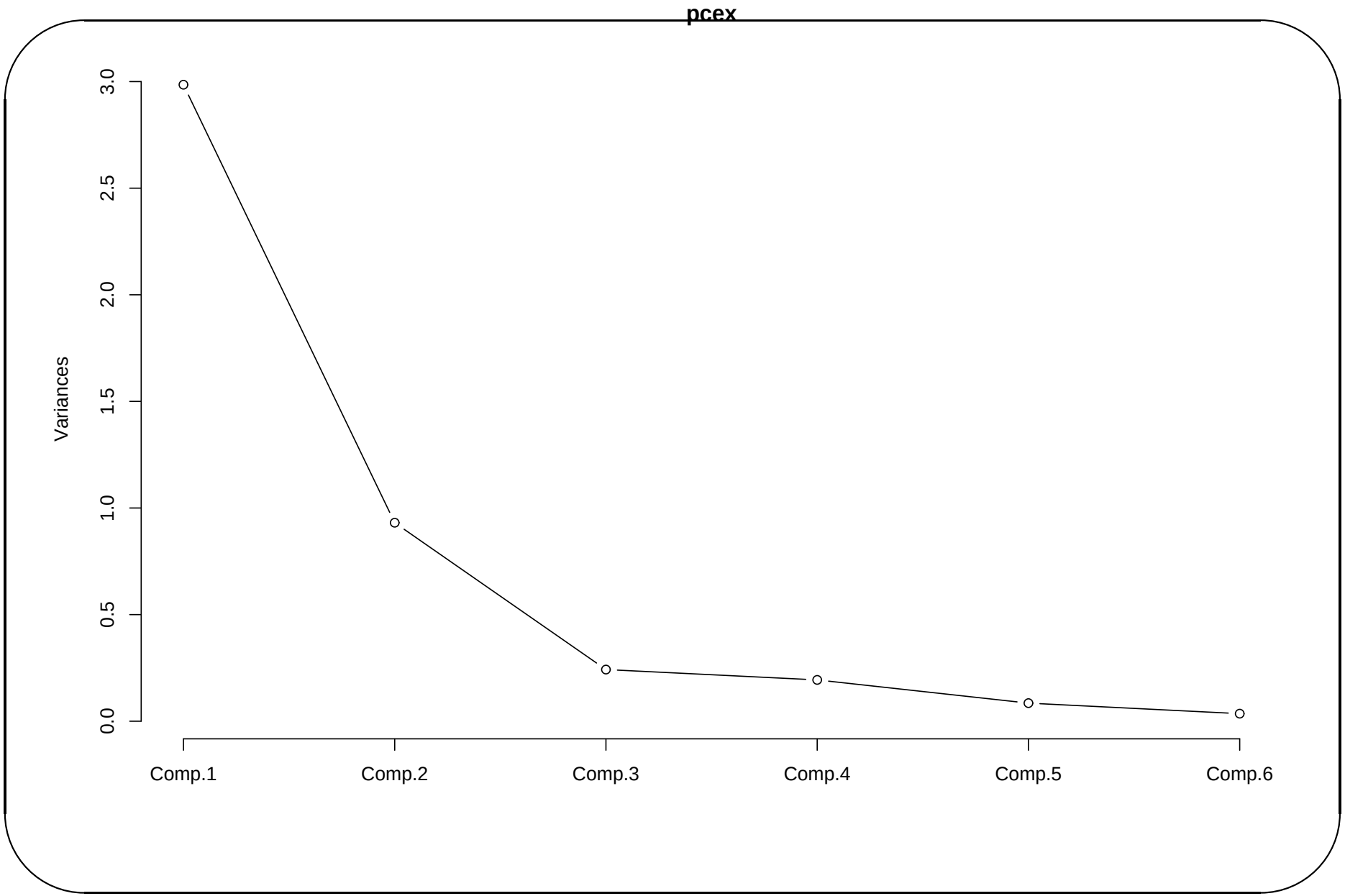
	Length	Left	Right	Bottom	Top	Diag
Length	0.14	0.03	0.02	-0.1	-0.02	0.08
Left	0.03	0.13	0.11	0.2	0.11	-0.21
Right	0.02	0.11	0.16	0.3	0.13	-0.24
Bottom	-0.10	0.22	0.28	2.1	0.16	-1.04
Top	-0.02	0.11	0.13	0.2	0.64	-0.55
Diag	0.08	-0.21	-0.24	-1.0	-0.55	1.33

