3 Amostragem multivariada

3.1. Introdução

Com os conceitos de álgebra vetorial introduzidos no capítulo 2, pode-se aprofundar na interpretação geométrica das estatísticas descritivas $\overline{\underset{\sim}{X}}$, **S** e **R**. A maioria das explicações usam a representação das colunas de **X**, como p pontos no espaço n dimensional. Será introduzida neste instante a pressuposição de que as observações constituem uma amostra aleatória. De uma forma simplificada, amostra aleatória significa (i) que as medidas tomadas em diferentes itens (unidades amostrais ou experimentais) são não relacionadas uma com as outras, e (ii) que a distribuição conjunta das p variáveis permanece a mesma para todos os itens. Essa estrutura de amostra aleatória é que justifica uma escolha particular de distância e dita a geometria para a representação n dimensional dos dados. Finalmente, quando os dados podem ser tratados como uma amostra aleatória à inferência estatística terá por base um sólido fundamento.

3.2. Geometria amostral

Uma observação multivariada é uma coleção de medidas em p variáveis tomadas na mesma unidade amostral ou experimental. No capítulo 1, item 1.3.1, as n observações obtidas foram dispostas em um arranjo (Matriz) **X** por,

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

em que cada linha de **X** representa uma observação multivariada. Desde que o conjunto todo de mensurações é muitas vezes uma particular realização de variáveis aleatórias, diz-se que os dados representam uma amostra de tamanho n de uma população p variada.

Os dados podem ser plotados por um gráfico com p coordenadas. As colunas de **X** representam n pontos no espaço p dimensional. Esse tipo de gráfico fornece informações de locação dos pontos e de variabilidade. Se os pontos pertencem a uma esfera, o vetor de médias amostrais, $\overline{\underset{\sim}{X}}$, é o centro de balanço ou de massa. Se a variabilidade ocorre em mais de uma direção, pode-se detectar pela matriz de covariância, **S**. Uma medida numérica única de variabilidade é fornecida pelo determinante da matriz de covariância.

Exemplo 3.1

Calcule o vetor média $\overline{\underset{\sim}{X}}$ para a matriz **X** apresentada a seguir. Plote os n=3 pontos no espaço p=2 (bidimensional) e localize $\overline{\underset{\sim}{X}}$ no diagrama resultante.

$$X = \begin{bmatrix} 2 & 1 \\ -3 & 0 \\ -2 & 2 \end{bmatrix}$$

A média amostral é dada por:

$$\overline{\underset{\sim}{X}} = \begin{bmatrix} [(2+(-3)+(-2)]/3 \\ (1+0+2)/3 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

O primeiro ponto é dado por $\underset{\sim}{X}_1' = [2 \quad 1]$, o segundo por $\underset{\sim}{X}_2' = [-3 \quad 0]$, e o terceiro por $\underset{\sim}{X}_3' = [-2 \quad 2]$. A Figura 3.1 mostra os pontos juntamente com $\overline{\underset{\sim}{X}}$, centro de massa ou de balanço. obtidos pela matriz **X**.

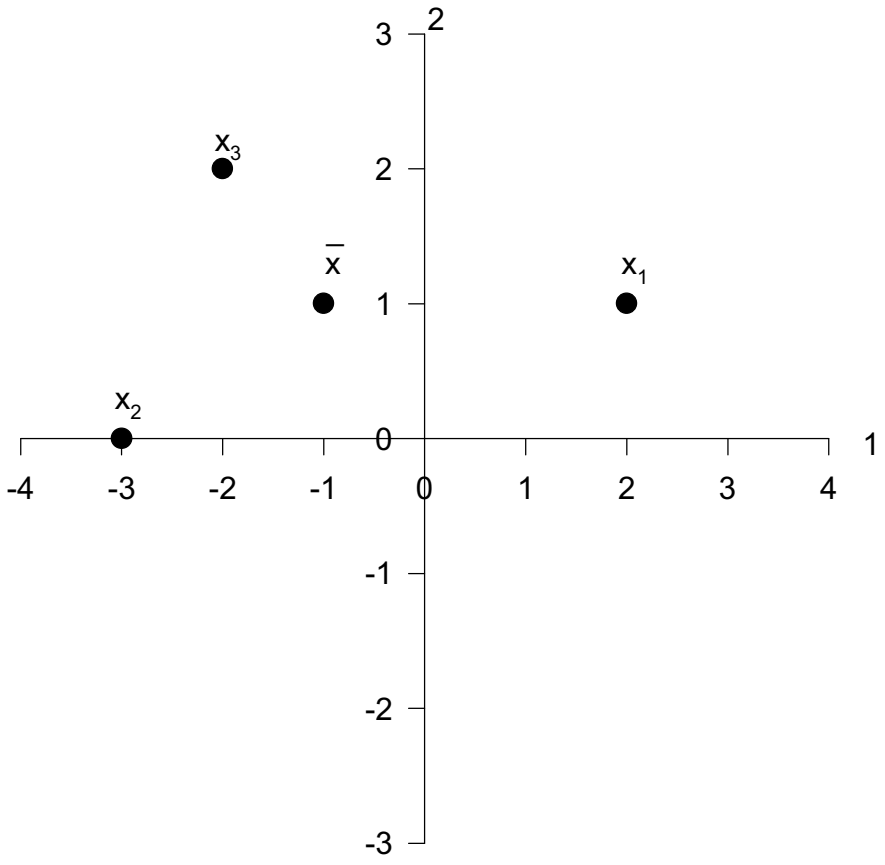


Figura 3.1. Diagrama com n=3 pontos no espaço bidimensional (p=2) mostrando o centro de massa, $\underset{\sim}{\overline{X}}$.

