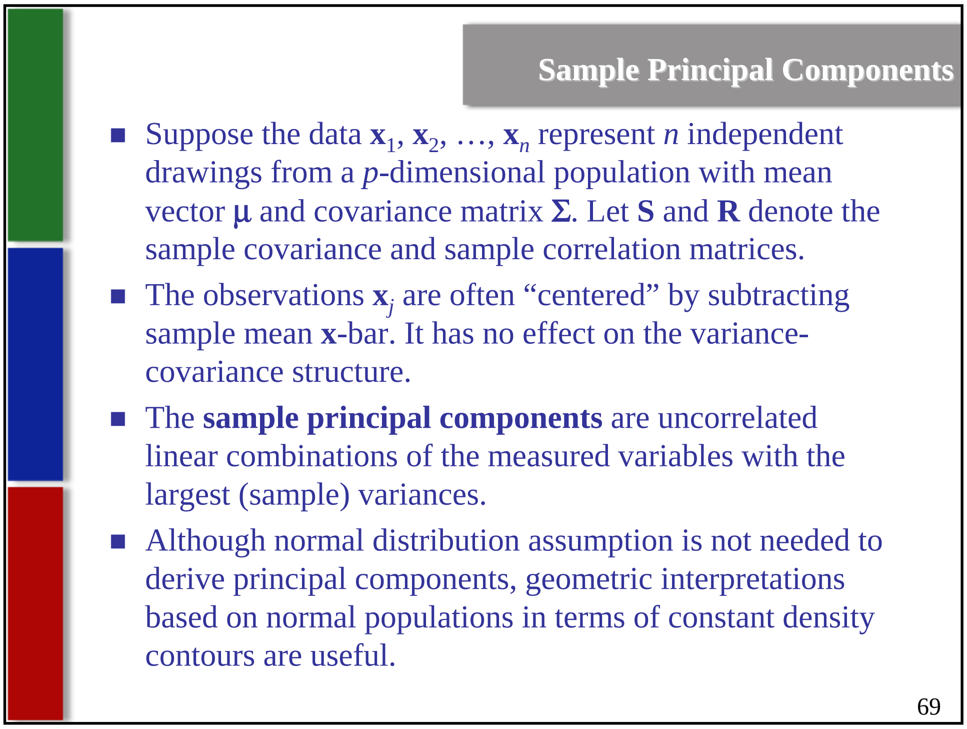
Principal Component Analysis

Introduction

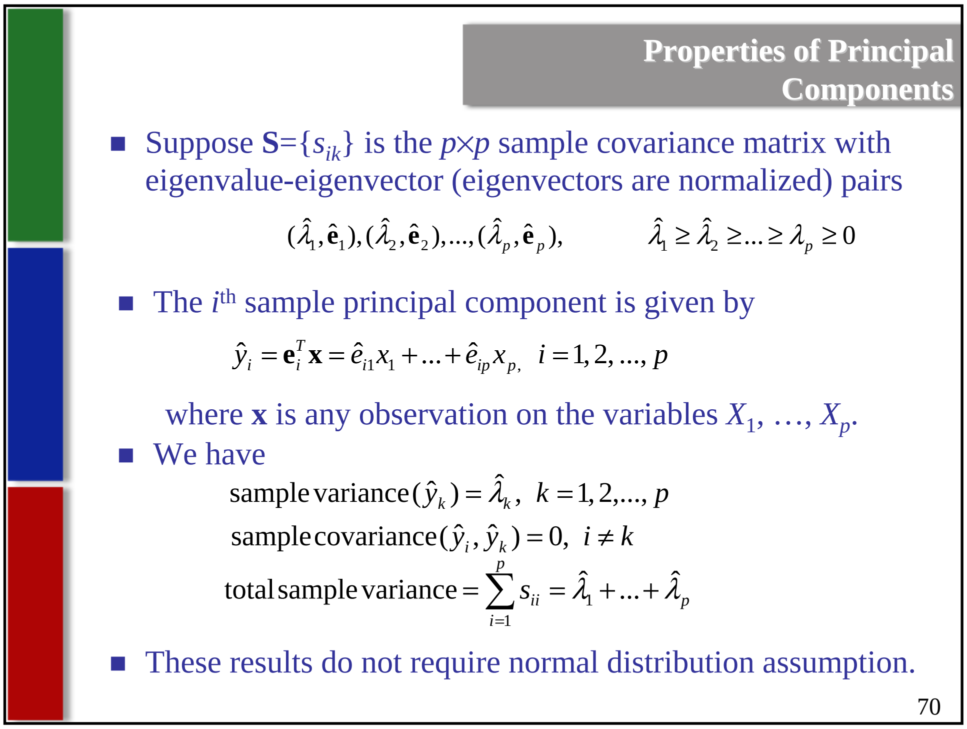
- A principal component analysis (PCA) is concerned with explaining the variance-covariance structure of a set of variables through a few *linear* combinations of these variables for (1) data reduction, and (2) interpretation.
- Principal components (PC) may be inputs to linear regression and cluster analysis.



