1. Aspects of Multivariate Analysis

1.1 Introduction

This course is considered with statistical methods designed to elicit information from the data sets with many different variables. Because the data include simultaneous measurements on many variables, this body of methodology is called **multivariate analysis**.

- The need to understand the relationships between many variables makes multivariate analysis an inherently difficult subject.

- Most of our emphasis will be on the analysis of measurements obtained without actively controlling or manipulating any of the variables on which the measurement are made.

- Many multivariate methods are based upon an underlying probability model known as the multivariate normal distribution.

- Multivariate analysis is a "mixed bag". It is difficult to establish a classification scheme for multivariate techniques that both widely accepted and indicates the appropriateness of the techniques.

The objectives of scientific investigations to which multivariate methods most naturally lend themselves include the following

- *Data reduction or structural simplification*
- *Sorting and grouping*
- *Investigation of the dependence among variables*
- *prediction*
- *Hypothesis construction and testing*

If the results disagree with informed opinion, do not admit a simple logical interpretation, and do not show up clearly in a graphical presentation, they are probably wrong. There is no magic about numerical methods, and many ways in which they can break down. They are a value aid to the interpretation of data, not sausage machines automatically transforming bodies of numbers into packet of scientific fact.

1.2 Application of Multivariate Techniques

Data reduction or simplification

- Using data on several variables related to cancer patient responses to radio-therapy, a simple measure of patient response to radiotherapy was constructed.
- Track records from many nations were used to develop an index of performance for both male and female athletes.
- Multispectral image data collected by a high -altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions.
- Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants.
- A matrix of tactic similarities was developed from aggregate data derived from professional mediators. From this matrix the number of dimensions by which professional mediator judge the tactics they use in resolving disputes was determined.

Sorting and grouping

- Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing(or planned) computer utilization.

- Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from nonalcoholics.

- Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from disease.

- The U.S. Internal Revenue Service uses data collected from tax returns to sort taxpayers into two groups: those that will be audited and those will not.

Investigation of the dependence among variables

- Data on several variables were used to identify factors that were responsible for client success in hiring external consultants.

- Measurement of variables related to innovation, on the one hand, and variables related to the business environment and business organization, on the other hand, were used to discover why some firms are product innovators and some firm are not.

- Measurements of pulp fiber characteristics and subsequent measurements of characteristics of the paper made from them are used to examine the relations between pulp fiber properties and the resulting paper properties. The goal is to determine those fiber that lead to higher quality paper.

- The association between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executive were used to assess the relation between risk-taking behavior and performance.

Prediction

- The association between test scores, and several high school performance variables, and several college performance variables were used to develop predictors of success in college.

- Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments.

- Measurement on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers.

- cDNA microarray experiment(gene expression data) are increasing used to study the molecular variation among cancer tumors. A reliable classification of tumor is essential for successful diagnosis and treatment of cancer.

Hypotheses testing

- Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends.

- Experimental data on several variables were used to see whether the nature of the instruction makes any difference in perceives risk, as quantified by test scores.

- Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories.

- Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innovation.

1.3 The Organization of Data

Array

Multivariate data arise whenever an investigator, seeking to understand a social or physical phenomenon, selects $p \geq 1$ of variables or characters to record. The values of these variables are all recorded for each distinct *item, individual* or *experimental unit*

$$x_{jk} = \text{measurement of the } k\text{th variable on the } j\text{th item}$$

Consequently, n measurements on p variables can be displayed as a rectangular array, called $\mathbf{X}$, of n rows and p columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Example 1.1 (A data array) A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided., among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding number on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales):	42	52	48	58
Variable 2 (number of books):	4	5	4	3

Descriptive Statistics

A large data set is bulky, and its very mass poses a serious obstacle to any attempt to visually extract pertinent information. Much of the information contained in the data can be assessed by calculating certain summary numbers, known as *descriptive statistics.*

- The arithmetic average or sample mean, is a descriptive statistics that provides a measure of location — that is, a "central value" for a set of numbers.

- The average of the squares of the distances of all of the number from mean provides a measure of the spread, or variation, in numbers.

- *Sample mean*

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{j1} \quad \text{or} \quad \bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{jk} \quad k = 1, 2, \ldots, p$$

- *Sample variance*

$$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2, \quad k = 1, 2, \ldots, p.$$

- *Sample standard deviation* $\sqrt{s_{kk}}$.

- *Sample covariance*

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

- *Sample correlation coefficient (or Pearson's product-moment correlation coefficient)*

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}}$$

1. The value of r must between -1 and $+1$ inclusive.

2. Here r measures the strength of the linear association.
   - $r = 0$: implies linear independent, lack of linear association between the components.
   - $r < 0$: implies a tendency for one value in the pair to be larger than its average when the other is smaller than its average.
   - $r > 0$: implies a tendency for one value of the pair to be large when the other value is large and also for both values to be small together.

3. The value of r_{ik} remain unchange if the measurements of ith variable are changed to $y_{ji} = ax_{ji} + b, j = 1, 2, \ldots, n$, and the value of the kth variable are changed to $y_{jk} = cx_{jk} + d, j = 1, 2, \ldots, n$, provide that the constants a and c have the same sign.

Example 1.2 (The arrays $\bar{x}, S_n$ and R for bivariate data) Consider the data introduced in Example 1.1.Each receipt yields a pair of measurements, total dollar sales, and number of books sold. Find the array $\bar{\mathbf{x}}, \mathbf{S}_n$ and $\mathbf{R}$.

Graphical Techniques

Variable 1	(x_1):	3	4	2	6	8	2	5
Variable 2	(x_2):	5	5.5	4	7	10	5	7.5

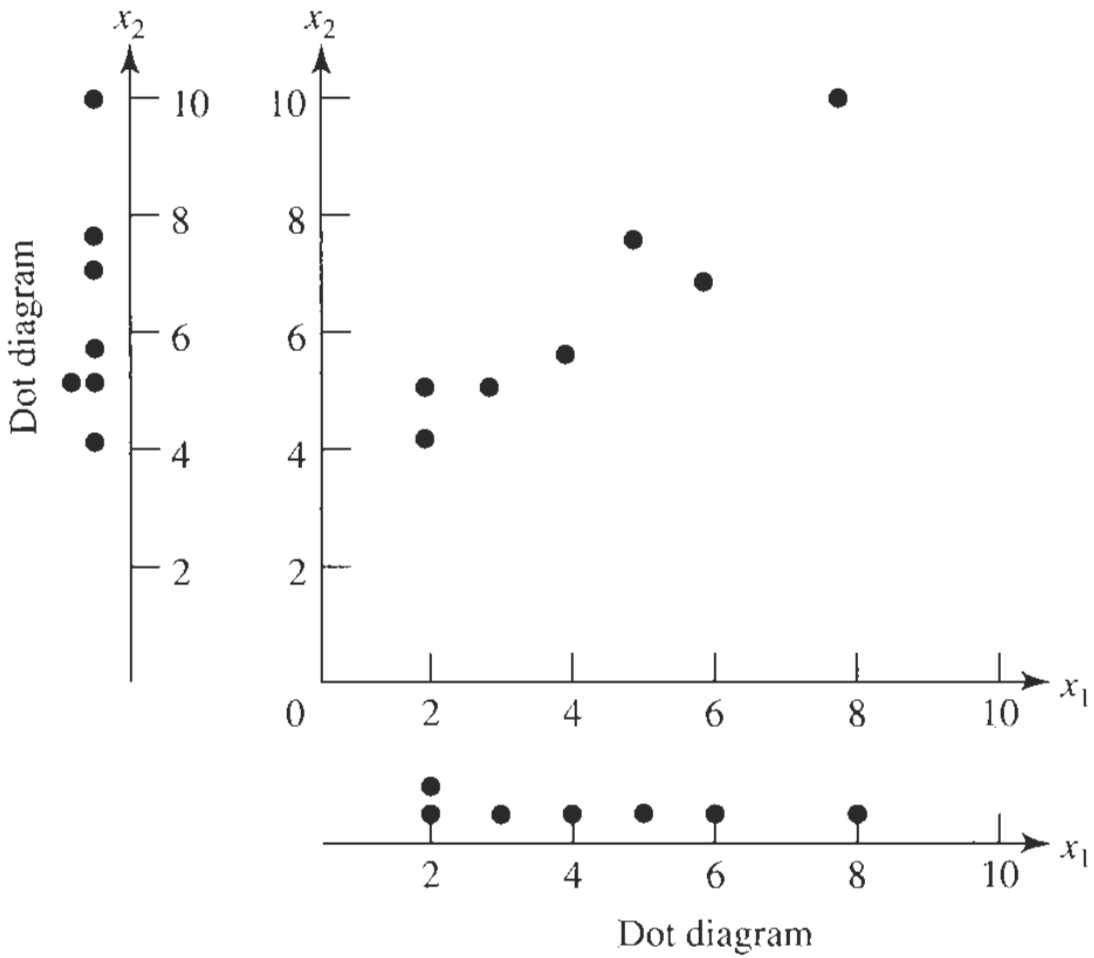


Figure 1.1 A scatter plot and marginal dot diagrams.

