—WORKSHOP—

Applied Classical and Modern Multivariate Statistical Analysis

Module 5: Principal Component Analysis

Weixing Song, Juan Du

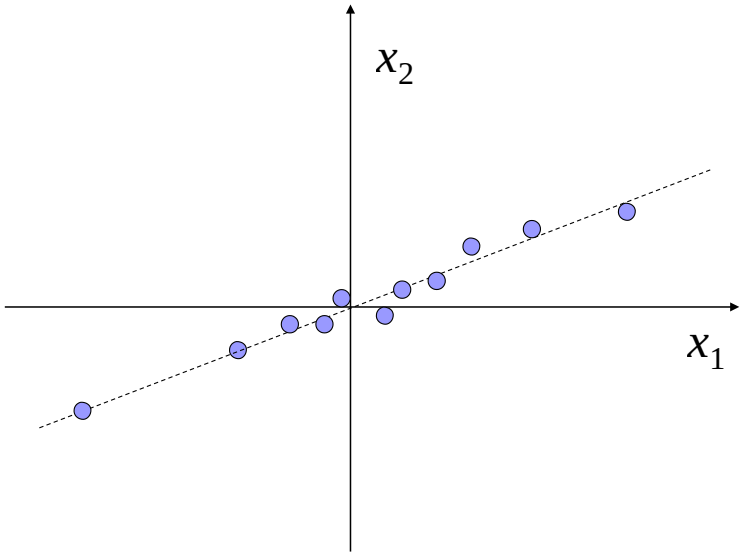
Department of Statistics
Kansas State University

Outline

Outline

What are principal components?

Principal components of $X_1, X_2, \ldots, X_p$ are some special linear combinations of these variables whose variability is close to the variability of $X_1, X_2, \ldots, X_p$.



Why principal components?

Data reduction; Interpretation.

Note: "Analyses of principal components are more of a means to an end rather than an end in themselves."

Outline

Let $(X_1, X_2, \ldots, X_p)'$ be a p-dimensional random vector. Its covariance and the correlation matrices are Σ and $\boldsymbol{\rho}$, respectively.

Principal Components

The principal components of $\mathbf{X} = (X_1, \ldots, X_p)'$ are the following p linear combinations of $X_1, \ldots, X_p$

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

such that

- $Y_1 = \mathbf{a}_1'\mathbf{X}$ maximizes $\mathrm{Var}(\mathbf{a}_1'\mathbf{X})$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$;
- $Y_2 = \mathbf{a}_2'\mathbf{X}$ maximizes $\mathrm{Var}(\mathbf{a}_2'\mathbf{X})$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $\mathrm{Cov}(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_1'\mathbf{X}) = 0$;
- $\cdots\cdots$;
- $Y_p = \mathbf{a}_p'\mathbf{X}$ maximizes $\mathrm{Var}(\mathbf{a}_p'\mathbf{X})$ subject to $\mathbf{a}_p'\mathbf{a}_p = 1$ and $\mathrm{Cov}(\mathbf{a}_p'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = 0$ for $j = 1, 2, \ldots, p-1$.

Population Principal Compoents

Let Σ be the covariance matrix of $\mathbf{X} = (X_1, \ldots, X_p)'$. Its eigenvalue-eigenvector pairs are $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the i-th principal component is given by

$$Y_i = \mathbf{e}_i \mathbf{X} = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p, \quad i = 1, 2, \ldots, p.$$

With these choices,

$$\mathrm{Var}(Y_i) = \mathbf{e}_i' \Sigma \mathbf{e}_i = \lambda_i, \quad \mathrm{Cov}(Y_i, Y_k) = \mathbf{e}_i' \Sigma \mathbf{e}_k = 0,$$

$i, k = 1, 2, \ldots, p, i \neq k$. If some λ_i's are equal, the choices of the corresponding coefficient vectors $\mathbf{e}_i$, and hence Y_i are not unique.