Sample Principal Components

- Suppose the data $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$ represent n independent drawings from a p-dimensional population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let **S** and **R** denote the sample covariance and sample correlation matrices.
- The observations $\mathbf{x}_j$ are often "centered" by subtracting sample mean **x**-bar. It has no effect on the variance-covariance structure.
- The **sample principal components** are uncorrelated linear combinations of the measured variables with the largest (sample) variances.
- Although normal distribution assumption is not needed to derive principal components, geometric interpretations based on normal populations in terms of constant density contours are useful.

69



Properties of Principal Components

- Suppose $\mathbf{S}=\{s_{ik}\}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector (eigenvectors are normalized) pairs

$$(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p), \qquad \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$$

- The i^{th} sample principal component is given by

$$\hat{y}_i = \mathbf{e}_i^T \mathbf{x} = \hat{e}_{i1} x_1 + \ldots + \hat{e}_{ip} x_{p,} \quad i = 1, 2, \ldots, p$$

where **x** is any observation on the variables X_1, …, X_p.

- We have

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

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

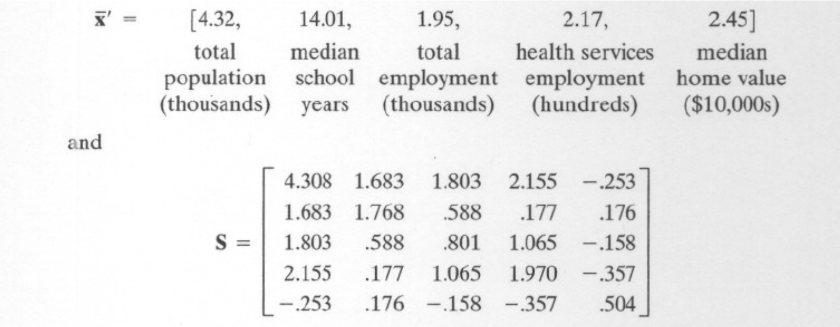
$$\text{total sample variance} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \ldots + \hat{\lambda}_p$$

- These results do not require normal distribution assumption.

70

Example 8.3

- A census provided information on five socioeconomic variables. The data are from 14 tracts.
- The summary statistics are:



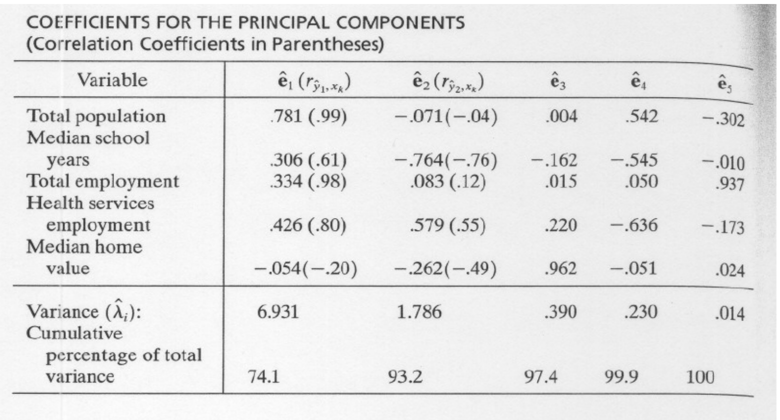
$$
\begin{array}{lccccc}
\bar{\mathbf{x}}' = & [4.32, & 14.01, & 1.95, & 2.17, & 2.45] \\
 & \text{total} & \text{median} & \text{total} & \text{health services} & \text{median} \\
 & \text{population} & \text{school} & \text{employment} & \text{employment} & \text{home value} \\
 & \text{(thousands)} & \text{years} & \text{(thousands)} & \text{(hundreds)} & \text{(\$10,000s)}
\end{array}
$$