Variable 1	(x_1):	5	4	6	2	2	8	3
Variable 2	(x_2):	5	5.5	4	7	10	5	7.5

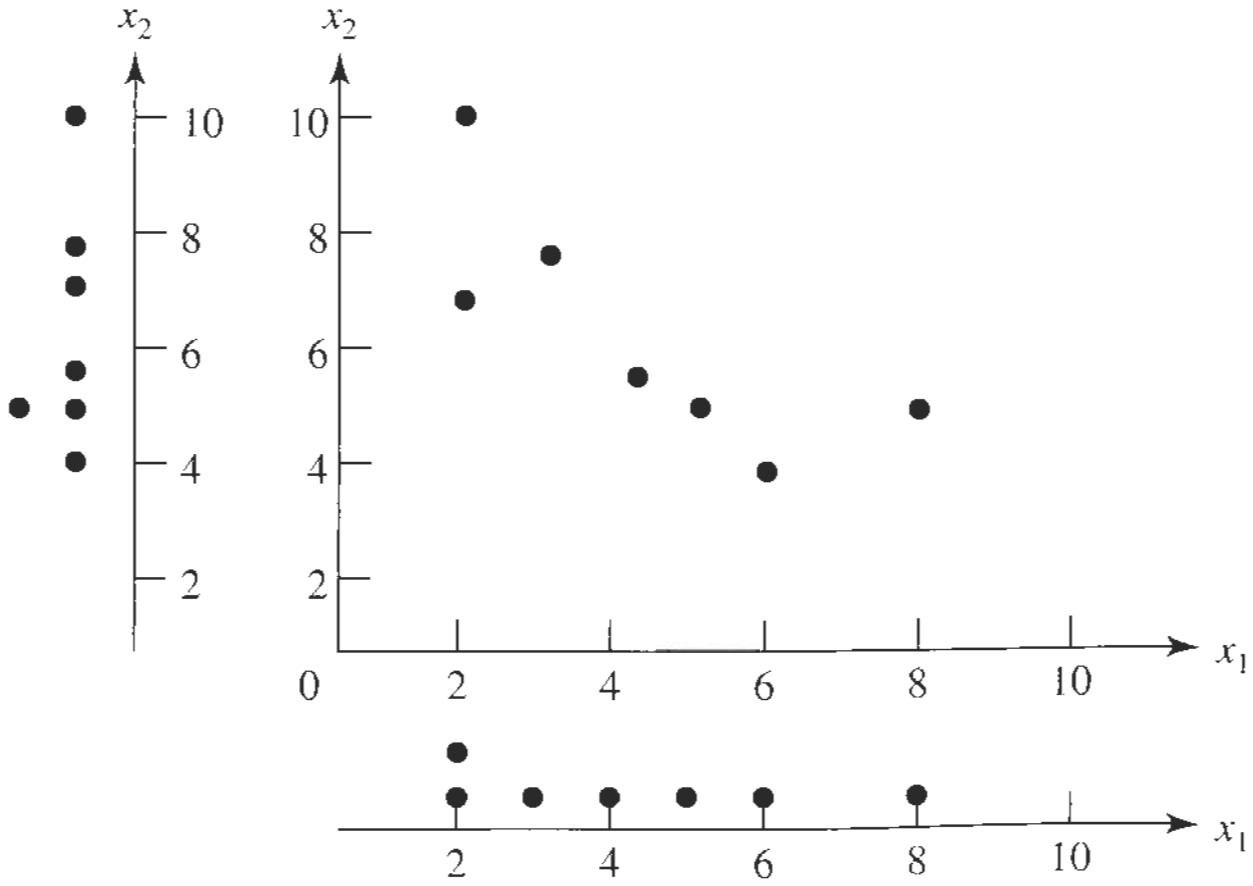


Figure 1.2 Scatter plo and dot diagrams for rearranged data.

Example 1.3 (The effect of unusual observations on sample correlations) Some financial data representing jobs and productivity for the 16 largest publishing firms appeared in an article in *Forbes* magazine on April 30,1990. The data for the pair of variable x_1 = employees(jobs) and x_2=profit per employee (productivity) are graphed in Figure 1.3. We have labeled two "unusual" observations. Dun&Bradstreet is the largest firm in term of number of employees, but is "typical" in terms of profits per employee. Time Warner has a "typical" number of employees, but comparatively small (negative) profit per employee.

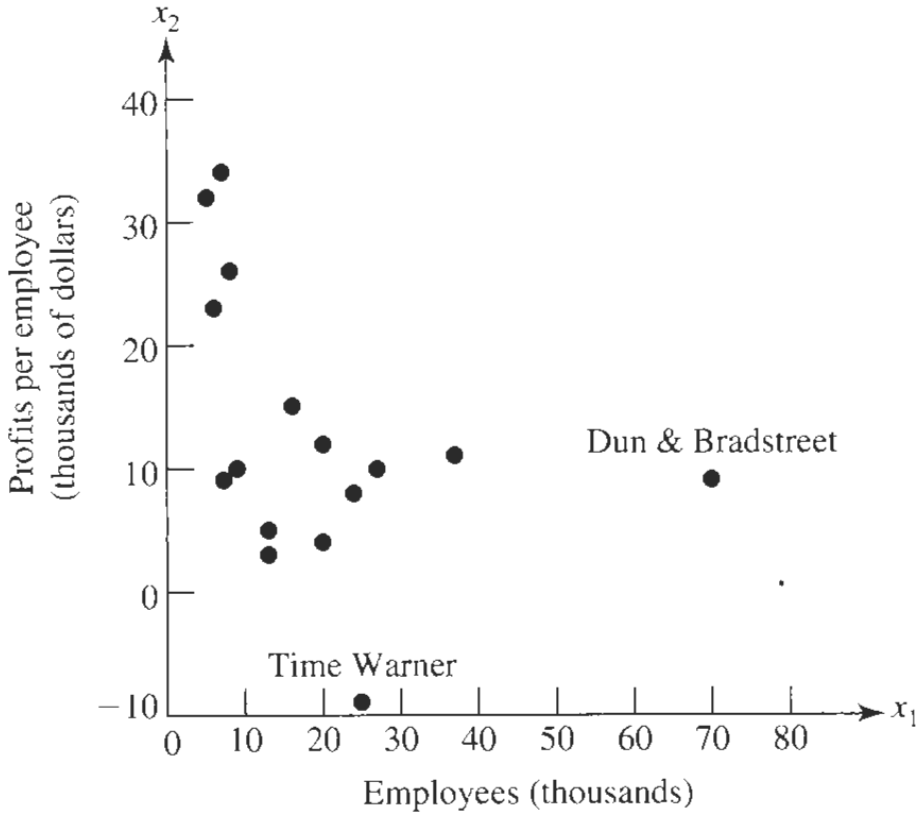


Figure 1.3 Profits per employee and number of employees for 16 publishing firms.

The sample correlation coefficient computed from the values of x_1 and x_2 is

$$r_{12} = \begin{cases} -0.39 & \text{for all 16 firms} \\ -0.56 & \text{for all firms but Dun and \& Bradstreet} \\ -0.39 & \text{for all firms but Time Warner} \\ -0.50 & \text{for all firms but Dun\&Bradstreet and Time Warner} \end{cases}$$

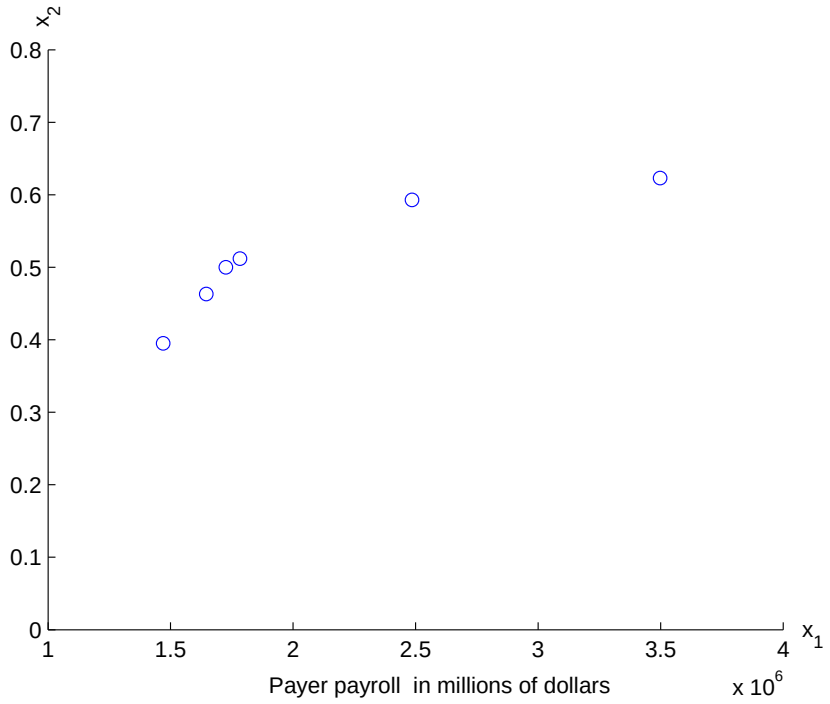
It is clear that atypical observations can have a considerable effect on the sample correlation coefficient.

Example 1.4 (A scatter plot for baseball data) In a July 17, 1978, article on money in sports, *Sports Illustrated* magazine provided data on x_1 =player payroll for National League East baseball teams.

We have added data on x_2 =won-lost percentage for 1977. The results are given in Table 1.1

Table 1.1 1977 Salary and Final Record for the National League East

Team	x_1 = player payroll	x_2 = won–lost percentage
Philadelphia Phillies	3,497,900	.623
Pittsburgh Pirates	2,485,475	.593
St. Louis Cardinals	1,782,875	.512
Chicago Cubs	1,725,450	.500
Montreal Expos	1,645,575	.463
New York Mets	1,469,800	.395



Example 1.5 (Multiple scatter plot for paper strength measurement) Paper is manufactured in continuous sheets several feet wide. Because of the orientation of fibers within the paper, it has a different strength when measured in the direction produced by the machine than when measured across, or at right angles to, the machine direction. The measured values includes

- $x_1 =$ density (grams/cublic centimeter)
- $x_2 =$ strength (pounds) in the machine direction
- $x_3 =$ strength (pounds) in the cross direction