Correlation Between PCA and Individual Variables

Let $Y_1 = \mathbf{e}_1' \mathbf{X}, \cdots, Y_p = \mathbf{e}_p' \mathbf{X}$ be the principal components obtained from the covariance matrix Σ, then

$$\rho_{i,k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \quad i, k = 1, 2, \ldots, p.$$

A Relationship between Variances and Eigenvalues

Let $\mathbf{X} = (X_1, \ldots, X_p)'$ have covariance matrix Σ, with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = \mathbf{e}_1'\mathbf{X}$, ..., $Y_p = \mathbf{e}_p'\mathbf{X}$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i).$$

Based on the above result, we have

$$\text{Proportion of total population variance due to the } k\text{-th principal component} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Example 5.1 Suppose the random variables X_1, X_2, X_3 have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Find the principal components, and the correlation coefficients between each variable and principal component.

See R-code

An Interpretation of Principal Components

Suppose $X \sim N_p(\boldsymbol{\mu}, \Sigma)$. The contour of its density function is given by the $\boldsymbol{\mu}$-centered ellipsoids

$$(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

which have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, where $(\lambda_i, \mathbf{e}_i)$ are the eigenvalue-eigenvector pairs of Σ, $i = 1, 2, \ldots, p$.

WLOG, assume that $\mu = 0$.

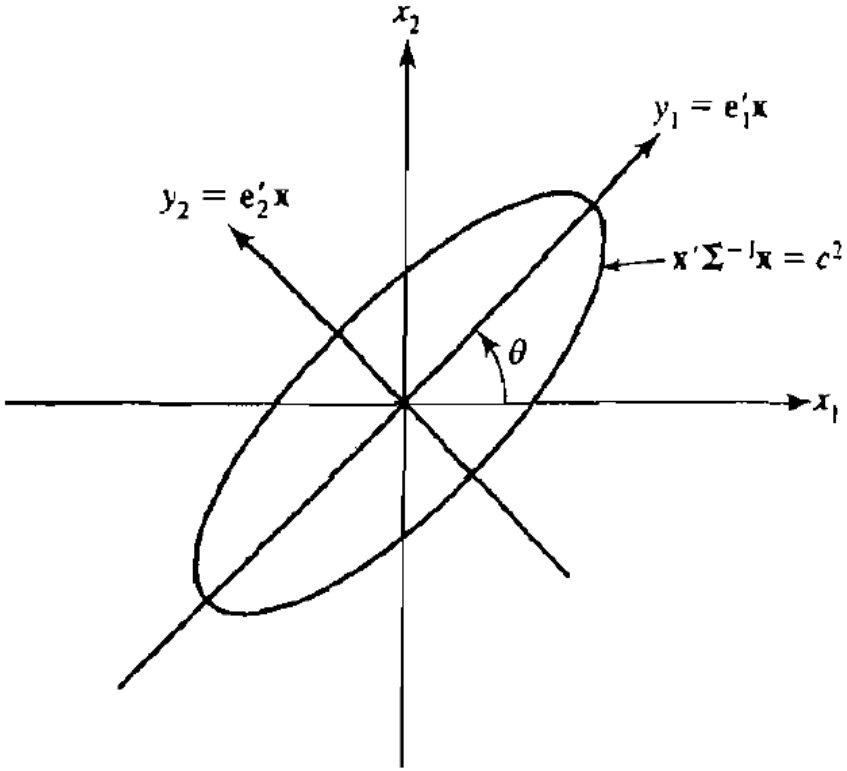
Note that the ellipsoid can be also written as

$$\frac{1}{\lambda_1}(\mathbf{e}_1'\mathbf{x})^2 + \frac{1}{\lambda_2}(\mathbf{e}_2'\mathbf{x})^2 + \cdots + \frac{1}{\lambda_p}(\mathbf{e}_p'\mathbf{x})^2 = c^2.$$

Recall that $\mathbf{e}_1'\mathbf{x}, \mathbf{e}_2'\mathbf{x}, \ldots, \mathbf{e}_p'\mathbf{x}$ are the PCs of $\mathbf{x}$. It is easy to see that the PCs lie in the following ellipsoid

$$\frac{1}{\lambda_1}y_1^2 + \frac{1}{\lambda_2}y_2^2 + \cdots + \frac{1}{\lambda_p}y_p^2 = c^2.$$

That is, **the principal components lie in the directions of the axes of a constant density ellipsoid.**



$\boldsymbol{\mu} = \mathbf{0}$
$\rho = .75$

Figure 5.1 The constant density ellipse $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector $\mathbf{X}$ having mean $\mathbf{0}$.