Uma representação alternativa é obtida através da consideração de p pontos no espaço n dimensional. Os elementos das linhas de **X** são utilizados como coordenadas.

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

$$= \begin{bmatrix} \underset{\sim}{y}_1 & \underset{\sim}{y}_2 & \cdots & \underset{\sim}{y}_k & \cdots & \underset{\sim}{y}_p \end{bmatrix}$$

As coordenadas do k-ésimo ponto $\underset{\sim}{y}_k' = \begin{bmatrix} x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix}$ é determinada pela n-upla de todos as medidas da k-ésima variável. É conveniente representar $\underset{\sim}{y}_k'$ como vetor ao invés de pontos.

Exemplo 3.2

Plote os dados da matriz **X**, com p=2 vetores no espaço tridimensional (n=3)

$$X = \begin{bmatrix} 2 & 1 \\ -3 & 0 \\ -3 & 2 \end{bmatrix}$$

$\underset{\sim}{y}_1' = \begin{bmatrix} 2 & -3 & -2 \end{bmatrix}$ $\quad$ $\underset{\sim}{y}_2' = \begin{bmatrix} 1 & 0 & 2 \end{bmatrix}$

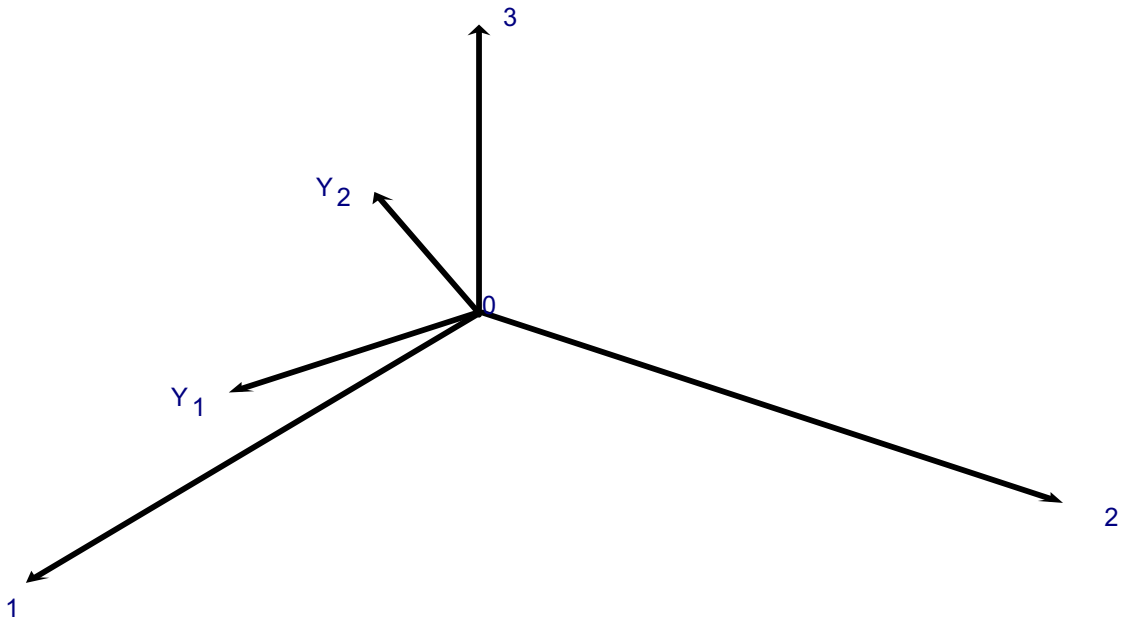


Figura 3.2. Diagrama da matriz de dados **X** como p=2 vetores no espaço tridimensional.

Muita das expressões algébricas que serão encontradas na análise multivariada, podem ser relacionadas às noções geométricas de ângulos, comprimento (norma) e volumes. Isto é importante, pois representações geométricas facilitam a compreensão e conduz a novas visões. Infelizmente, o ser humano está limitado a visualizar objetos no espaço tridimensional, e as representações da matriz **X** não serão úteis se n>3. No entanto, os relacionamentos geométricos e os conceitos estatísticos associados, descritos para o espaço tridimensional ou bidimensional, permanecem válidos para dimensões maiores.