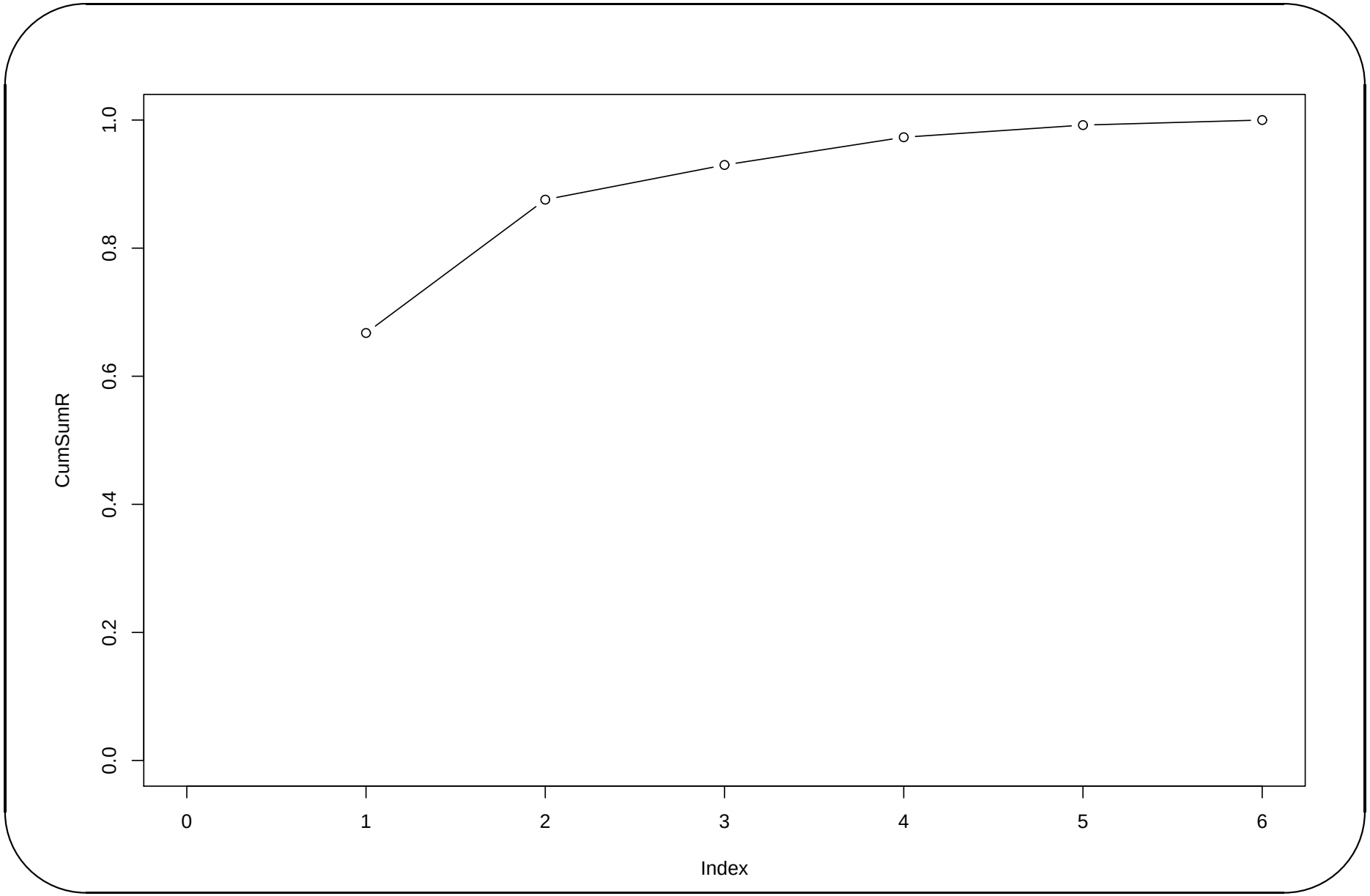
Importance of components:

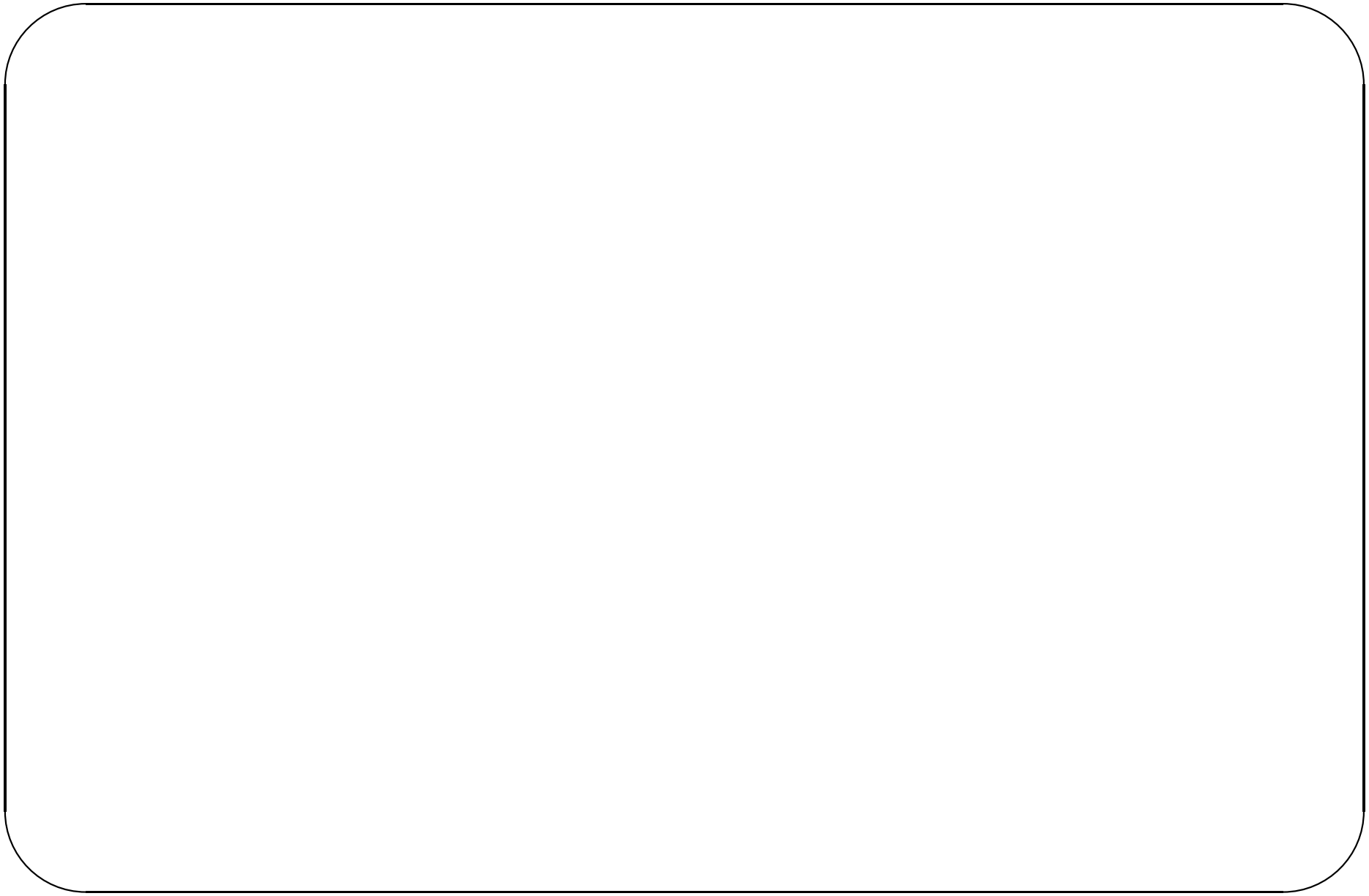
	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Stand. dev.	1.73	0.96	0.49	0.44	0.29	0.19
Proportion of Var	0.67	0.21	0.05	0.04	0.02	0.01
Cumulative Prop.	0.67	0.88	0.93	0.97	0.99	1.00