Principal Components for Standardized Variables

If $\mathbf{X}$ has mean $\boldsymbol{\mu}$ and covariance matrix Σ, then the standardized vector $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ has mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\rho}$ (which, indeed, is the correlation matrix of $\mathbf{X}$), where $\mathbf{V} = \text{diag}(\sigma_{11}, \sigma_{22}, \cdots, \sigma_{pp})$.

PCAs Based on Standardized Variables

Suppose $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$ are the eigenvalue-eigenvector pairs for $\boldsymbol{\rho}$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. The i-th principal component of the standardized variables $\mathbf{Z}' = [Z_1, \ldots, Z_p]$ with $\text{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$, is given by

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

Moreover,

$$\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p$$

and

$$\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i}, \quad i, k = 1, 2, \ldots, p.$$

Outline

The sample principal components are defined as those linear combinations which have maximum sample variance with the restriction that the coefficient vectors are of unit length.

To be specific,

- First Sample PC: the linear combination $\mathbf{a}_1'\mathbf{x}_j$ that maximizes the sample variance $\mathbf{a}_1'\mathbf{x}_j$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$;
- Second Sample PC: the linear combination $\mathbf{a}_2'\mathbf{x}_j$ that maximizes the sample variance $\mathbf{a}_2'\mathbf{x}_j$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and zero sample covariance for the pairs $(\mathbf{a}_1'\mathbf{x}_j, \mathbf{a}_2'\mathbf{x}_j)$;
- ......
- p-th Sample PC: the linear combination $\mathbf{a}_p'\mathbf{x}_j$ that maximizes the sample variance $\mathbf{a}_p'\mathbf{x}_j$ subject to $\mathbf{a}_p'\mathbf{a}_p = 1$ and zero sample covariance for the pairs $(\mathbf{a}_k'\mathbf{x}_j, \mathbf{a}_p'\mathbf{x}_j)$, $k < p$.

Recall that the n-values $\mathbf{a}_1'\mathbf{x}_j$ of linear combination $\mathbf{a}_1'\mathbf{x}$ have sample variance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$, and the pairs of values $(\mathbf{a}_1'\mathbf{x}_j, \mathbf{a}_2'\mathbf{x}_j)$ have sample covariance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$.

Sample PCs

Let $\mathbf{S}$ be the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, the i-th sample principal component is given by

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p, \quad i = 1, 2, \ldots, p$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ and $\mathbf{x}$ is any observation on the variables $X_1, \ldots, X_p$. Also

$$\text{Sample variance}(\hat{y}_k) = \hat{\lambda}_k, \ldots k = 1, 2, \ldots, p$$

$$\text{Sample covariance}(\hat{y}_i, \hat{y}_k) = 0, i \neq k.$$

In addition

$$\text{Total sample variance} = \sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i$$

and

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p.$$

Remarks:

- The observations $\mathbf{x}_j$ are often "centered" by substracting $\bar{\mathbf{x}}$. This has no effect on the sample covariance matrix $\mathbf{S}$ and gives the i-th principal component
$$\hat{y}_i = \hat{\mathbf{e}}_i(\mathbf{x} - \bar{\mathbf{x}}), \quad i = 1, 2, \ldots, p$$
for any observation vector $\mathbf{x}$.
- The values of the i-th principal component at each $\mathbf{x}_j$ are
$$\hat{y}_{ji} = \hat{\mathbf{e}}_i'(\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1, 2, \ldots, n.$$
note that $\bar{\hat{y}}_i = 0$.
- The interpretation of sample PCs.
  - Realizations of population principal components.
  - The sample principal components can be viewed as the result of translating the origin of the original coordinate system to $\bar{\mathbf{x}}$ and then rotating the coordinate axes until they pass through the scatter in the directions of maximum variances.