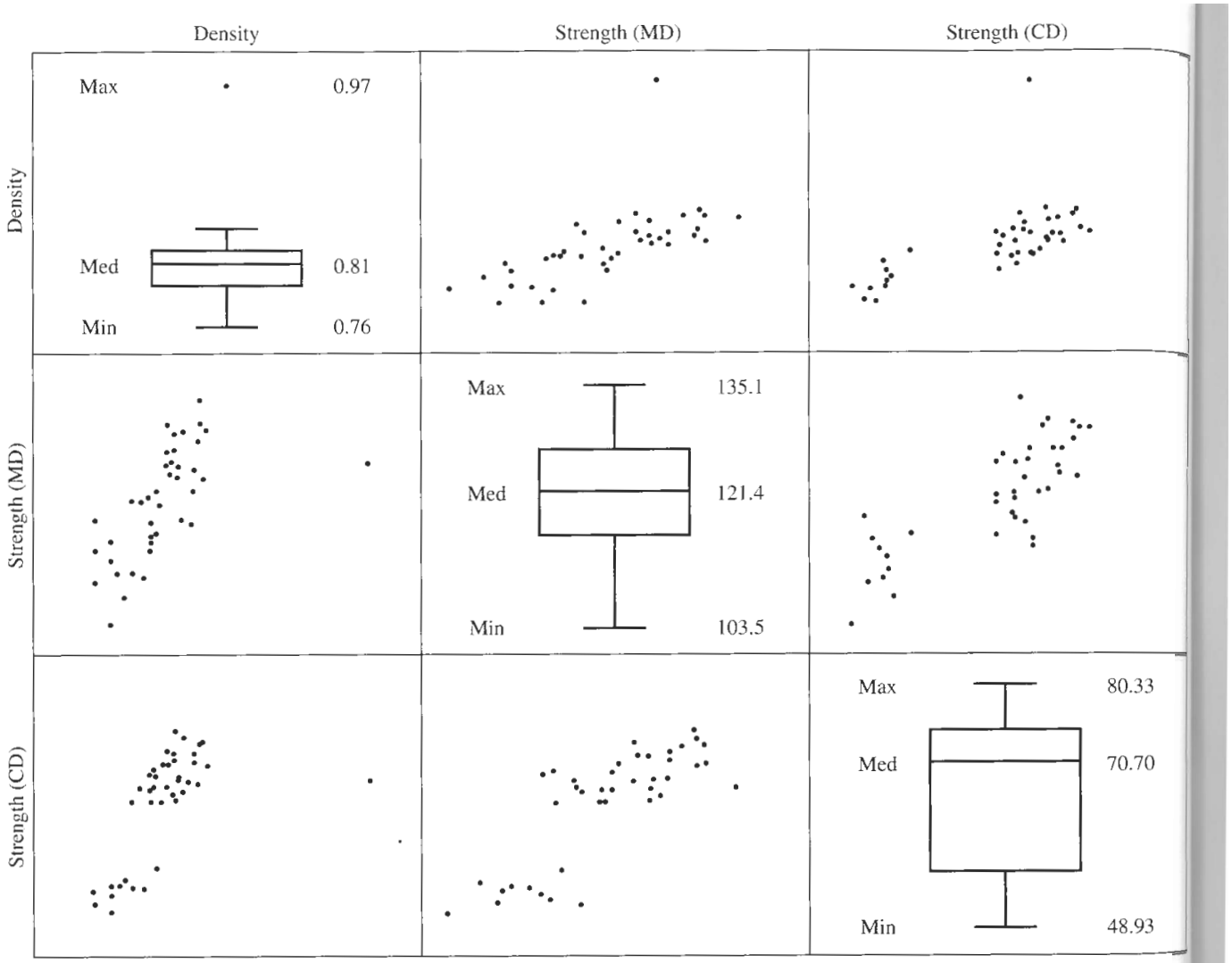


Figure 1.5 Scatter plots and boxplots of paper-quality data from Table 1.2.

Example 1.6 (Looking for lower-dimensional structure) A zoologist obtained measurement on $n = 25$ lizard known scientifically as *Cophosaurus texanus*. The weight, or mass, is given in millimeters. The data are displayed in Table 1.3.

Table 1.3 Lizard Size Data

Lizard	Mass	SVL	HLS	Lizard	Mass	SVL	HLS
1	5.526	59.0	113.5	14	10.067	73.0	136.5
2	10.401	75.0	142.0	15	10.091	73.0	135.5
3	9.213	69.0	124.0	16	10.888	77.0	139.0
4	8.953	67.5	125.0	17	7.610	61.5	118.0
5	7.063	62.0	129.5	18	7.733	66.5	133.5
6	6.610	62.0	123.0	19	12.015	79.5	150.0
7	11.273	74.0	140.0	20	10.049	74.0	137.0
8	2.447	47.0	97.0	21	5.149	59.5	116.0
9	15.493	86.5	162.0	22	9.158	68.0	123.0
10	9.004	69.0	126.5	23	12.132	75.0	141.0
11	8.199	70.5	136.0	24	6.978	66.5	117.0
12	6.601	64.5	116.0	25	6.890	63.0	117.0
13	7.622	67.5	135.0				

Source: Data courtesy of Kevin E. Bonine.

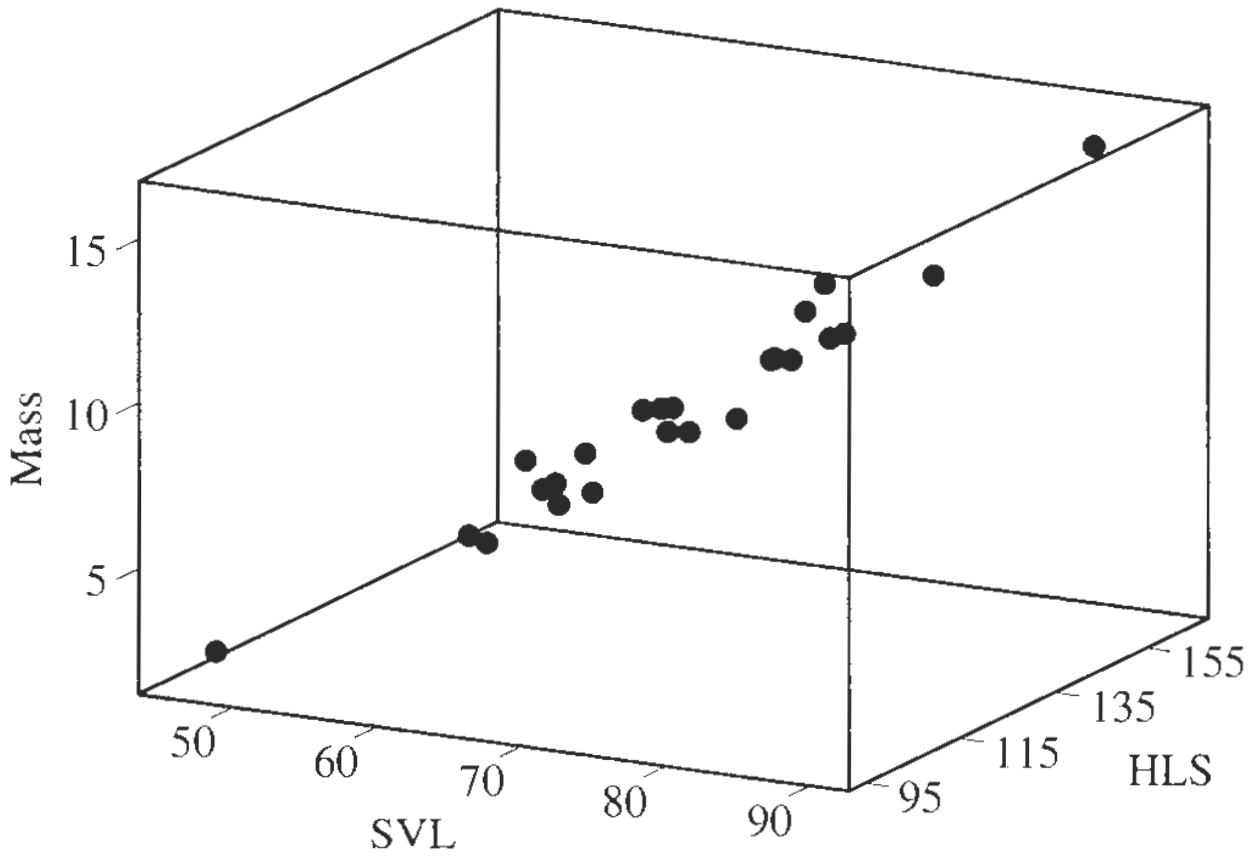


Figure 1.6 3D scatter
plot of lizard data fro
Table 1.3.

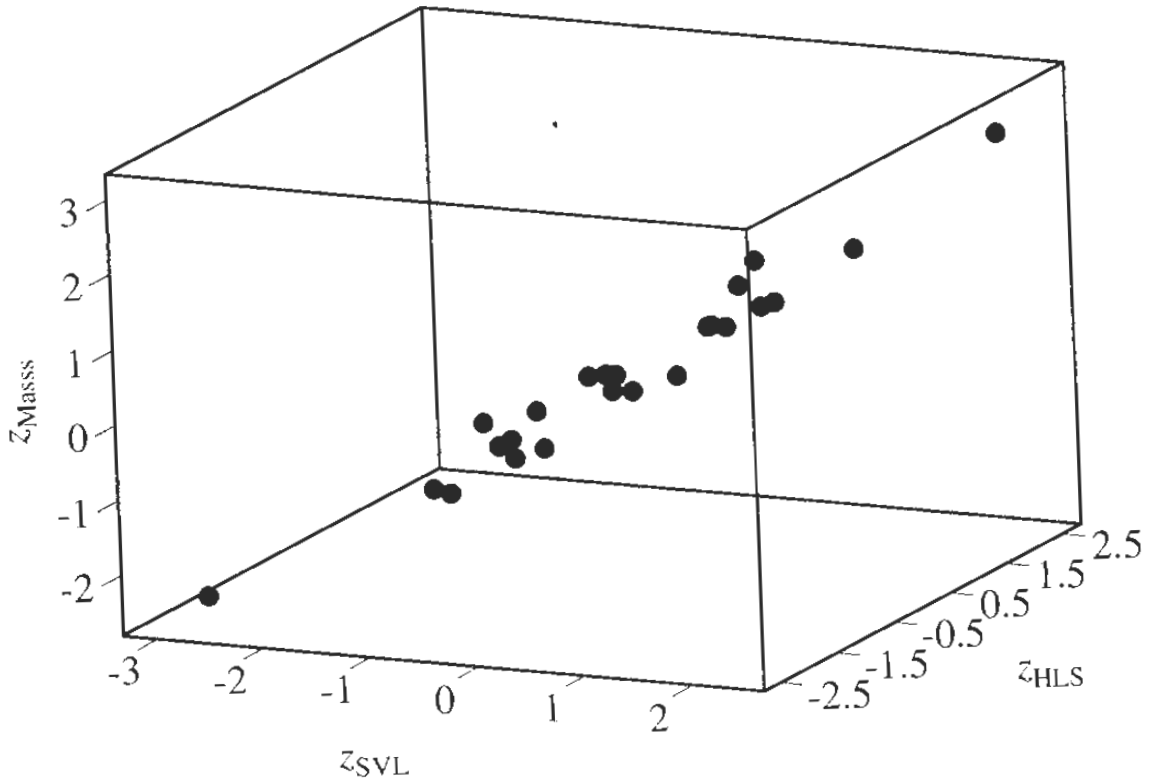


Figure 1.7 3D scatter plot of standardized lizard data.