Loadings:

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Length			-0.33	0.56	0.75	
Left	0.11		-0.26	0.46	-0.35	-0.77
Right	0.14		-0.35	0.42	-0.54	0.63
Bottom	0.77	-0.56	-0.22	-0.19		
Top	0.20	0.66	-0.56	-0.45	0.10	
Diagonal	-0.58	-0.49	-0.59	-0.26		


pcex
Variances
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6


CumSumR
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
Index



Recall that the main idea of PC transformations is to find the most informative projections that maximize variances. The most information is given by the first eigenvector. In particular:

$$y_1 = -0.0x_1 + 0.11x_2 + 0.14x_3 + 0.77x_4 + 0.20x_5 - 0.58x_6$$
$$y_2 = -0.56x_4 + 0.66x_5 - 0.49x_6$$

Hence, the first PC is essentially the difference between the bottom frame variable and the diagonal. The second PC is best described by the difference between the top frame variable and the sum of bottom frame and diagonal variables.

Introduction to Factor Analysis

- Factor analysis is a generic name given to a class of a multivariate statistical methods whose primary purpose is to define the underlying structure in a data matrix.
- It addresses the problem of analyzing the structure of the interrelationships (correlations) among a large number of variables by defining a set of common underlying dimensions known as factors.
- With factor analysis the researcher can first identify the separate dimensions of the structure and then determine the extend to which each variable is explained by each

dimension. Once these dimensions and the explanation of each variable are determined, the two primary uses for FA -summarization and data reduction- can be achieved.

- In summarizing the data FA derived underlying dimensions, that when interpreted and understood, describe the data in a much smaller number of concepts than the original individual variables.
- Data reduction can be achieved by calculating scores for each underlying dimension and substituting them for the original variables.

- FA differs from the dependence techniques (e.g. multiple regression) in which one or more variables are explicitely considered the criterion or dependent variable and all others are the predictor or independent variables.
- FA is an interdependence technique in which all variables are simultaneously considered each related to all others. The variates (factors) are formed to maximize their explanation of the entire variable set, not to predict a dependent variable(s).
- Factor analytic techniques can achieve their purpose from either an explanatory or confirmatory prospective.

Explanatory implies the search for structure among a set of variables or as a data reduction method. In this case there is no a priori constraints on the estimation of components or the number of components to be extracted. This is the most appropriate FA in many cases.

In the confirmatory case the researcher may wish to test hypothesis involving issues such as which variables should be grouped together on a factor or the precise number of factors. In this case the FA can be use to asses the degree to which the data meet the expected structure.

Example

Assume that through a qualitative research a retail firm has identified 80 different characteristics of retail stores, and their service that consumers have mentioned as affecting their choice among stores. The retailer wants to understand how consumers make decisions but feels that it cannot evaluate 80 separate characteristics, or develop action plans for this many variables because they are too specific. Instead, it would like to know if the consumers think in more general evaluative dimensions rather than in just specific items. To identify these dimensions, the retailer could commission a survey asking for consumer evaluations on each of these specific items.

Factor Analysis then would be used to identify the underlying evaluating dimensions. Specific items that correlate highly are assumed to be a member of that broader dimension. These dimensions become composites of specific variables, which in turn allow the dimensions to be interpreted and described.

E.g. product prices, product quality, product assortment, store personnel, service and store atmosphere might be identified by FA as the most important dimensions. Each of these dimensions contains specific items that are a facet of the broader evaluative dimensions. From these findings the retailer may then use the dimensions (factors) to define broad areas for planning and action.

Objectives of Factor Analysis

The general purpose of factor analytic techniques is to find a way to condense (summarized) the information contained in a number of original variables into a smaller set of new, composite dimensions or factors, with a minimum loss of information.

More specifically, the factor analytic techniques can satisfy either two objectives: (1) identifying structure through data summarization, or (2) data reduction.

1. FA can identify the structure of relationships among either variables or observations by examining either the correlations between the variables or the correlations between the observations. For example suppose we have data on 100 respondents in terms of 10 characteristics. If the objective of the research were to summarize the characteristics, the FA will be applied to a correlation matrix of the variables.

