Organization of Multivariate Data

- Multivariate data arise whenever an investigator select a number p (≥1) of variables or characters to record. The values of these variables are all recorded for n distinct items, individuals, …
- Organization
  - ➢ x_{jk} = measurement of the kth variable on the jth item
  - ➢ n measurements on p variables can be displayed as follows:

	Variable 1	Variable 2	⋯	Variable k	⋯	Variable p
Item 1:	x_{11}	x_{12}	⋯	x_{1k}	⋯	x_{1p}
Item 2:	x_{21}	x_{22}	⋯	x_{2k}	⋯	x_{2p}
⋮	⋮	⋮		⋮		⋮
Item j:	x_{j1}	x_{j2}	⋯	x_{jk}	⋯	x_{jp}
⋮	⋮	⋮		⋮		⋮
Item n:	x_{n1}	x_{n2}	⋯	x_{nk}	⋯	x_{np}

  - ▪ The data can be displayed as a rectangular array:

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

➢Example: a selection of 4 receipts from a university bookstore

- ▪ Variable 1 (dollar sales): 42 52 48 58
  Variable 2 (number of books): 4 5 4 3
- ▪ $$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

- Some descriptive statistics (summary numbers)
  - ➢ sample mean: $\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$
  - ➢ sample variance: $s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2$
  - ➢ sample standard deviation: $\sqrt{s_{kk}}$
    - ▪ it uses the same units as the observations
  - ➢ sample covariance: $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)$
    - ▪ it measures the association between 2 variables
    - ▪ it reduces to the sample variance when i=k

➢ sample correlation:

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

- it measures the strength of "linear" association
- $-1 \le r \le 1$; $r=0 \Rightarrow$ no linear association
- r can be viewed as sample covariance of standardized data
- r remains unchanged if the variables are linearly transformed

➢ Arrays of basic descriptive statistics

Sample means $\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$

Sample variances and covariances $\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$

Sample correlations $\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$

- S_n and R are symmetric and positive semi-definite matrices

❖ **Reading**: Textbook, 1.3

Geometry of the Sample

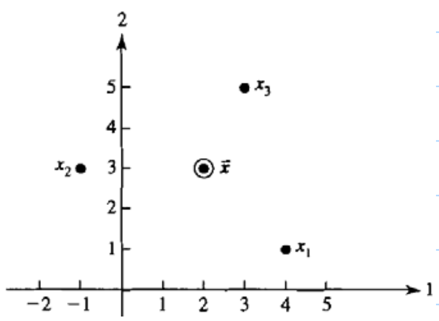
- A sample of size n from a p-variate "population": collection of measurements on p different variables taken on n items/trials

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- Approach 1: rows of X as n points in p-dimensional space

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

➢ scatter plot of n points in p-dim space provide information on the location & variability of the points



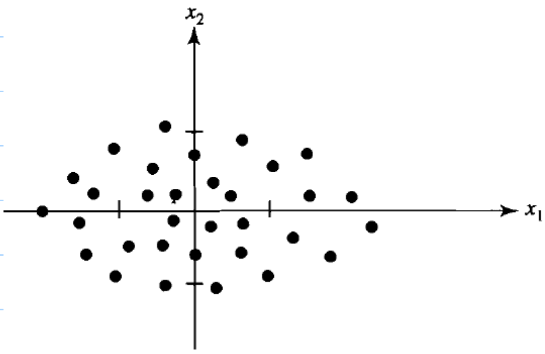
- distance: most multivariate techniques are based upon the concept of distance
  - Q: how to define distance between 2 multivariate data points?
  - Euclidean distance of two points

$$P = (x_1, x_2, \ldots, x_p) \qquad Q = (y_1, y_2, \ldots, y_p)$$

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

    - unsatisfactory for most statistical purpose because each coordinate contributes equally to the calculation of Euclidean distance (Note: the coordinates represent measurements subject to random fluctuations of differing magnitudes)
    - Q: how to account for difference in variation?
      Ans: weighting
  - *statistical distance* accounting for difference in variation & the present of correlation

- different variation & zero correlation



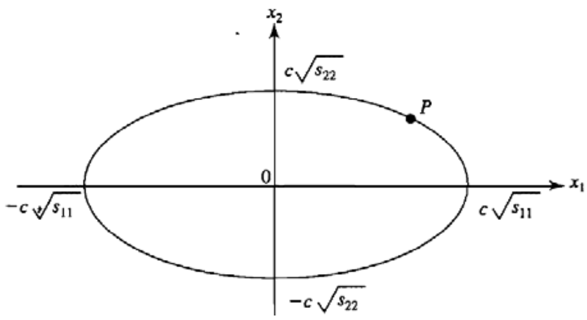
- values which are given deviation from the original in the x_1 direction are not as "surprising" or "unusual" as are values equidistant from the original in the x_2 direction

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

- all points with same distance from the original form an hyperellipsoid

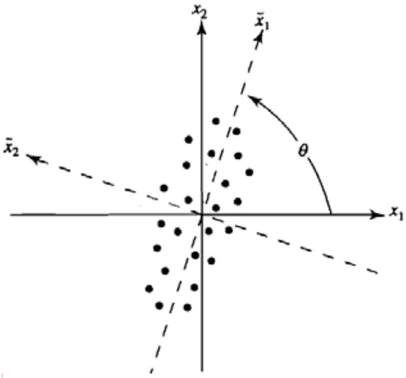
$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2$$

major axis/minor axis $= (s_{11}/s_{22})^{1/2}$



- If $s_{11} = \ldots = s_{pp}$, (*)=Euclidean distance

- different variation & nonzero correlation

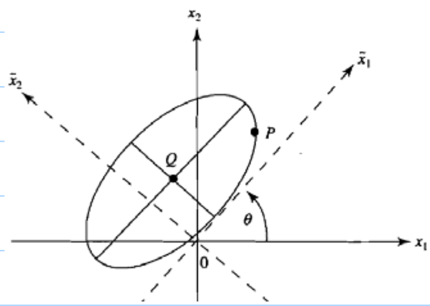


- the points exhibits a tendency to be large or small together
- Q: What is a meaningful measure of distance for the case? Ans: rotate coordinate system
- $\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$