É possível, em função do exposto, prover uma interpretação geométrica ao processo de encontrar a média amostral. O vetor $\underset{\sim}{1}$ (nx1) será definido por $\underset{\sim}{1}$'=[1 1 … 1]. O vetor $\underset{\sim}{1}$ forma um ângulo igual com cada um dos eixos coordenados,

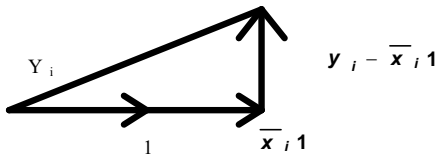
de tal forma que $\left(1/\sqrt{n}\right)\underset{\sim}{1}$ tenha comprimento unitário e mesmo ângulo de direção. Considerando o vetor $\underset{\sim}{y}_i^{'} = \left[x_{1i} \quad x_{2i} \quad \cdots \quad x_{ni}\right]$, cuja projeção em $\left(1/\sqrt{n}\right)\underset{\sim}{1}$ é:

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

Pois, a projeção geral de $\underset{\sim}{x}$ em $\underset{\sim}{y}$ é dada por:

$$\text{Proj}\,(\underset{\sim}{X}\ \text{em}\ \underset{\sim}{Y}) = \frac{\underset{\sim}{x}'\underset{\sim}{y}}{\left|\underset{\sim}{y}\right|}\underset{\sim}{y}$$

Dessa forma $\bar{x}_i = \underset{\sim}{y}_i^{'}\left(\frac{1}{n}\underset{\sim}{1}\right)$ corresponde a um múltiplo de 1, obtido a partir da projeção de $\underset{\sim}{y}_i^{'}$ em um vetor $\underset{\sim}{1}$, de acordo com o esquema a seguir.



em que, $\underset{\sim}{y}_i - \bar{x}_i \underset{\sim}{1}$ é perpendicular a $\bar{x}_i \underset{\sim}{1}$. Observe, também, que $\underset{\sim}{e}_i = \underset{\sim}{y}_i - \bar{x}_i \underset{\sim}{1}$ é definido como desvios da i-ésima variável em relação a sua média amostral, e consiste nos elementos apresentados a seguir:

$$\underset{\sim}{e}_i = \underset{\sim}{y}_i - \bar{x}_i \underset{\sim}{1} = \begin{bmatrix} x_{i1} - \bar{x}_i \\ x_{i2} - \bar{x}_i \\ \\ x_{in} - \bar{x}_i \end{bmatrix},$$

A decomposição de $\underset{\sim}{y}_i$, nos vetores média e desvio da média está apresentada esquematicamente na Figura 3.3 para p=2 e n=3.

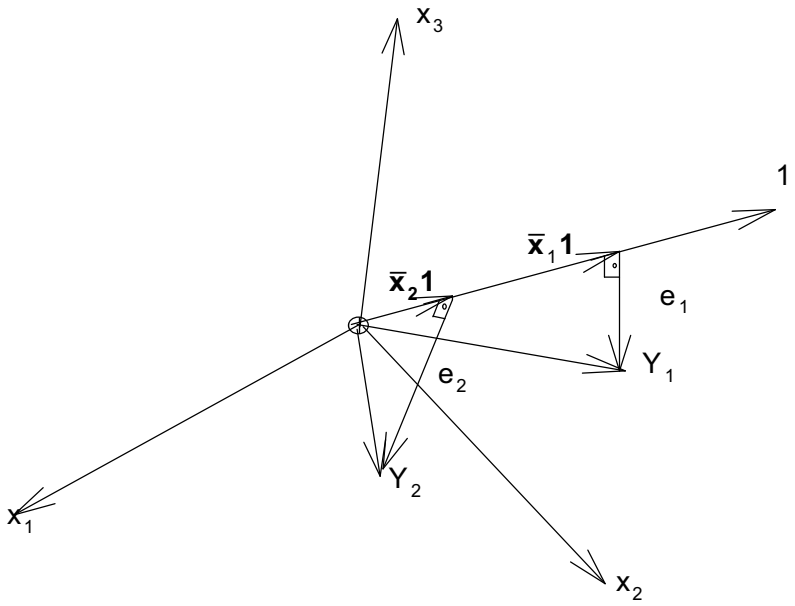


Figura 3.3. Decomposição de $\underset{\sim}{y}_i$ em componentes de média $\bar{x}_i\underset{\sim}{1}$ e componentes de desvio $\underset{\sim}{e}_i = \underset{\sim}{y}_i - \bar{x}\underset{\sim}{1}$.

Exemplo 3.3

Faça a decomposição de $\underset{\sim}{y}_i$ em componentes de média $\bar{x}_i\underset{\sim}{1}$ e componentes de desvio $\underset{\sim}{e}_i = \underset{\sim}{y}_i - \bar{x}\underset{\sim}{1}$, i=1, 2, para os dados do exemplo 3.2.

$$X = \begin{bmatrix} 2 & 1 \\ -3 & 0 \\ -3 & 2 \end{bmatrix} \qquad \underset{\sim}{y}_1' = \begin{bmatrix} 2 & -3 & -2 \end{bmatrix} \qquad \underset{\sim}{y}_2' = \begin{bmatrix} 1 & 0 & 2 \end{bmatrix}$$

$$\bar{x}_1 = \frac{2 + (-3) + (-2)}{3} = -1 \qquad \bar{x}_2 = \frac{1 + 0 + 2}{3} = 1$$

$$\bar{x}_1 \underset{\sim}{1} = -1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} \qquad \bar{x}_2 \underset{\sim}{1} = 1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\underset{\sim}{e}_1 = \underset{\sim}{y}_1 - \bar{x}_1 \underset{\sim}{1} = \begin{bmatrix} 2 \\ -3 \\ -2 \end{bmatrix} - \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \\ -1 \end{bmatrix}$$

$$\underset{\sim}{e}_2 = \underset{\sim}{y}_2 - \bar{x}_2 \underset{\sim}{1} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