   This is the most common type of FA and is referred to as **R factor analysis**.

R FA analyzes a set of variables to identify the dimensions that are not easily observed (latent).

The FA which is applied to the correlation matrix of the individual observations based on their characteristics is known as **Q factor analysis**. This is not a very commonly used FA. Instead, *cluster analysis* to group individual observations is used.

2. FA can also

- identify representative variables from a much larger set of variables for use in subsequent multivariate analyses, or

- create an entirely new set of variables, much smaller in number, to partially or completely replace the original set of variables for inclusion is subsequent techniques.

  In both instances, the purpose is to retain the nature and character of the original variables, but reduce their number to simplify the subsequent multivariate analysis.

  The researcher is always looking for the most parsimonious set of variables to include in the analysis.

Example

Consider a simple example of factor analysis which considers the following nine store image elements:

V_1	Price level	V_6	Assortment depth
V_2	Store personnel	V_7	Assortment width
V_3	Return policy	V_8	In-store service
V_4	Product availability	V_9	Store atmosphere
V_5	Product Quality		

The correlation matrix of these variables is given by:

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
V_1	1.0								
V_2	0.427	1.0							
V_3	0.302	0.771	1.0						
V_4	0.470	0.497	0.427	1.0					
V_5	0.765	0.406	0.307	0.472	1.0				
V_6	0.281	0.445	0.423	0.713	0.325	1.0			
V_7	0.354	0.490	0.471	0.719	0.378	0.724	1.0		
V_8	0.242	0.719	0.733	0.428	0.240	0.311	0.435	1.0	
V_9	0.372	0.737	0.774	0.479	0.326	0.429	0.466	0.710	1.0

Question

Are all these elements separate in their evaluative properties or do they group into some more general areas of evaluation?

Visual inspection of the original correlation does not reveal any specific pattern. There are scattered high correlations, but variable grouping is not apparent. The application of factor analysis results in the grouping of variables as reflected in the table below:

	V_3	V_8	V_9	V_2	V_6	V_7	V_4	V_1	V_5
V_3	1.0								
V_8	0.733	1.0							
V_9	0.774	0.710	1.0						
V_2	0.741	0.719	0.787	1.0					
V_6	0.423	0.311	0.429	0.445	1.0				
V_7	0.471	0.435	0.468	0.490	0.724	1.0			
V_4	0.427	0.428	0.479	0.497	0.713	0.719	1.0		
V_1	0.302	0.242	0.372	0.427	0.281	0.354	0.470	1.0	
V_5	0.307	0.240	0.326	0.406	0.325	0.378	0.472	0.765	1.0

	V_3	V_8	V_9	V_2	V_6	V_7	V_4	V_1	V_5
V_3	1.0								
V_8	0.733	1.0							
V_9	0.774	0.710	1.0						
V_2	0.741	0.719	0.787	1.0					
V_6					1.0				
V_7					0.724	1.0			
V_4					0.713	0.719	1.0		
V_1								1.0	
V_5								0.765	1.0

- The first four variables all related to the in-store experience of shoppers are grouped together.

- Three variables describing the product assortment and availability are grouped together.

- Finally, product quality and price levels are grouped.

  Each group represents a set of highly interrelated variables that may reflect a more general evaluative dimension. These groups might be labeled as *in-store experience*, *product offerings* and *value*.

Factor & Principal Components Analysis

- Compute the eigenvalues and eigenvectors of the variance-covariance (or correlation) matrix. The eigenvalues correspond to the variance of the factors.
- Calculate the *Proportion of Variance* of each factor. A big proportion of Variance implies a significant factor.
- Calculate the *Cumulative Proportion of Variance* and find how many factors to keep.

- Use a screeplot (eigenvalues vs Cumulative Proportion of Variance) to visualize the cut point.
- Calculate the eigenvectors to derive the combinations of the various variables used in each factor (loadings).
- If a correlation matrix is used then derive the loadings through a transformation.