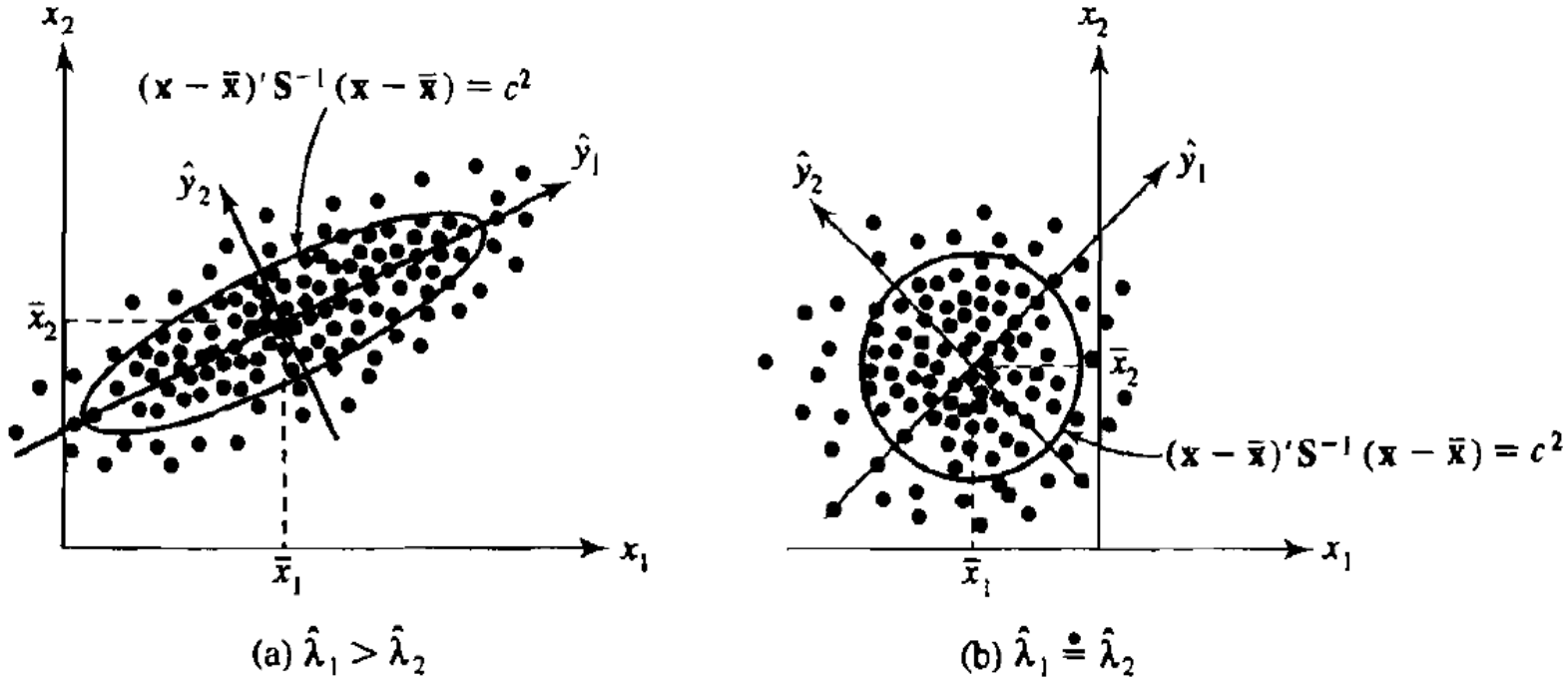


Figure 8.4 Sample principal components and ellipses of constant distance.

Example 5.2 (Summarizing sample variability with two sample PCs) (See R-code and output)

A census provided information. bv tract, on five socioeconomic variables for the Madison, Wisconsin, area.

$$\bar{\mathbf{x}}' = \underset{\substack{\text{total}\\ \text{population}\\ \text{(thousands)}}}{[4.47,} \quad \underset{\substack{\text{professional}\\ \text{degree}\\ \text{(percent)}}}{3.96,} \quad \underset{\substack{\text{employed}\\ \text{age over 16}\\ \text{(percent)}}}{71.42,} \quad \underset{\substack{\text{government}\\ \text{employment}\\ \text{(percent)}}}{26.91,} \quad \underset{\substack{\text{median}\\ \text{home value}\\ \text{(\$100,000)}}}{1.64]}$$

and

$$\mathbf{S} = \begin{bmatrix} 3.397 & -1.102 & 4.306 & -2.078 & 0.027 \\ -1.102 & 9.673 & -1.513 & 10.953 & 1.203 \\ 4.306 & -1.513 & 55.626 & -28.937 & -0.044 \\ -2.078 & 10.953 & -28.937 & 89.067 & 0.957 \\ 0.027 & 1.203 & -0.044 & 0.957 & 0.319 \end{bmatrix}$$

Can the sample variation be summarized by one or two principal components?