Observa-se que: $\bar{x}_1 \underset{\sim}{1}$ e $\underset{\sim}{e}_1$, $\bar{x}_1 \underset{\sim}{1}$ e $\underset{\sim}{e}_2$, são perpendiculares.

$$(\bar{x}_1 \underset{\sim}{1})'(\underset{\sim}{y}_1 - \bar{x}_1 \underset{\sim}{1}) = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix} \times \begin{bmatrix} 3 \\ -2 \\ -1 \end{bmatrix} = -3 + 2 + 1 = 0$$

A decomposição é:

$$\underset{\sim}{y}_1 = \begin{bmatrix} 2 \\ -3 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} + \begin{bmatrix} 3 \\ -2 \\ -1 \end{bmatrix}; \text{ e } \underset{\sim}{y}_2 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}.$$

Os vetores de resíduos podem ser plotados a partir da origem, como apresentado na Figura 3.4, para os resíduos do exemplo 3.3.

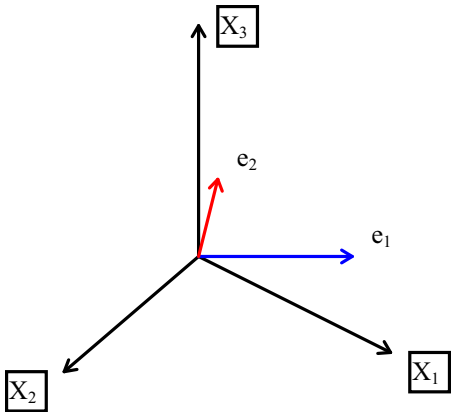


Figura 3.4. Vetores de desvios $\underset{\sim}{e}_i$ do exemplo 3.3.

Considere o comprimento ao quadrado dos vetores de desvios, obtidos por (2.2):

$$|\underset{\sim}{e}_i|^2 = \underset{\sim}{e}_i . \underset{\sim}{e}_i = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 \qquad (3.1)$$

Observa-se por (3.1) que o comprimento ao quadrado dos vetores de desvios é proporcional à variância da i-ésima variável. Equivalentemente, o comprimento é proporcional ao desvio padrão. Vetores longos representam maiores variabilidades que os vetores mais curtos.

Para dois vetores desvios $\underset{\sim}{e}_i$ e $\underset{\sim}{e}_k$:

$$\underset{\sim}{e}_i'\underset{\sim}{e}_k = \sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) \tag{3.2}$$

De (2.3) e denotando o ângulo θ_{ik} como o ângulo formado pelos vetores $\underset{\sim}{e}_i$ e $\underset{\sim}{e}_k$, tem-se:

$$\mathrm{Cos}(\theta_{ik}) = \frac{\underset{\sim}{e}_i'\underset{\sim}{e}_k}{\sqrt{\underset{\sim}{e}_i'\underset{\sim}{e}_i} \times \sqrt{\underset{\sim}{e}_k'\underset{\sim}{e}_k}} \tag{3.3}$$

Usando (3.1) e (3.2) é fácil verificar que (3.3) é:

$$r_{ik} = \mathrm{Cos}(\theta_{ik}) = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} \tag{3.4}$$

O coseno do ângulo formado entre dois vetores desvios é igual ao coeficiente de correlação amostral. Portanto, se os dois vetores de desvios possuem a mesma orientação, o coeficiente de correlação será próximo de 1. Se os dois vetores

estão próximos de serem perpendiculares, a correlação amostral será próxima de zero. Se os dois vetores forem orientados em direções opostas, o coeficiente de correlação amostral será próximo de -1. Os conceitos de comprimento e ângulos permitem que se faça interpretações das estatísticas amostrais geometricamente, e auxiliam na compreensão dos seus significados.

3.3. Amostras aleatórias e esperanças do vetor média e da matriz de covariância amostral.

Com a finalidade de estudar a variabilidade amostral de estatísticas como $\bar{\underset{\sim}{X}}$ e **S** com a finalidade de se fazer inferências, é necessário fazer pressuposições a respeito das variáveis cujos valores observados constituem um conjunto de dados **X**.

Supondo que os dados não foram ainda observados, mas pretende-se obter n mensurações em p variáveis. Antes de serem mensurados, os valores não podem em geral ser preditos exatamente. Conseqüentemente, estes são tratados como variáveis aleatórias. Neste contexto, os elementos (j, k) da matriz de dados representam realizações de uma variável aleatória, $\mathbf{X}_{jk}$. Cada conjunto de medidas $\underset{\sim}{X}_j$ em p variáveis é um vetor aleatório.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}_1' \\ \underset{\sim}{X}_2' \\ \vdots \\ \underset{\sim}{X}_j' \\ \vdots \\ \underset{\sim}{X}_n' \end{bmatrix} \tag{3.5}$$

Uma amostra aleatória pode ser definida por: “Se o vetor coluna $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ em (3.5), representa independentes observações com distribuição conjunta com densidade $f(\underset{\sim}{x})=f(x_1, x_2, ..., x_p)$, então $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ é uma amostra aleatória. Se a função conjunta de densidade é igual ao produto das marginais $f(\underset{\sim}{x}_1) . f(\underset{\sim}{x}_2) . ..., . f(\underset{\sim}{x}_n)$, sendo $f(\underset{\sim}{x}_j)=f(x_{j1}, x_{j2}, ..., x_{jp})$, então, $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ é uma amostra aleatória.”