Example 1.7 (Looking for group structure in three dimensions) Referring to Example 1.6, it is interesting to see if male and female lizard occupy different parts of three dimensional space containing the size data. The gender, by row, for the lizard data in Table 1.3 are

f m f f m f m f m f m f m m m m f m m m f f m f f

er than females.

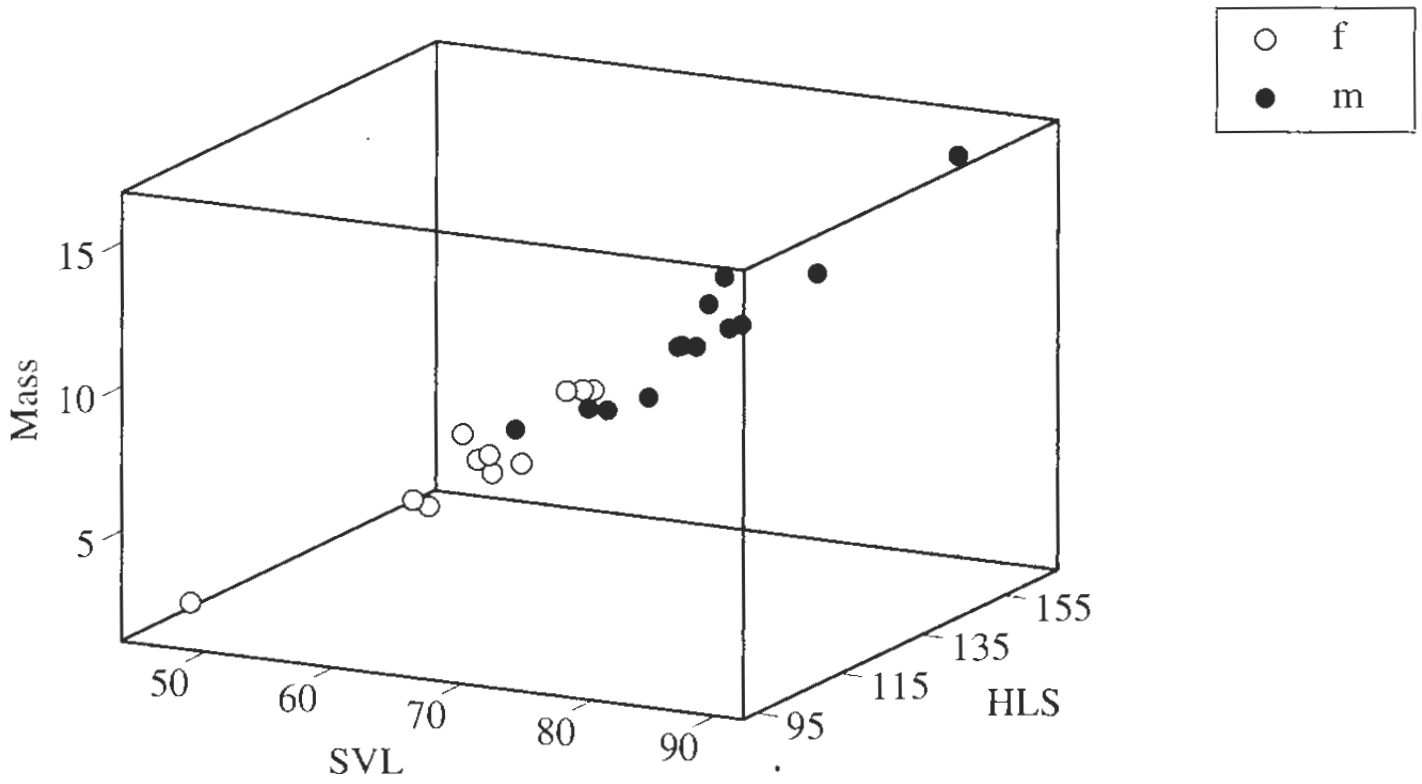


Figure 1.8 3D scatter plot of male and female lizards. ■

Data Display and Pictorial Representations

Linking Multiple Two-Dimensional Scatter Plots

Example 1.8 (Linked scatter plots and brushing) To illustrate *linked* two-dimensional scatter plots, we refer to the paper-quality data in Example 1.5. These data represent measurements on the variables x_1 =density, x_2 =strength in the machine direction, and x_3 =strength in the cross direction.

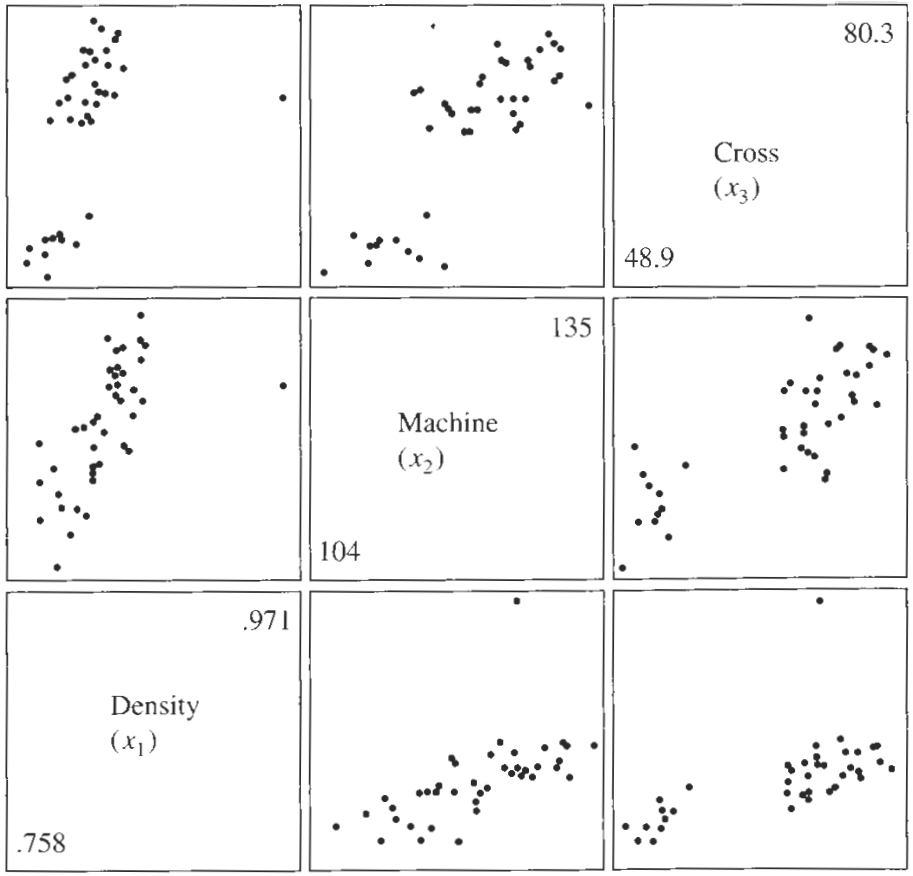
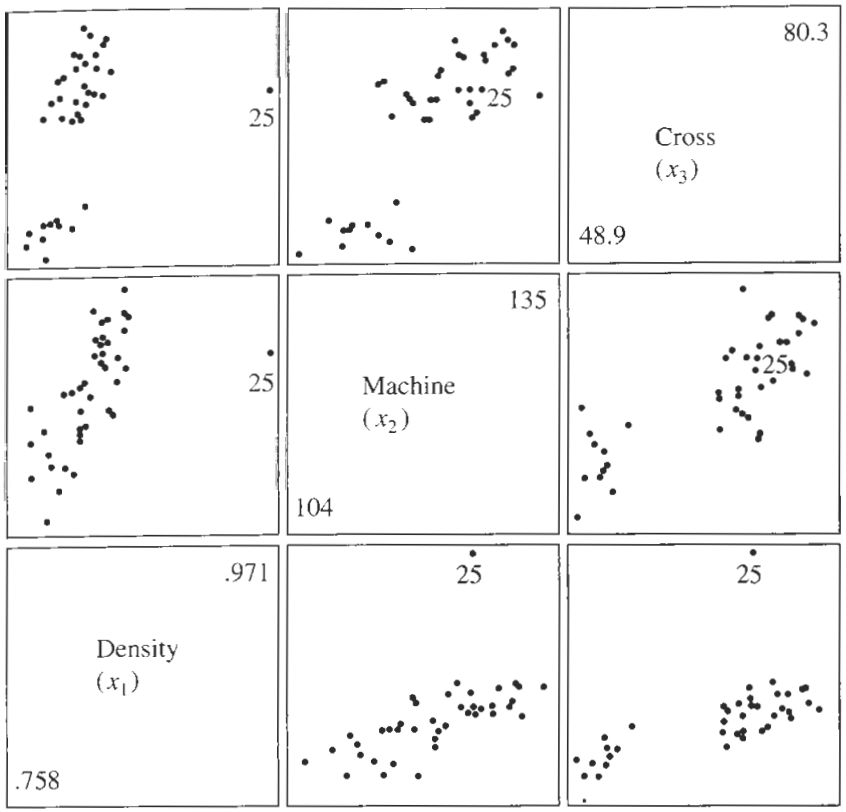
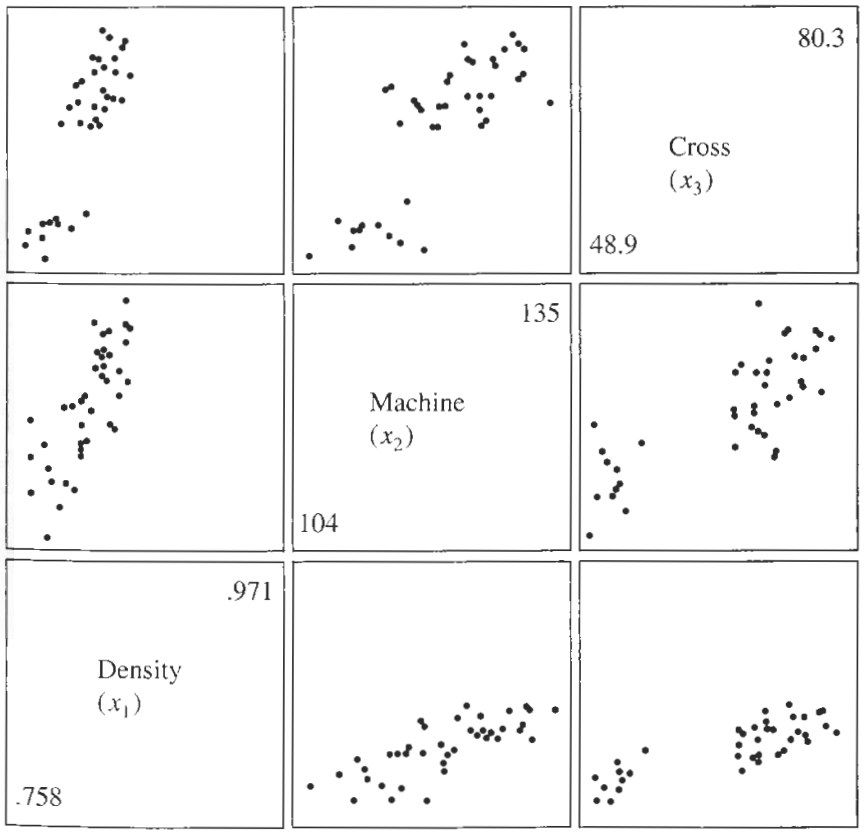


Figure 1.9 Scatter plots for the paper-quality data of Table 1.2.