Example 5.2 (Continued)
(Johnson and Wichern(2007))

Coefficients for the Principal Components
(Correlation Coefficients in Parentheses)

Variable	$\hat{\mathbf{e}}_1\,(r_{\hat{y}_1,x_k})$	$\hat{\mathbf{e}}_2\,(r_{\hat{y}_2,x_k})$	$\hat{\mathbf{e}}_3$	$\hat{\mathbf{e}}_4$	$\hat{\mathbf{e}}_5$
Total population	−0.039(−.22)	0.071(.24)	0.188	0.977	−0.058
Profession	0.105(.35)	0.130(.26)	−0.961	0.171	−0.139
Employment (%)	−0.492(−.68)	0.864(.73)	0.046	−0.091	0.005
Government employment (%)	0.863(.95)	0.480(.32)	0.153	−0.030	0.007
Medium home value	0.009(.16)	0.015(.17)	−0.125	0.082	0.989
Variance ($\hat{\lambda}_i$):	107.02	39.67	8.37	2.87	0.15
Cumulative percentage of total variance	67.7	92.8	98.1	99.9	1.000

1

PCs Based on Sample Correlation Matrix

Similar to the treatment in population principal component analysis, variables measured on different scales or on a common scale with widely differing ranges are often standardized.

Standardization is accomplished by constructing

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) = \begin{bmatrix} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1, 2, \ldots, n.$$

The sample covariance of $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ is the sample correlation $\mathbf{R}$.

Principal Components from $\mathbf{R}$

Suppose $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are standardized observations with covariance matrix $\mathbf{R}$, the i-th sample principal component is

$$\hat{y}_i = \hat{\mathbf{e}}_i \mathbf{z} = \hat{\mathbf{e}}_{i1} z_1 + \cdots + \hat{\mathbf{e}}_{ip} z_p, \quad i = 1, 2, \ldots, p$$

where $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ is the i-th eigenvalue-eigenvector pair of $\mathbf{R}$ with $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Also

$$\text{Smaple variance}(\hat{y}_i) = \hat{\lambda}_i, \quad i = 1, 2, \ldots, p,$$
$$\text{Sample covariance}(\hat{y}_i, \hat{y}_k) = 0, \quad i \neq k.$$

In addition,

$$\text{Total (standardized) sample variance} = \text{tr}(\mathbf{R}) = p = \hat{\lambda}_1 + \cdots + \hat{\lambda}_p$$

and

$$r_{\hat{y}_i, z_k} = \hat{e}_{ik} \sqrt{\hat{\lambda}_i}, \quad i, k = 1, 2, \ldots, p.$$

The Number of PCs

No definite answers to this question!

Factors to be considered:

- The amount of total sample variance explained;
- The relative sizes of the variances of the sample components;
- The subject-matter interpretations of the components.

A useful visual technique to determine an appropriate number of principal components is a scree plot. With the eigenvalues ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus i.

The number of components is taken to be the point at which the remaining eigenvalues are relative small and all about the same size.

Example 5.3 (This example will be revisited in the next section) The weekly rates of return for five stocks (JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, and ExxonMobil) listed on the New York Stock Exchange were determined for the period January 2004 through December 2005. The weekly rates of return are defined as (current Friday closing price - previous Friday closing price)/(previous Friday closing price), adjusted for stock splits and dividends. The observations in 103 successive weeks appear to be independently distributed, but the rates of return across stocks are correlated, since, as one expects, stocks tend to move together in response to general economic conditions.
Let x_1, x_2, x_3, x_4, x_5 denote the observed weekly rates of return for JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, and ExxonMobil, respectively. Find the PCs based on the sample correlation coefficient.