Algumas conclusões podem ser obtidas da distribuição de $\overline{\underset{\sim}{X}}$ e **S** sem pressuposições sobre a forma da distribuição conjunta das variáveis. Dessa forma, considere $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ como uma amostra aleatória de uma distribuição conjunta com vetor média μ e matriz de covariância Σ. então, $\overline{\underset{\sim}{X}}$ é um estimador não viciado de $\underset{\sim}{\mu}$ e sua matriz de covariância é $\frac{1}{n}\Sigma$. Isto é,

$E(\overline{\underset{\sim}{X}}) = \underset{\sim}{\mu}$ (vetor média populacional)

$Cov(\overline{\underset{\sim}{X}}) = \frac{1}{n}\Sigma$ (Matriz de covariância populacional dividida pelo tamanho da amostra).

PROVA:

$$\overline{\underset{\sim}{X}} = (\underset{\sim}{X}_1 + \underset{\sim}{X}_2 + ... + \underset{\sim}{X}_n)/n$$

$$E(\underset{\sim}{\bar{X}}) = \left(\tfrac{1}{n}\underset{\sim}{X}_1 + \tfrac{1}{n}\underset{\sim}{X}_2 + \cdots + \tfrac{1}{n}\underset{\sim}{X}_n\right)$$

$$= E\left(\tfrac{1}{n}\underset{\sim}{X}_1\right) + E\left(\tfrac{1}{n}\underset{\sim}{X}_2\right) + \cdots + E\left(\tfrac{1}{n}\underset{\sim}{X}_n\right)$$

$$= \frac{1}{n}\left[nE\left(\underset{\sim}{X}_j\right)\right] = \frac{1}{n} \times n \times \underset{\sim}{\mu}$$

$$\therefore E(\underset{\sim}{\bar{X}}) = \underset{\sim}{\mu}$$

Para provar o valor da covariância, pode-se observar que:

$$(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu})\,(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu})' = \left(\frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)\right) \times \left(\frac{1}{n}\sum_{\ell=1}^{n}\left(\underset{\sim}{X}_\ell - \underset{\sim}{\mu}\right)\right)' = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)\left(\underset{\sim}{X}_\ell - \underset{\sim}{\mu}\right)'$$

Então,

$$\mathrm{Cov}\left(\underset{\sim}{\bar{X}}\right) = E\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}\right)\left(\underset{\sim}{\bar{X}} - \underset{\sim}{\mu}\right)' = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}E\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)\left(\underset{\sim}{X}_\ell - \underset{\sim}{\mu}\right)'$$

Sendo $j \neq \ell$ e considerando que $E\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)\left(\underset{\sim}{X}_\ell - \underset{\sim}{\mu}\right)'$ é igual a zero, devido a covariância entre os elementos independentes $\underset{\sim}{X}_j$ e $\underset{\sim}{X}_\ell$ ser nula, então,

$$\mathrm{Cov}\left(\underset{\sim}{\bar{X}}\right) = \frac{1}{n^2}\sum_{j=1}^{n}E\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)'$$

Desde que $\Sigma = E\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)'$ é a covariância populacional comum dos componentes $\underset{\sim}{X}_j$, têm-se:

$$\mathrm{Cov}\left(\underset{\sim}{\bar{X}}\right) = \frac{1}{n^2}\sum_{j=1}^{n} E\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)\left(\underset{\sim}{X}_j - \underset{\sim}{\mu}\right)' = \frac{1}{n^2}\left(\Sigma + \Sigma + \cdots + \Sigma\right) =$$

$$= \frac{1}{n^2}(n\Sigma) = \frac{1}{n}\Sigma$$

3.4. Variância Generalizada

Com uma única variável, a variância da amostra é usada para descrever a variação nas mensurações desta variável. Quando p variáveis são observadas em cada unidade da amostra ou do experimento, a variação é descrita pela matriz de variância e covariância amostral.

$$\mathbf{S} = \begin{bmatrix} \mathbf{S_{11}} & \mathbf{S_{12}} & \cdots & \mathbf{S_{1p}} \\ \mathbf{S_{21}} & \mathbf{S_{22}} & \cdots & \mathbf{S_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S_{p1}} & \mathbf{S_{p2}} & \cdots & \mathbf{S_{pp}} \end{bmatrix}$$

A matriz de covariância amostral contém p variâncias e ½p(p-1) covariâncias, potencialmente diferentes. Algumas vezes, no entanto, deseja-se expressar a variação por um único valor numérico. Uma escolha deste valor é o

determinante de **S**, o qual reduz à variância amostral usual para o caso de uma única variável (p=1). Este determinante é denominado de variância amostral generalizada.