Example

- The correlation matrix R:

```
1.0   0.427 0.302 0.470 0.765 0.281 0.354 0.242 0.372
0.427 1.0   0.771 0.497 0.406 0.445 0.490 0.719 0.737
0.302 0.771 1.0   0.427 0.307 0.423 0.471 0.733 0.774
0.470 0.497 0.427 1.0   0.472 0.713 0.719 0.428 0.479
0.765 0.406 0.307 0.472 1.0   0.325 0.378 0.240 0.326
0.281 0.445 0.423 0.713 0.325 1.0   0.724 0.311 0.429
0.354 0.490 0.471 0.719 0.378 0.724 1.0   0.435 0.466
0.242 0.719 0.733 0.428 0.240 0.311 0.435 1.0   0.710
0.372 0.737 0.774 0.479 0.326 0.429 0.466 0.710 1.0
```

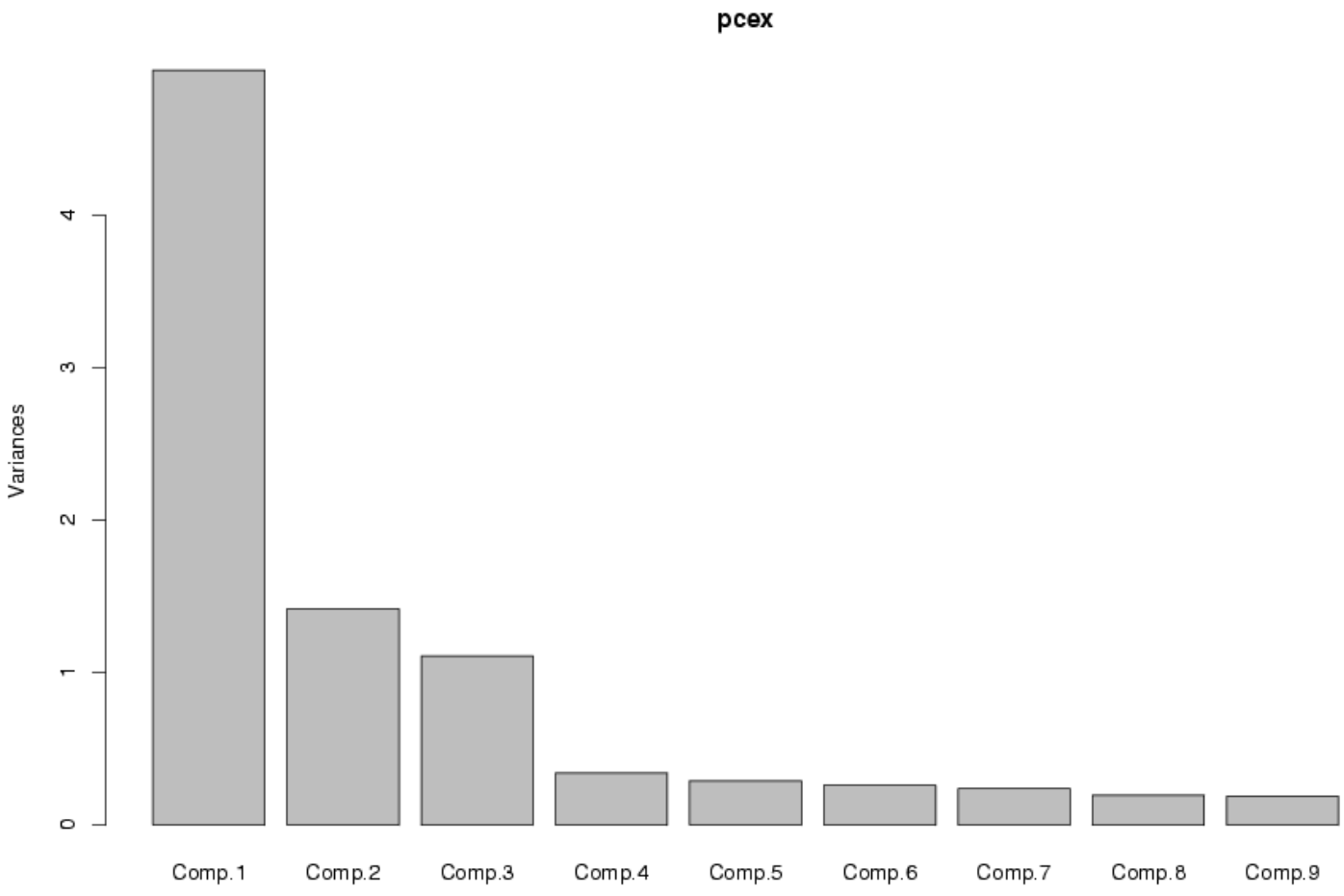
- The summary of applying a PCA to R:

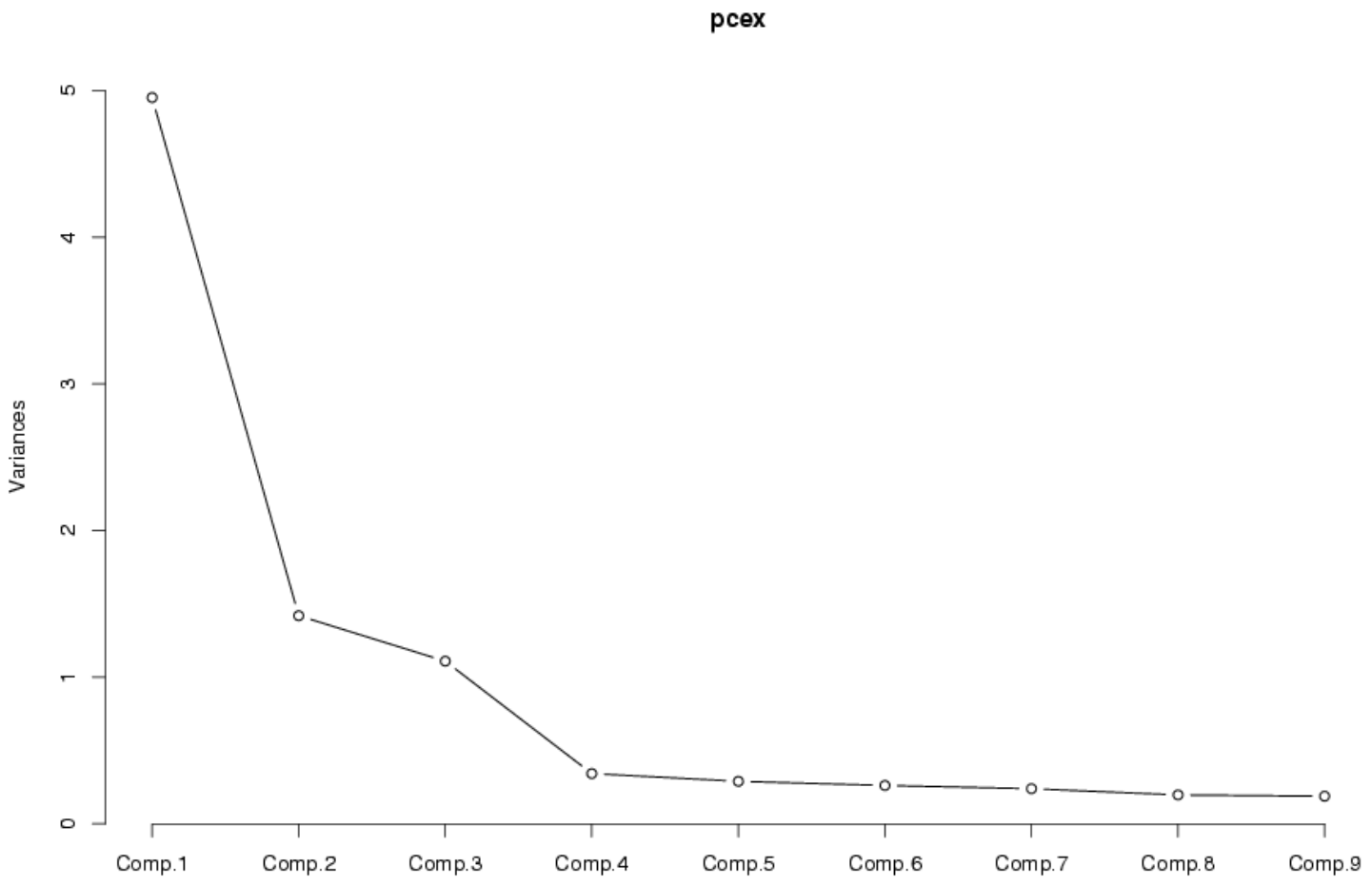
```
Importance of components:
Component   1    2    3     4    5   6    7    8    9
SD         2.22 1.19 1.05 0.59 0.54 0.51 0.49 0.44 0.43
Propor Var 0.55 0.16 0.12 0.04 0.03 0.03 0.03 0.02 0.02
Cum Propor 0.55 0.71 0.83 0.87 0.90 0.93 0.96 0.98 1.00
```