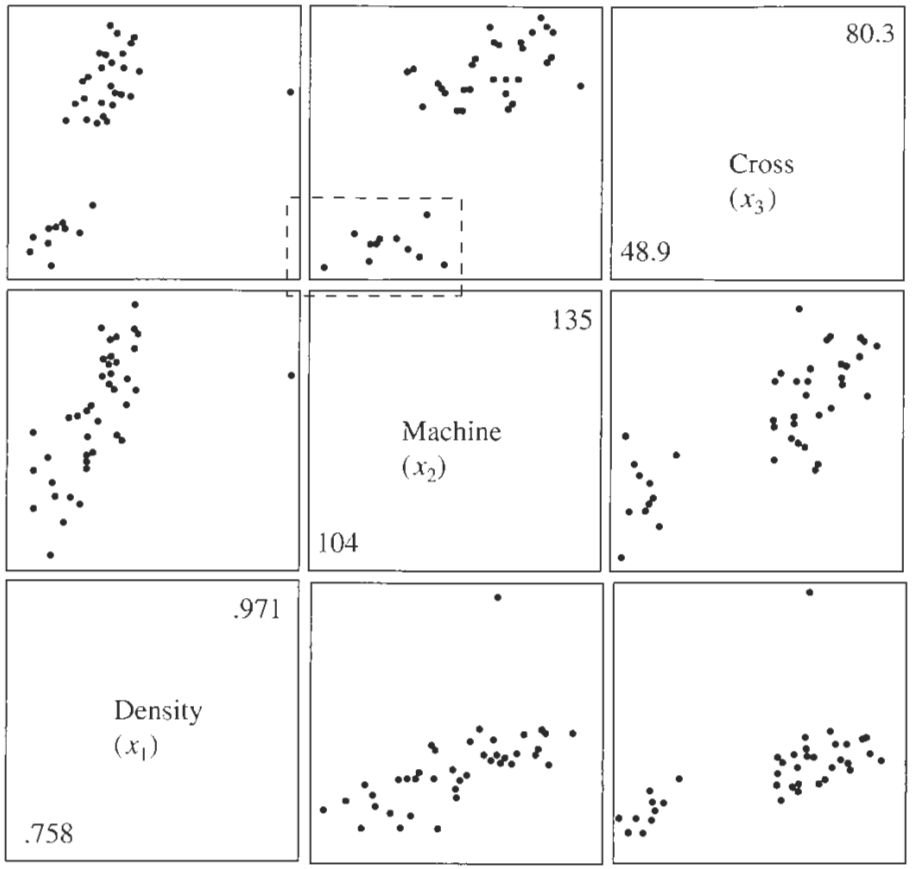

25
25
80.3
Cross
(x_3)
48.9
25
135
Machine
(x_2)
104
25
.971
Density
(x_1)
.758
25
25

(a)

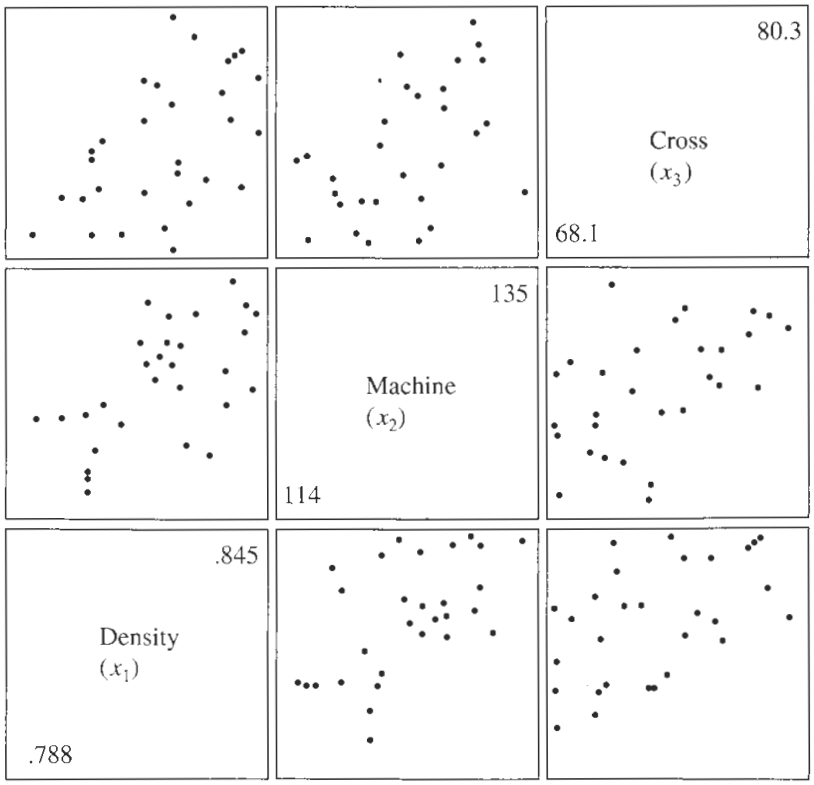


(b)

Figure 1.10 Modified scatter plots for the paper-quality data with outlier (25) (a) selected and (b) deleted.



(a)



(b)

Figure 1.11 Modified scatter plots with (a) group of points selected and (b) points, including specimen 25, deleted and the scatter plots rescaled.

Example 1.9 (Rotated plots in three dimensions) Four different measurements of lumber stiffness are given. Specimen (broad) 16 and possibly specimen (broad) 9 are identified as unusual observations.

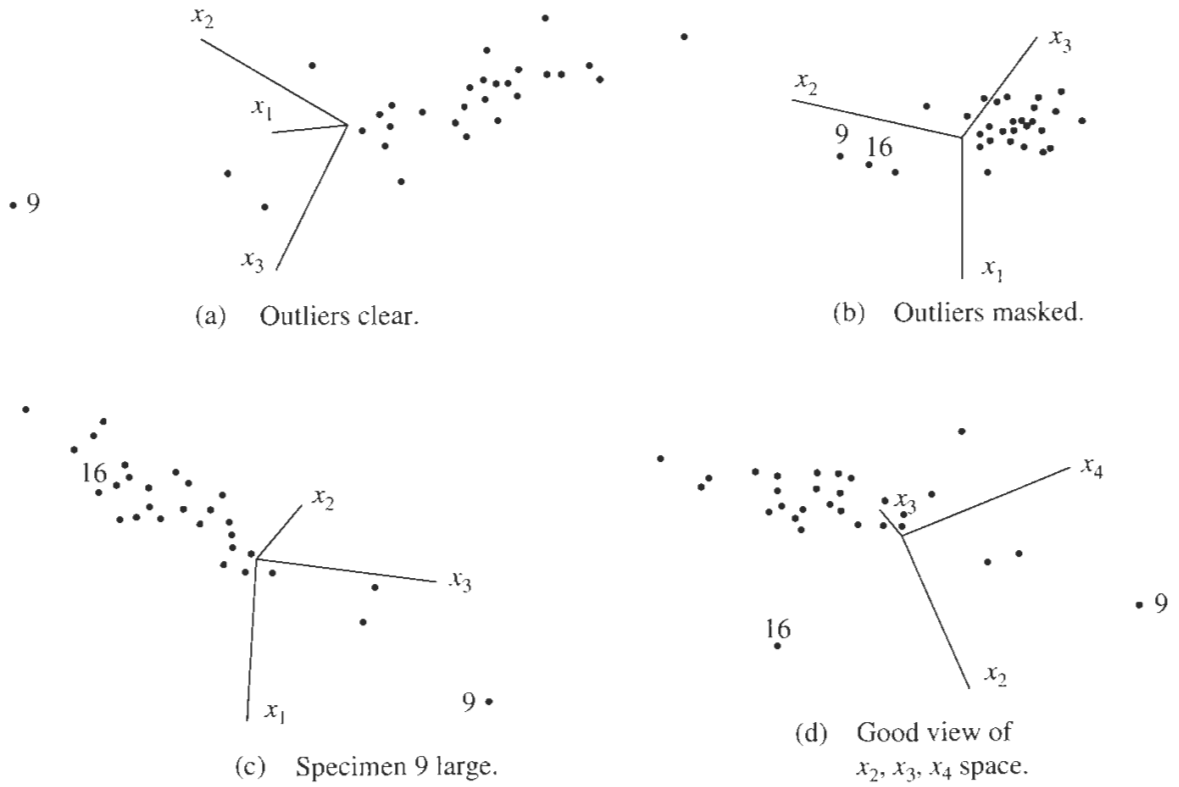


Figure 1.12 Three-dimensional perspectives for the lumber stiffness data.

Graphs of Growth Curves

Example 1.10 (Array of growth curves) The Alaska Fish and Game Department monitor grizzly bears with the goad of maintaining a health population. Bears are shot with a dart to induce sleep and weighted on a scale hanging from a tripod. Measurements of length are taken with a steel tape. The following Table gives the weights (wt) in kilograms and lengths (lngth) in centimeters of seven female bears at 2,3,4 and 5 years of age.

The noticeable exception to a common patter is the curve for bear 5. Is this an outlier or just natural variation in the population ? In the field, bears are weighted on a scale that reads pounds. Further inspection revealed that, in this case, an assistant later failed to convert the field reading to kilogram s when creating the electronic database. The correct weights are (45, 66, 84, 112) kilograms.