Variância amostral Generalizada=|**S**| (3.6)

Exemplo 3.4

O peso de espiga PE ($\mathbf{X}_1$), e o número de espigas NE ($\mathbf{X}_2$), foi avaliado em 28 variedades de milho em Sete Lagoas, MG. A matriz de covariância amostral **S**, obtida dos dados é:

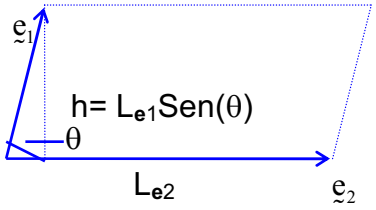
$$\mathbf{S}=\begin{bmatrix} \mathbf{2,905} & \mathbf{9,096} \\ \mathbf{9,096} & \mathbf{90,817} \end{bmatrix}$$

A variância generalizada neste caso é:

Variância amostral Generalizada = |**S**| = 2,905x90,817 - $9,096^2$ = 181,0862

A variância amostral generalizada se constitui numa forma de escrever toda a informação de todas as variâncias e covariâncias como um único valor numérico. Obviamente, quando p>1 é possível que algumas informações amostrais sejam perdidas no processo. A interpretação geométrica, no entanto, poderá mostrar a força e as fraquezas desta estatística descritiva.

Considerando-se o volume (área) gerado no plano definido por dois vetores de desvios $\underset{\sim}{e}_1 = \underset{\sim}{Y}_1 - \bar{x}_1\underset{\sim}{1}$ e $\underset{\sim}{e}_2 = \underset{\sim}{Y}_2 - \bar{x}_2\underset{\sim}{1}$. Seja L_{e1} e L_{e2} os comprimentos dos vetores $\underset{\sim}{e}_1$ e $\underset{\sim}{e}_2$, respectivamente. Da geometria têm-se:



A área do trapezóide é L_{e1} x Sen(θ) x L_{e2}, podendo ser expressa por:

$$\text{Área} = L_{e1} L_{e2} \sqrt{1 - \cos^2(\theta)}$$

Mas,

$$L_{e1} = \sqrt{\sum_{j=1}^{n} (X_{j1} - \bar{X}_1)^2} = \sqrt{(n-1)S_{11}}$$

$$L_{e2} = \sqrt{\sum_{j=1}^{n} (X_{j2} - \bar{X}_2)^2} = \sqrt{(n-1)S_{22}}$$

$\text{Cos}(\theta) = r_{12}$

Portanto,

$$\text{Área} = (n-1)\sqrt{S_{11}S_{22}\left(1-r_{12}^{2}\right)} \tag{3.7}$$

Por outro lado,

$$|S| = \begin{vmatrix} S_{11} & S_{21} \\ S_{12} & S_{22} \end{vmatrix} = \begin{vmatrix} S_{11} & \sqrt{S_{11}}\sqrt{S_{22}}r_{12} \\ \sqrt{S_{11}}\sqrt{S_{22}}r_{12} & S_{22} \end{vmatrix}$$

$$= S_{11}S_{22} - S_{11}S_{22}r_{12}^{2} = S_{11}S_{22}(1-r_{12}^{2}) \tag{3.8}$$

Se (3.7) e (3.8) forem comparados, pode-se observar que

$$|\mathbf{S}| = (\text{Área})^2/(n-1)^2$$

Esta expressão pode ser generalizada para p vetores desvios por indução:

$$\text{Variância amostral Generalizada} = |\mathbf{S}| = (\text{Volume})^2.(n-1)^{-p} \tag{3.9}$$

A equação (3.9) mostra que a variância amostral é proporcional ao quadrado do volume gerado pelos p vetores desvios. A Figura 3.5 (a) e (b) mostra as regiões trapezoidais geradas com p=3 vetores resíduos, correspondentes a “grandes” e “pequenas” variâncias amostrais generalizadas, respectivamente.