and

$$
\mathbf{S} = \begin{bmatrix} 4.308 & 1.683 & 1.803 & 2.155 & -.253 \\ 1.683 & 1.768 & .588 & .177 & .176 \\ 1.803 & .588 & .801 & 1.065 & -.158 \\ 2.155 & .177 & 1.065 & 1.970 & -.357 \\ -.253 & .176 & -.158 & -.357 & .504 \end{bmatrix}
$$

Example 8.3 (Cont.)

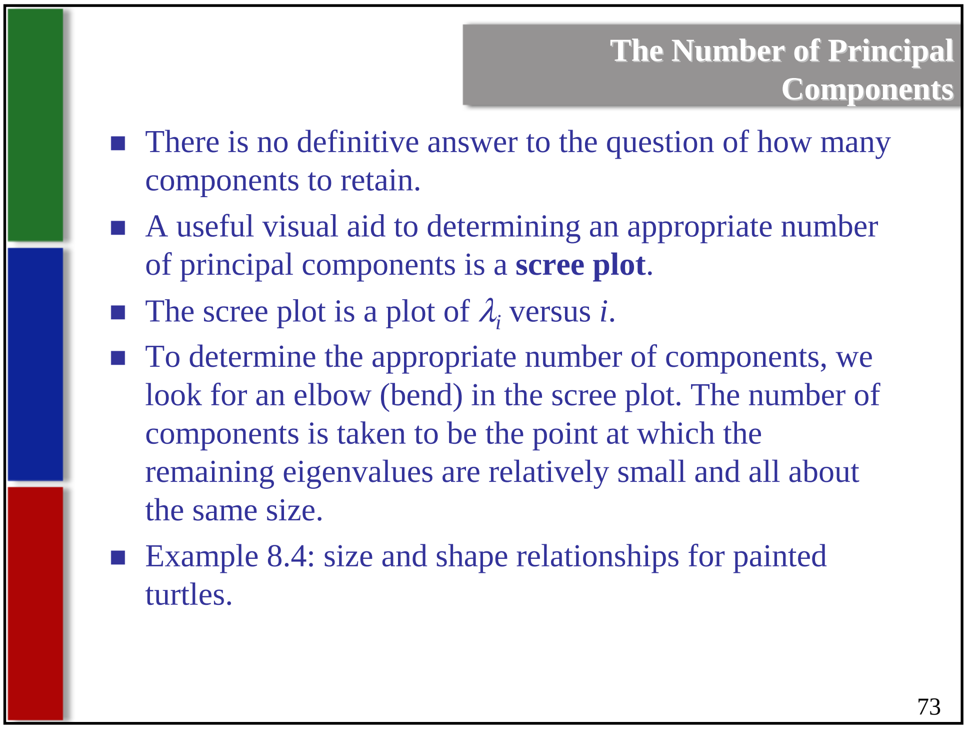
- The results of principal component analysis are:



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	.781 (.99)	−.071(−.04)	.004	.542	−.302
Median school years	.306 (.61)	−.764(−.76)	−.162	−.545	−.010
Total employment	.334 (.98)	.083 (.12)	.015	.050	.937
Health services employment	.426 (.80)	.579 (.55)	.220	−.636	−.173
Median home value	−.054(−.20)	−.262(−.49)	.962	−.051	.024
Variance ($\hat{\lambda}_i$):	6.931	1.786	.390	.230	.014
Cumulative percentage of total variance	74.1	93.2	97.4	99.9	100

- Sample variation is summarized very weel by two principal components. Reduction in the data from 14 observations on 5 variables to 14 observations on 3 PCs is reasonable



The Number of Principal Components

- There is no definitive answer to the question of how many components to retain.
- A useful visual aid to determining an appropriate number of principal components is a **scree plot**.
- The scree plot is a plot of λ_i versus i.
- To determine the appropriate number of components, we look for an elbow (bend) in the scree plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size.
- Example 8.4: size and shape relationships for painted turtles.