  $\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$
- $$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}} = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$
- the appearance of $2a_{12}x_1x_2$ is necessitated by the correlation r_{12}
- $d(P,Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$

- all points that are a constant distance from the point Q is an ellipse centered at Q. Its major and minor axes are parallel to the $\tilde{x}_1$ and $\tilde{x}_2$ axes.

$$a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2 = c^2$$



- In general, for p-dim points

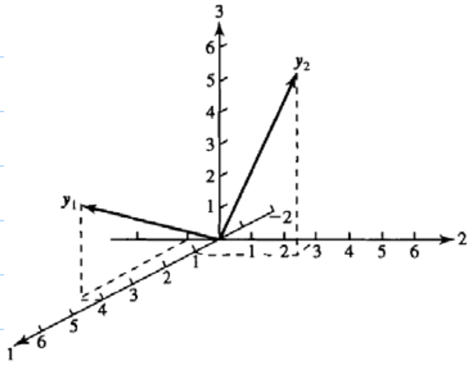
$$d(P,Q) = \sqrt{[a_{11}(x_1 - y_1)^2 + a_{22}(x_2 - y_2)^2 + \cdots + a_{pp}(x_p - y_p)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + 2a_{13}(x_1 - y_1)(x_3 - y_3) + \cdots + 2a_{p-1,p}(x_{p-1} - y_{p-1})(x_p - y_p)]}$$

$$= \left([x_1 - y_1, \ldots, x_p - y_p] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ \cdot \\ \cdot \\ \cdot \\ x_p - y_p \end{bmatrix} \right)^{1/2}$$

- the matrix $[a_{ij}]$ is related to the sample variance-covariance matrix

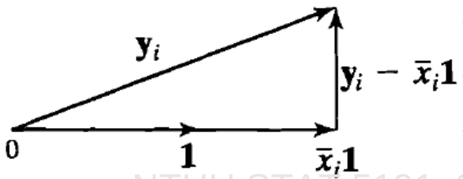
- Approach 2: columns of X as p vectors in n-dimensional space (c.f. approach 1)

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\mathbf{y}_1 \mid \mathbf{y}_2 \mid \cdots \mid \mathbf{y}_p]$$



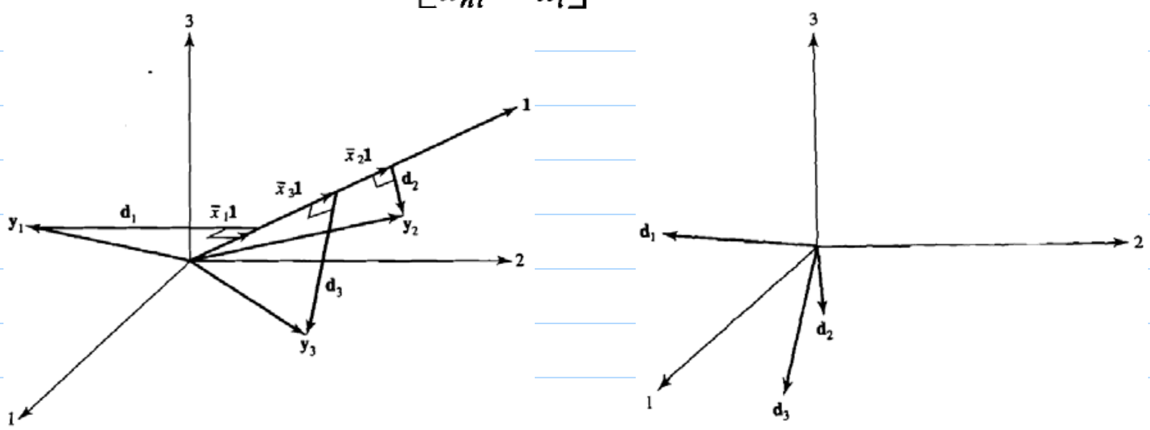
  - benefit of approach 2: many of the algebraic expressions we shall encounter in multivariate analysis can be related to the geometrical notion of *length*, *angle*, *volume*.
  - sample mean

$$\mathbf{y}_i'\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\frac{1}{\sqrt{n}}\mathbf{1} = \frac{x_{1i} + x_{2i} + \cdots + x_{ni}}{n}\mathbf{1} = \bar{x}_i\mathbf{1}$$



  - sample variance

$$\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1} = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix} \qquad L^2_{\mathbf{d}_i} = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2$$



  - sample covariance and correlation

$$\mathbf{d}_i'\mathbf{d}_k = L_{\mathbf{d}_i}L_{\mathbf{d}_k}\cos(\theta_{ik}) = \sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos(\theta_{ik})$$

- visualization of objects in 3-dim is useful to illustrate certain statistical concepts in terms of only 2 or 3 vectors of any n-dim

❖ **Reading**: Textbook, 3.2, 3.3, 1.6