- The loadings of the factors (eigenvectors):

```
Loadings:
Com.1  Com.2 Com.3 Com.4 Com.5 Com.6 Com.7 Com.8 Com.9
-0.27 -0.44  0.47  0.01  0.21 -0.24 -0.42  0.14  0.46
-0.38  0.24  0.16  0.15 -0.16  0.47 -0.60  0.02 -0.39
-0.36  0.36  0.10  0.31 -0.24 -0.06  0.18 -0.59  0.45
-0.35 -0.27 -0.30 -0.36  0.53  0.26  0.04 -0.47 -0.08
-0.27 -0.47  0.42  0.02 -0.33  0.18  0.54  0.00 -0.31
-0.31 -0.20 -0.52  0.47  0.01  0.27  0.14  0.45  0.28
-0.34 -0.18 -0.42 -0.24 -0.50 -0.54 -0.21 -0.04 -0.17
-0.33  0.41  0.10 -0.63 -0.06  0.15  0.20  0.42  0.27
-0.37  0.30  0.12  0.27  0.48 -0.48  0.21  0.18 -0.39
```

- The screeplot:




pcex
Variances
0
1
2
3
4
5
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7
Comp.8
Comp.9

- Factor Analysis of the correlation matrix R (4 factors)

```
factanal(covmat = R, factors = 4)
Loadings:
      Factor1 Factor2 Factor3 Factor4
 [1,]  0.172   0.163   0.934
 [2,]  0.770   0.285   0.266
 [3,]  0.872   0.248   0.122
 [4,]  0.247   0.752   0.325   0.129
 [5,]  0.170   0.243   0.745
 [6,]  0.227   0.838   0.114
 [7,]  0.293   0.764   0.190
 [8,]  0.816   0.186           0.506
 [9,]  0.782   0.273   0.203
```

	Factor1	Factor2	Factor3	Factor4
SS loadings	2.890	2.189	1.713	0.291
Proportion Var	0.321	0.243	0.190	0.032
Cumulative Var	0.321	0.564	0.755	0.787

- Factor Analysis of the correlation matrix R (3 factors)

```
factanal(covmat = R, factors = 3)
Loadings:
     Factor1 Factor2 Factor3
 [1,] 0.172   0.160   0.955
 [2,] 0.786   0.275   0.260
 [3,] 0.852   0.242   0.123
 [4,] 0.270   0.748   0.319
 [5,] 0.173   0.249   0.728
 [6,] 0.227   0.824   0.116
 [7,] 0.304   0.769   0.187
 [8,] 0.806   0.201
 [9,] 0.795   0.261   0.201
```

	Factor1	Factor2	Factor3
SS loadings	2.902	2.160	1.722
Proportion Var	0.322	0.240	0.191
Cumulative Var	0.322	0.562	0.754

- Factor Analysis of the correlation matrix R (2 factors)

```
Loadings:
     Factor1 Factor2
 [1,] 0.260   0.443
 [2,] 0.794   0.354
 [3,] 0.851   0.271
 [4,] 0.258   0.839
 [5,] 0.235   0.469
 [6,] 0.214   0.785
 [7,] 0.290   0.780
 [8,] 0.792   0.233
 [9,] 0.800   0.319

                Factor1 Factor2
SS loadings       2.940   2.700
```

```
                Proportion Var   0.327   0.300
                Cumulative Var   0.327   0.627
```

- Factor Analysis of the correlation matrix R (1 factor)

```
Loadings:
      Factor1                              Factor1
 [1,] 0.466           SS loadings          4.369
 [2,] 0.871           Proportion Var       0.485
 [3,] 0.858
 [4,] 0.627
 [5,] 0.455
 [6,] 0.563
 [7,] 0.625
 [8,] 0.792
 [9,] 0.850
```