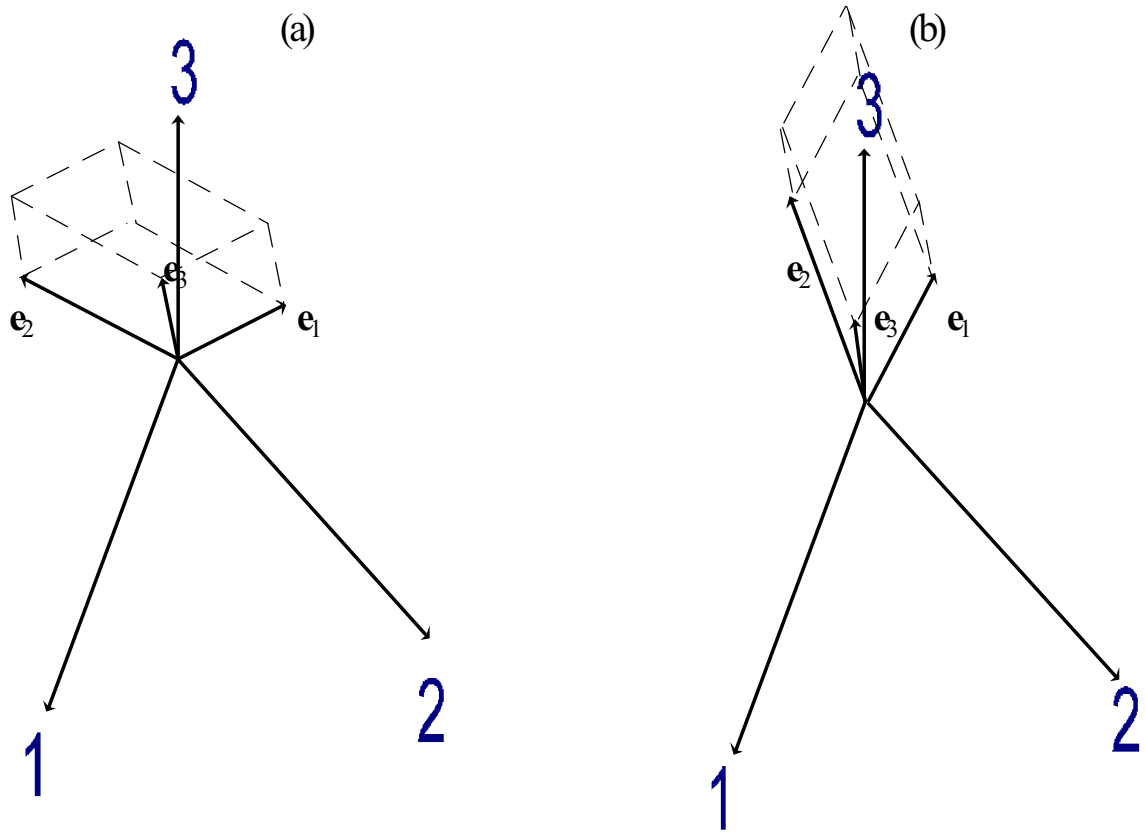


Figura 3.5. (a) grande variância amostral generalizada, e (b) pequena variância amostral generalizada, para p=3.

Para um tamanho amostral fixo, é óbvio que |S| cresce com o aumento do comprimento dos vetores de desvios $\underset{\sim}{e}_i$ (ou $\sqrt{(n-1)S_{ii}}$). Em adição, o volume aumentará para um comprimento fixado, se os vetores residuais forem movidos até possuírem ângulos retos. Por outro lado se um ou mais dos vetores residuais aproximar do hiperplano formado por outros vetores residuais, o volume diminuirá tendendo a zero.

Apesar da variância amostral generalizada possuir algumas interpretações geométricas formidáveis como as ilustradas na Figura 3.5, ela sofre alguns problemas

como estatística amostral capaz de sumarizar a informação contida na matriz **S**. Para ilustrar estas deficiências, considere as matrizes de covariâncias e os coeficientes de correlações apresentados a seguir.

$$S = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix} \qquad S = \begin{bmatrix} 10 & -8 \\ -8 & 10 \end{bmatrix} \qquad S = \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}$$

$$r_{12} = \frac{8}{\sqrt{10}\sqrt{10}} = 0,8 \qquad r_{12} = \frac{-8}{\sqrt{10}\sqrt{10}} = -0,8 \qquad r_{12} = \frac{0}{\sqrt{6}\sqrt{6}} = 0,0$$

$$|S| = 36 \qquad |S| = 36 \qquad |S| = 36$$

Apesar das três matrizes possuírem a mesma variância amostral generalizada (|**S**|=36), elas possuem estruturas de correlações distintas. Portanto, diferentes estruturas de correlações não são detectadas pela variância amostral generalizada. As situações em que p>2 podem ser ainda mais obscuras.

Muitas vezes é desejável mais informações do que um simples valor como |**S**| pode oferecer como resumo de **S**. Pode-se mostrar que |**S**| pode ser expresso como produto dos autovalores de **S** (|**S**|=$\lambda_1.\lambda_2....\lambda_p$). A elipsóide centrada na média é baseada em $\mathbf{S}^{-1}$, possui eixos de comprimento proporcionais a raiz quadrada de λ_i's de **S**, que reflete a variabilidade no sentido do i-ésimo autovalor. Esta elipsóide é apresentada a seguir.

$$\left(\underset{\sim}{X} - \overline{\underset{\sim}{X}}\right)' S^{-1} \left(\underset{\sim}{X} - \overline{\underset{\sim}{X}}\right) = c^2 \qquad (3.10)$$

Demonstra-se que o volume desta hiperelipsóide é proporcional à raiz quadrada de |**S**|. Desta forma, os autovalores, fornecem informações da variabilidade em todas as direções da representação no espaço p-dimensional dos dados. Portanto, é mais útil apresentar seus valores individuais do que seu produto. Este tópico será abordado com mais detalhe quando se discutir sobre os componentes principais.

A variância amostral generalizada será zero se um dos vetores residuais pertencer a um (hiper) plano formado por uma combinação linear dos outros, ou seja, quando as linhas da matriz de desvios, forem linearmente dependentes.

Exemplo 3.5

Mostre que |**S**|=0 para

$$X = \begin{bmatrix} 3 & 3 & 6 \\ 1 & 3 & 4 \\ 2 & 0 & 2 \end{bmatrix}$$

O vetor média é:

$$\underset{\sim}{\bar{X}}' = \begin{bmatrix} 2 & 2 & 4 \end{bmatrix}$$

Os vetores dos desvios são:

$$\underset{\sim}{X} - \underset{\sim}{1}\underset{\sim}{\bar{X}}' = \begin{bmatrix} \underset{\sim}{e}_1 & \underset{\sim}{e}_2 & \underset{\sim}{e}_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ -1 & 1 & 0 \\ 0 & -2 & -2 \end{bmatrix}$$

Verifica-se que $\underset{\sim}{e}_3' = \underset{\sim}{e}_1' + \underset{\sim}{e}_2'$, ou seja:

[2 0 -2] = [1 -1 0] +[1 1 -2] = [2 0 -2] c.q.d.