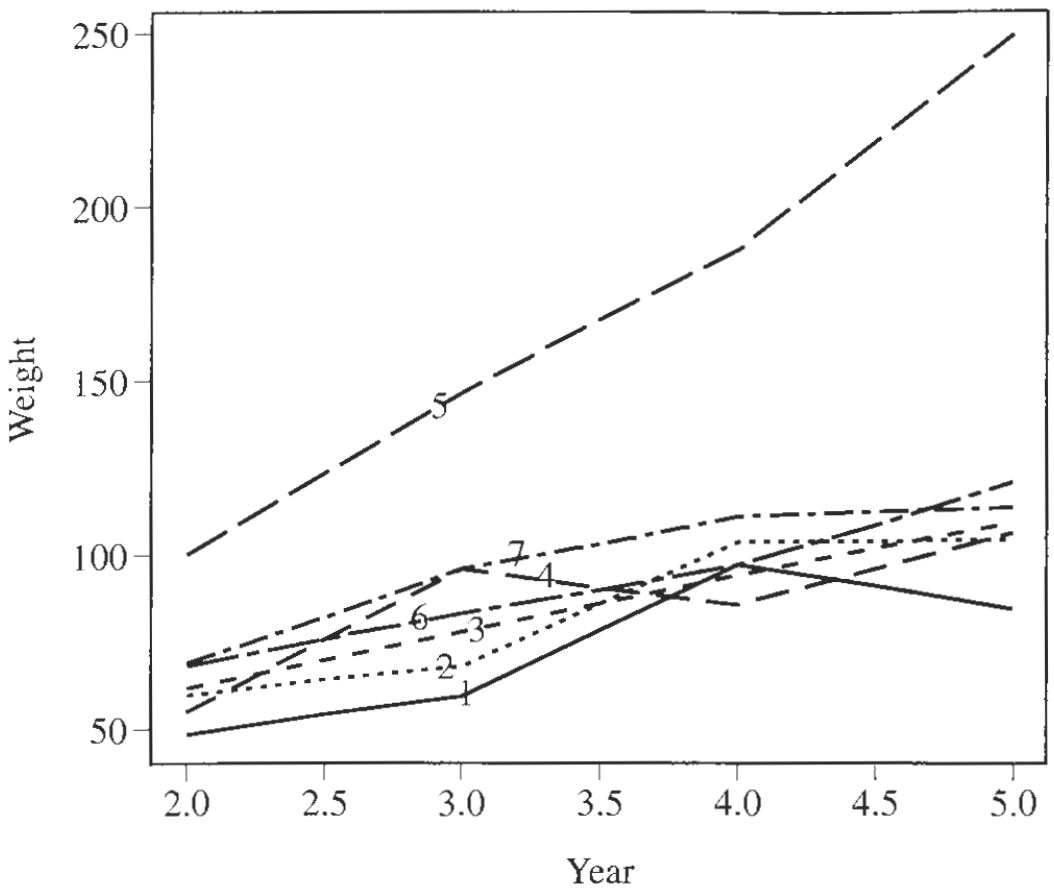


Figure 1.13 Combined growth curves for weight for seven female grizzly bears.

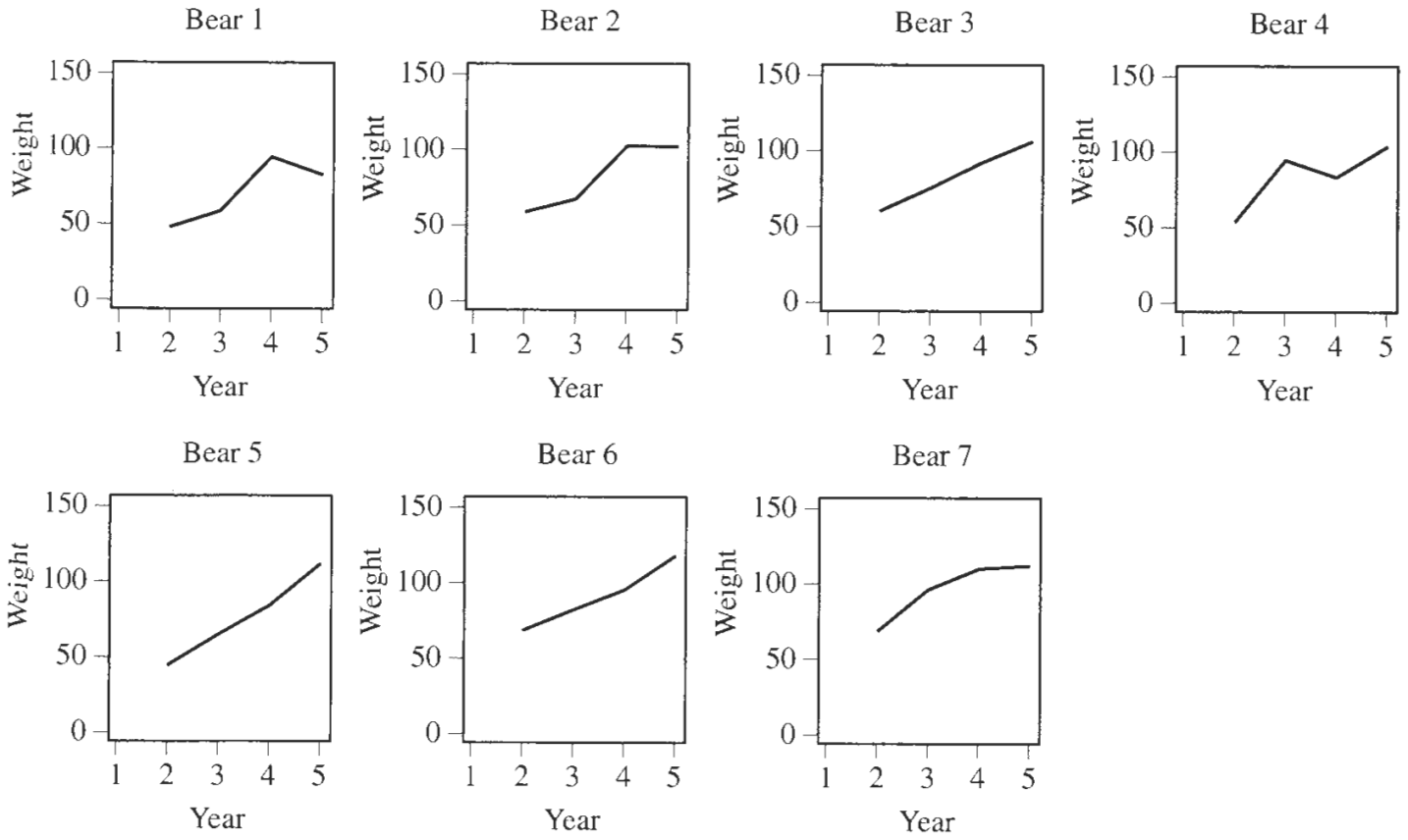


Figure 1.14 Individual growth curves for weight for female grizzly bears.

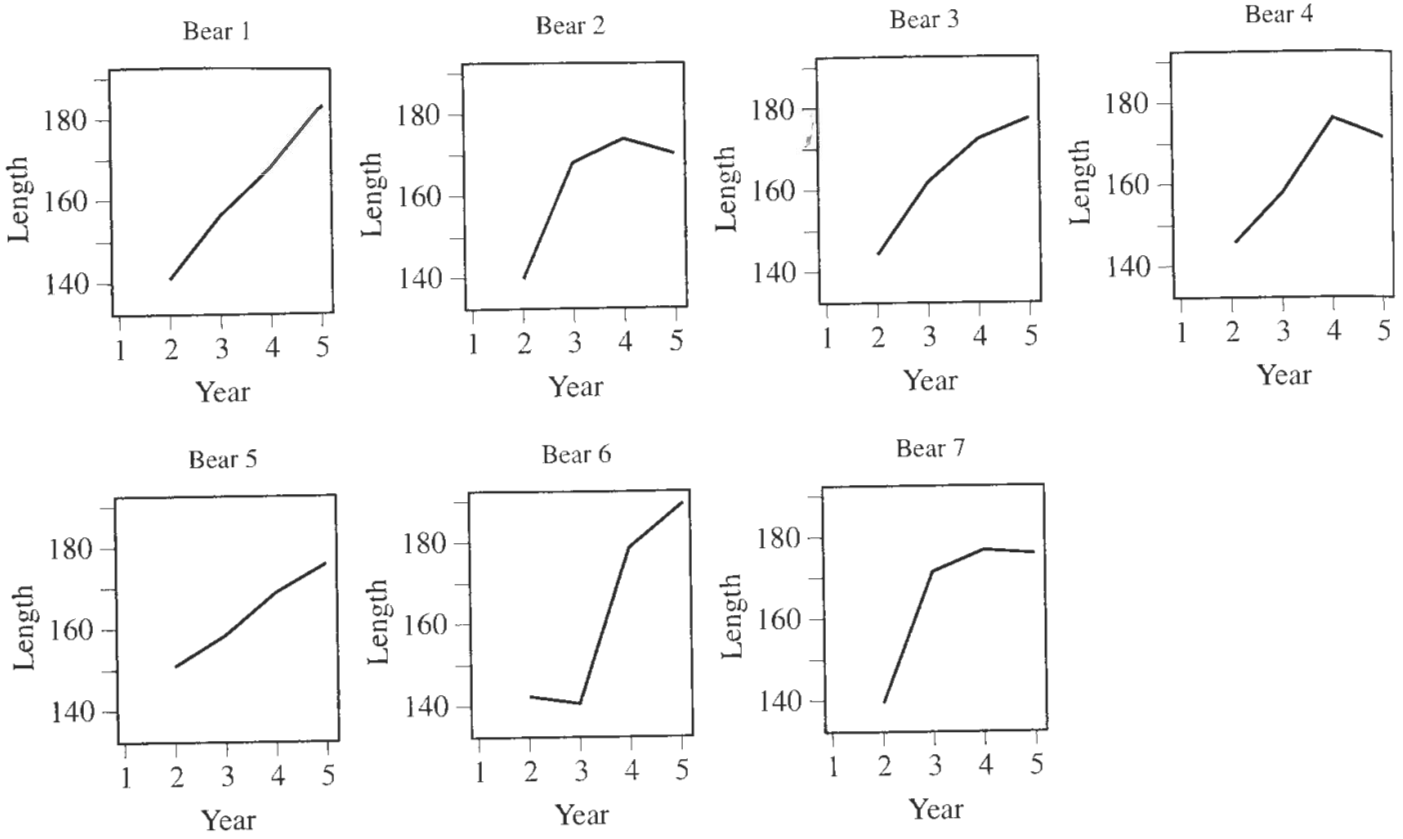


Figure 1.15 Individual growth curves for length for female grizzly bears.

