Chapter 1

1.3 Organization of Data

Arrays

$$\begin{array}{c} \\ \text{Item 1} \\ \text{Item 2} \\ \vdots \\ \text{Item } n \end{array} \begin{array}{c} \begin{array}{cccc} \text{Var 1} & \text{Var 2} & \cdots & \text{Var } p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{array}$$

Descriptive Statistics

Mean: $$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \qquad k = 1, \ldots, p$$

Variance: $$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \qquad k = 1, \ldots, p \quad \text{(note divide by } n\text{)}$$

Covariance: $$s_{ik} = \frac{1}{n}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \qquad i = 1, \ldots, p;\ k = 1, \ldots, p$$

Correlation: $$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}\, s_{kk}}} = \frac{\sum_j (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_j (x_{ji} - \bar{x}_i)^2}\ \sqrt{\sum_j (x_{jk} - \bar{x}_k)^2}} \qquad \begin{array}{l} i = 1, \ldots, p \\ k = 1, \ldots, p \end{array}$$

Sum of squares: $$w_{ik} = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \qquad i = 1, \ldots, p \quad j = 1, \ldots, p$$

$$\underset{\sim}{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad S_n = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} \qquad R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Graphical Techniques

- Scatterplot and Marginal Dot Diagrams
- Unusual observations' effect on correlation
- Scatterplot Matrix
- Looking for lower dimensional structure $z_{jk} = \frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}}$
- Rotated plots in 3 Dimensions
- Graphs of Growth Curves
  - Combined Graphs
  - Individual (array) graphs
- Star Graphs
- Chernoff Faces

Distance

2 dim: $P = (x_1, x_2) \quad O = (0,0) \quad d(O,P) = \sqrt{x_1^2 + x_2^2}$

p dim: $P = (x_1, x_2, \dots, x_p) \quad O = (0, \dots, 0) \quad d(O,P) = \sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$

All points a fixed distance c^2 from origin satisfy:

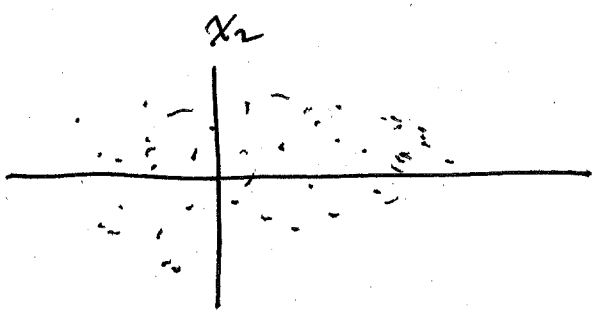
$$d^2(O,P) = x_1^2 + x_2^2 + \dots + x_p^2 = c^2 \equiv \text{hypersphere}$$

$P = (x_1, \dots, x_p) \quad Q = (y_1, \dots, y_p)$

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

Statistical Distance

Suppose n pairs of measurements on 2 variables each w/ mean 0: x_1, x_2 w/ $\bar{x}_1 = \bar{x}_2 = 0$



x_1, x_2 independent

$s_{11} > s_{22}$

Want to weight x_2 higher than x_1

$$x_1^* = \frac{x_1}{\sqrt{s_{11}}} \qquad x_2^* = \frac{x_2}{\sqrt{s_{22}}}$$

Statistical distance from $P = (x_1, x_2)$ from origin $O = (0,0)$:

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

If $s_{11} = s_{22}$, x_1, x_2 independent $\Rightarrow$ use Euclidean Distance

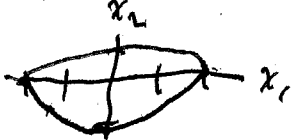
$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2 \equiv \text{Constant squared distance from origin}$$

EXAMPLE 1.4 $\quad \bar{x}_1 = \bar{x}_2 = 0 \quad s_{11} = 4 \quad s_{22} = 1 \quad x_1 \perp x_2$

$$d^2(O,P) = \frac{x_1^2}{4} + \frac{x_2^2}{1}$$

Points w/ constant distance 1:

$$\frac{x_1^2}{4} + \frac{x_2^2}{1} = c^2 = 1$$



x_1, x_2	Distance^2: $\frac{x_1^2}{4} + \frac{x_2^2}{1}$
$(0,1)$	1
$(0,-1)$	1
$(2,0)$	1
$(1,\sqrt{3}/2)$	1