Stars

Example 1.11 (Utility data as stars) Stars representing the first 5 of the 22 public utility firms data are shown in the following figure. There are eight variables; consequently, the stars are distorted octagons.

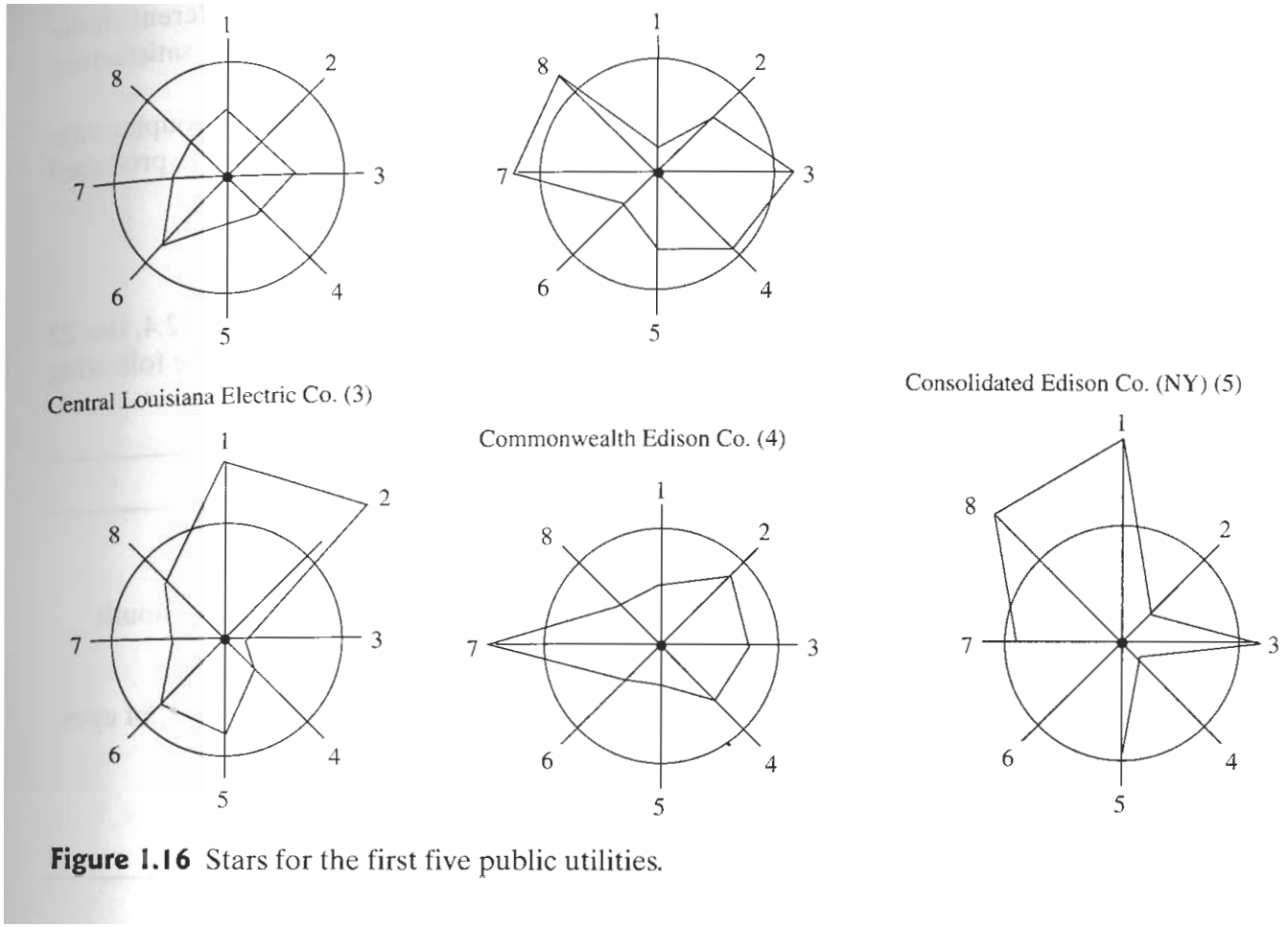


Figure 1.16 Stars for the first five public utilities.

Chernoff Faces

People react to faces. Chernoff suggest representing p-dimensional observation as a two-dimenional face whose characteristics (face shape, mouth curvature, nose length, eye size, pupil position, and so forth) are determined by the measurements on the p variables.

Chernoff faces appear to be most useful for verifying (1) an initial grouping suggested by subject-matter knowledge and intuition or (2) final groupings produced by clustering algorithms.

Example 1.12 (Utility data as Chernoff faces) The 22 public utility companies data were represented as chernoff faces. We have the following correspondences:

Variable		Facial characteristic
X_1: Fixed-charge coverage	$\leftrightarrow$	Half-height of face
X_2: Rate of return on capital	$\leftrightarrow$	Face width
X_3: Cost per kW capacity in place	$\leftrightarrow$	Position of center of mouth
X_4: Annual load factor	$\leftrightarrow$	Slant of eyes
X_5: Peak kWh demand growth from 1974	$\leftrightarrow$	Eccentricity $\left(\frac{\text{height}}{\text{width}}\right)$ of eyes
X_6: Sales (kWh use per year)	$\leftrightarrow$	Half-length of eye
X_7: Percent nuclear	$\leftrightarrow$	Curvature of mouth
X_8: Total fuel costs (cents per kWh)	$\leftrightarrow$	Length of nose

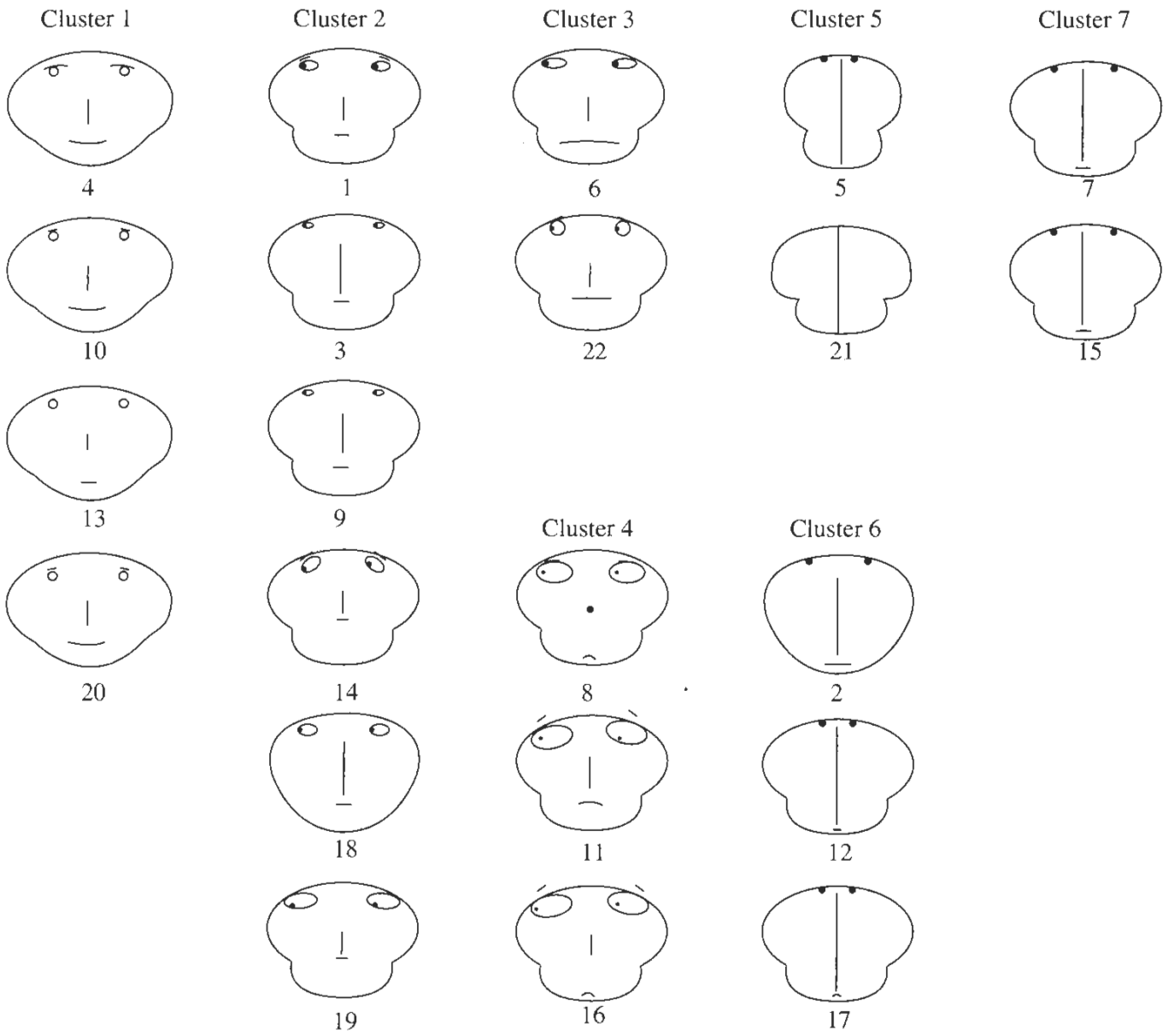


Figure 1.17 Chernoff faces for 22 public utilities.

Example 1.14 (Using Chernoff faces to show changes over time) The following figure illustrates an additional use of Chernoff faces. In the figure, the faces are used to track the financial well-being of a company over time. As indicated, each facial feature represent a single financial indicator, and the longitudinal changes in these indicators are thus evident at a glance

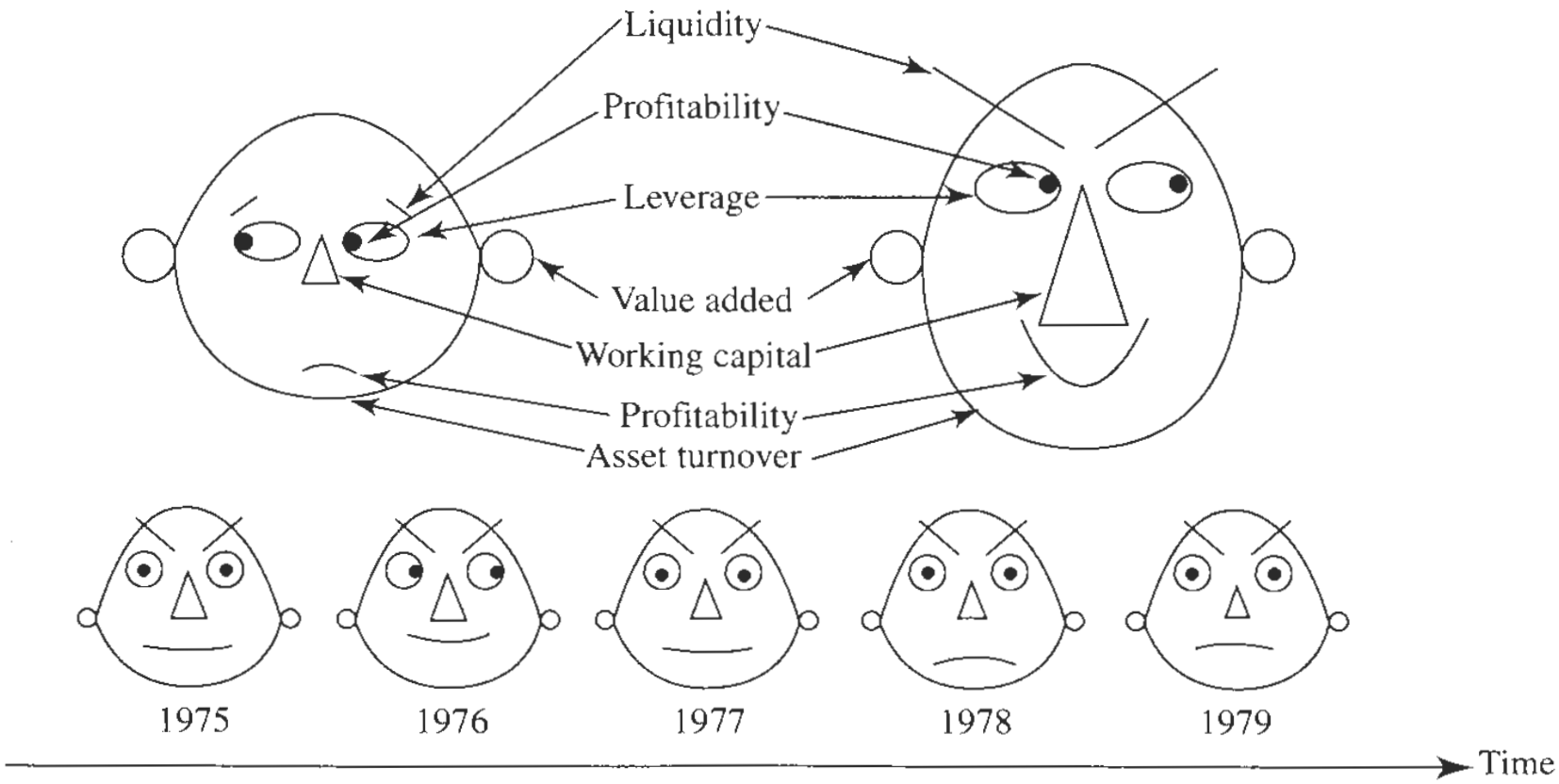


Figure 1.18 Chernoff faces over time.

1.5 Distance

- Straight-line, or Euclidean distance between $P = (x_1, x_2)$ and $O = (0, 0)$

$$d(O, P) = \sqrt{x_1^2 + x_2^2}$$

- In general, if the point $P = (x_1, x_2, \ldots, x_p)$ and $O = (0, 0, \ldots, 0)$

$$d(O, P) = \sqrt{x_1^2 + x_2^2 + \ldots + x_p^2}$$

- Straight-line or Euclidean distance is unsatisfactory for most statistical purposes. This is because each coordinates contributes equally to the calculation of Eucildean distance.

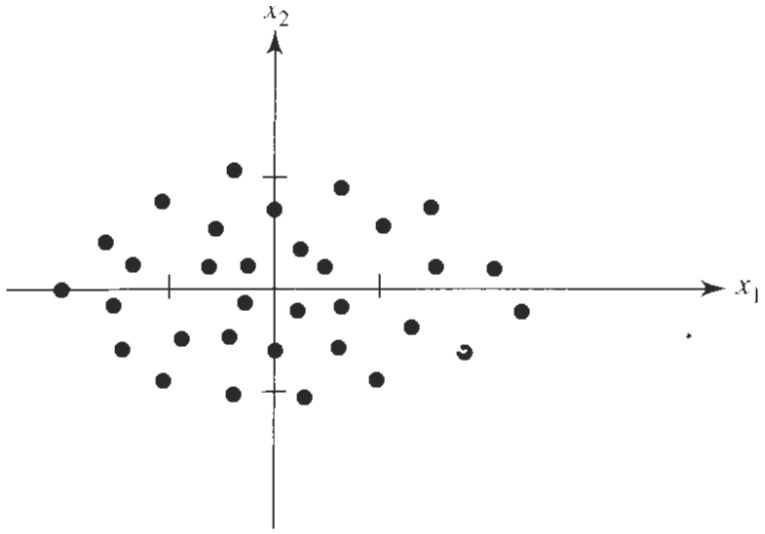


Figure 1.20 A scatter plot with greater variability in the x_1 direction than in the x_2 direction.

- **Statistical distance**

- When the coordinates represent measurements that are subject to random fluctuations of differing magnitudes, it is often to desirable to weight coordinates subject to a great deal of variability less heavily than those that are not highly variable.

- ***Standardize*** coordinates.

  Suppose we have n pairs of measurements on two variables x_1, x_2each having mean zero.

  $$x_1^* = x_1/\sqrt{s_{11}} \quad \text{and} \quad x_1^* = x_1/\sqrt{s_{22}}.$$

  Hence a statistical distance of the point $P = (x_1, x_2)$ from the origin $O = (0, 0)$ can be defined as

$$d(O, P) = \sqrt{(x_1^*)^2 + (x_2^*)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}.$$

- All points which have coordinates (x_1, x_2) and are constant square distance c^2 from origin muse satisfy

$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2$$

  and lie on an ellipse.

- The statistical distance from an arbitrary point $P = (x_1, x_2)$ to any *fixed* point $Q = (y_1, y_2)$.

$$d(P, Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}}.$$

- The extension of statistical distance to more than two dimensions $P = (x_1, x_2, \ldots, x_p)$ and $Q = (y_1, y_2, \ldots, y_p)$

$$d(P, Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}.$$

- All point P that are a constant squared distance from Q lie on a hyper-ellipsoid centered at Q whose major and minor axes are parallel to the coordinate axes.

The distances defined above does not include most of the important cases we shall encounter, because of the assumption of independent coordinates. See the following scatter plot

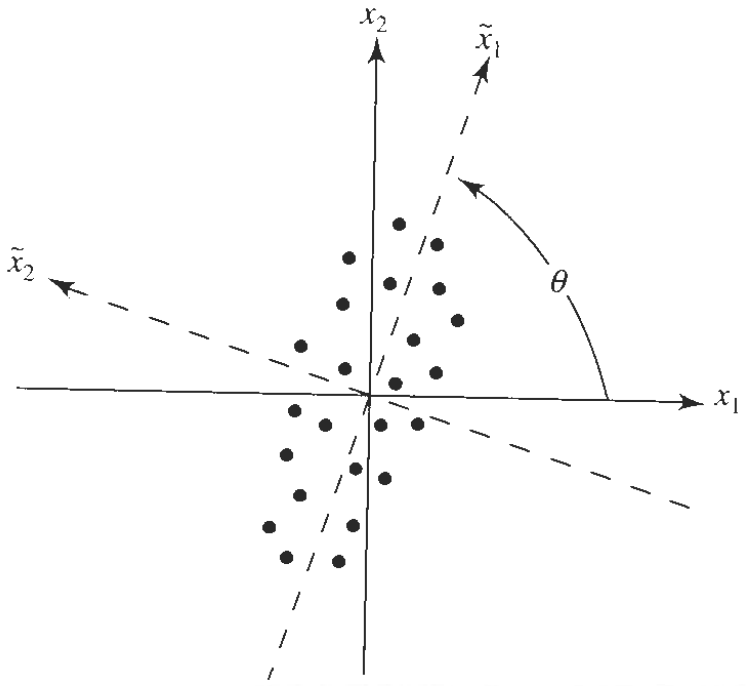


Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.

- Rotate x_1 and x_2 directions to directions $\tilde{x}_1$ and $\tilde{x}_2$.

- Define the distance from the point $P = (\tilde{x}_1, \tilde{x}_2)$ to the origin $O = (0, 0)$ as

$$d(O, P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}}.$$

  where $\tilde{s}_{11}$ and $\tilde{s}_{22}$ denote the sample variances computed with the $\tilde{x}_1$ and $\tilde{x}_2$ measurements.

- The relation between the original coordinates (x_1, x_2) and the rotated coordinates $(\tilde{x}_1, \tilde{x}_2)$ is provided by

$$\begin{aligned} \tilde{x}_1 &= x_1 \cos(\theta) + x_2 \sin(\theta) \\ \tilde{x}_2 &= -x_1 \sin(\theta) + x_2 \cos(\theta) \end{aligned}$$

- After some straightforward algebraic manipulations, the distance from $P = (\tilde{x}_1, \tilde{x}_2)$ to origin $O = (0, 0)$ can be written in term of the original coordinates x_1 and x_2

$$d(O, P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

where the a's are numbers such that the distance is nonnegative for all possible variables of x_1 and x_2.

- In general, the statistical distance of the point $P = (x_1, x_2)$ from the fixed point $Q = (y_1, y_2)$ for the situation in which the variables are correlated has the general form

$$d(P, Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$

- The generalization of the distance formulas to p dimensions

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} a_{ii}(x_i - y_i)^2 + \sum_{i \neq j}^{n} 2a_{ij}(x_i - y_i)(x_j - y_j)}$$

Any distance measure $d(P, Q)$ between two points P and Q is valid provided that it satisfies the following properties, where R is any other intermediate point:

(a) $d(P, Q) = d(Q, P)$

(b) $d(P, Q) > 0$ if $P \neq Q$

(c) $d(P, Q) = 0$ if $P = Q$

(d) $d(P, Q) \leq d(P, R) + d(R, Q)$