$P = (x_1, x_2)$ $Q = (y_1, y_2)$ Q fixed

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

p-dimensions $P = (x_1, \ldots, x_p)$ $Q = (y_1, \ldots, y_p)$ Q fixed

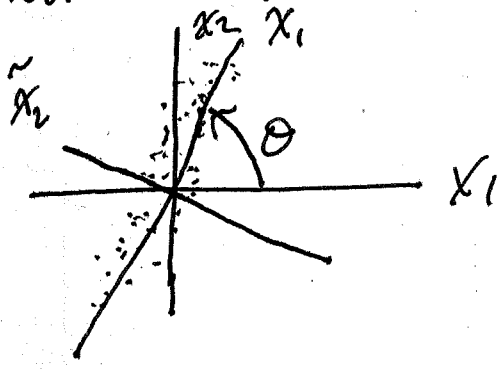
$$d(P,Q) = \sqrt{\frac{(x_1-y_1)^2}{s_{11}} + \cdots + \frac{(x_p-y_p)^2}{s_{pp}}} \qquad s_{ii} \equiv V\{x_i\}$$

1) Distance from P to the origin $\Rightarrow$ Let $y_1 = \ldots = y_p = 0$

2) If $s_{11} = s_{22} = \cdots = s_{pp}$ $\rightarrow$ Euclidean distance

Not independent data:

$r_{12} > 0$



$P = (\tilde{x}_1, \tilde{x}_2)$ $O = (0,0)$

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

$\tilde{s}_{ii}$ = Sample variance of $\tilde{x}_i$

$$\tilde{x}_1 = x_1 \cos\theta + x_2 \sin\theta$$

$$\tilde{x}_2 = -x_1 \sin\theta + x_2 \cos\theta$$

$$\tilde{x}_1^2 = x_1^2\cos^2\theta + x_2^2\sin^2\theta + 2x_1x_2\cos\theta\sin\theta$$

$$\tilde{x}_2^2 = x_1^2\sin^2\theta + x_2^2\cos^2\theta - 2x_1x_2\sin\theta\cos\theta$$

$$d(0,P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

$$a_{11} = \frac{\cos^2\theta}{\cos^2\theta(s_{11}) + 2\sin\theta\cos\theta(s_{12}) + \sin^2\theta\, s_{22}} + \frac{\sin^2\theta}{\cos^2\theta(s_{22}) - 2\sin\theta\cos\theta(s_{12}) + \sin^2\theta(s_{11})}$$

$$= \frac{\cos^2\theta}{d_1} + \frac{\sin^2\theta}{d_2}$$

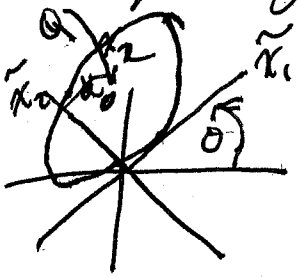
$$a_{22} = \frac{\sin^2\theta}{d_1} + \frac{\cos^2\theta}{d_2}$$

$$a_{12} = \frac{\cos\theta\sin\theta}{d_1} - \frac{\sin\theta\cos\theta}{d_2}$$

Distance from P to fixed $Q = (y_1, y_2)$

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

All points that are constant squared distance c^2 from Q satisfy $d^2(P,Q) = c^2 \equiv$ Ellipse centered @ Q.



In general, in p dimensions

$$d(O,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{pp}x_p^2 + 2a_{12}x_1x_2 + \cdots + 2a_{p-1,p}x_{p-1}x_p} \quad (*)$$

$$d(P,Q) = \sqrt{\sum_{i=1}^{p} a_{ii}(x_i - y_i)^2 + 2\sum_{i=1}^{p-1}\sum_{i'=i+1}^{p} a_{ii'}(x_i - y_i)(x_{i'} - y_{i'})} \quad (**)$$

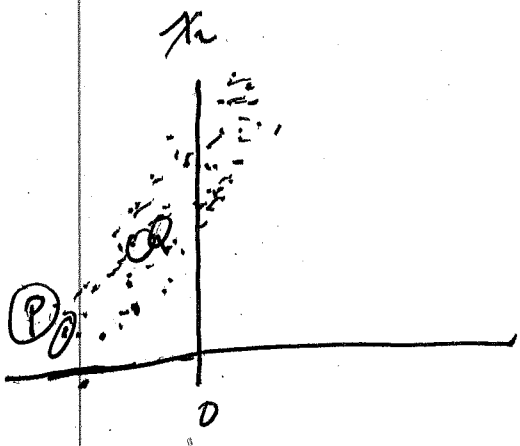
d^2's are quadratic forms (positive definite)

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{bmatrix}$$

Contours of constant distances computed from (*) and (**)

are <u>hyperellipsoids</u>. $p=3 \Rightarrow$ Football

Not visualizable if $p>3$



$Q \equiv$ Cent of gravity

Q could be closer to O in Euclidean distance than to P.

But Q closer to P in statistical distance

Valid Measures of distance must have:

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

$d(P,Q) > 0 \quad \text{if} \quad P \neq Q$

$d(P,Q) = 0 \quad \text{if} \quad P = Q$

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