Isto significa que um dos vetores resíduos, pertence ao plano gerado pelos outros dois. Desta forma o volume tridimensional é zero (degenerescência). Este caso é ilustrado na Figura 3.6 e demonstrado numericamente através da obtenção de |**S**|.

$$\mathbf{S} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{3} & \mathbf{3} \\ \mathbf{1} & \mathbf{3} & \mathbf{4} \end{bmatrix}$$

Pela definição (2.9), têm-se:

$$|S| = 1 \times \begin{vmatrix} 3 & 3 \\ 3 & 4 \end{vmatrix} \times (-1)^2 + 0 \times \begin{vmatrix} 0 & 1 \\ 3 & 4 \end{vmatrix} \times (-1)^3 + 1 \times \begin{vmatrix} 0 & 1 \\ 3 & 3 \end{vmatrix} \times (-1)^4 =$$

$$= 1.3.1 + 0 + 1.(-3).1 = 3 - 3 = 0$$

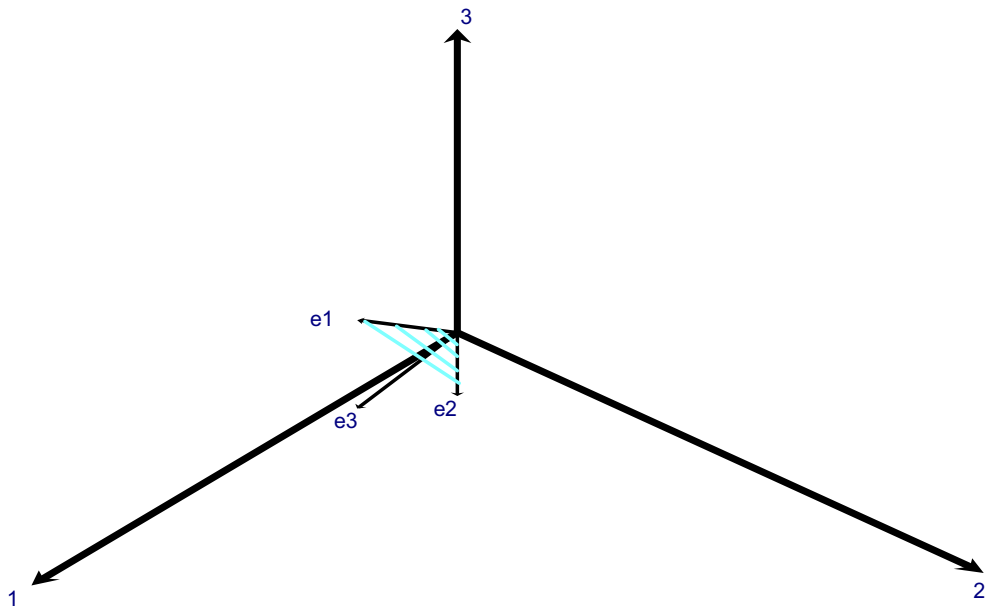


Figura 3.6 Caso em que |**S**|=0 (degenerescência) para o volume tridimensional.

Em qualquer análise estatística o resultado |**S**|=0 indica que existem variáveis redundantes, ou seja, que possuem a mesma informação, e que estas podem ser removidas do estudo. A matriz de covariância reduzida, será de posto completo e a variância generalizada diferente de zero. A questão de quais variáveis remover no caso de degenerescência não é fácil de responder e será abordado nos estudos de componentes principais. No entanto, quando há possibilidade de escolha, o pesquisador deve reter as medidas de uma variável (presumidamente) causal ao invés de uma característica secundária.

3.5.Variância generalizada de variáveis padronizadas

A variância amostral generalizada é influenciada pela diferença de variabilidade das mensurações das variáveis individuais, ou seja, caso a variância amostral de uma determinada variável (S_{ii}) seja grande ou pequena em relação às demais. O vetor residual correspondente $\underset{\sim}{e}_i = \underset{\sim}{Y}_i - \overline{x}_i \underset{\sim}{1}$ será muito longo ou muito curto, do ponto de vista geométrico e terá um papel importante na determinação do volume. É muitas vezes necessário, em função do exposto, padronizar os vetores residuais, de tal forma que eles tenham o mesmo comprimento.

A padronização destes vetores residuais é equivalente a transformar as variáveis originais x_{jk} pelos seus valores $\left(x_{jk} - \overline{x}_k\right) \Big/ \sqrt{S_{kk}}$. A matriz de covariância amostral das variáveis padronizadas será então igual a **R**, ou seja, igual a matriz de correlação das variáveis originais. Dessa forma pode-se definir:

Variância generalizada amostral das variáveis padronizadas=|**R**| (3.11)

Os vetores resíduos resultantes, cujos valores são dados por $\mathbf{e_{jk}} = \left(x_{jk} - \overline{x}_k\right) \Big/ \sqrt{S_{kk}}$, possuem todos comprimento igual a $\sqrt{\mathbf{n-1}}$. A variância generalizada amostral das variáveis padronizadas será grande se estes vetores forem perpendiculares e pequena se dois ou mais deles tiverem próximas da mesma direção. Em (3.4) foi visto que o coseno do ângulo θ_{ik} entre os vetores residuais $\underset{\sim}{e}_i$ e $\underset{\sim}{e}_k$, com $i \neq k$, é igual ao coeficiente de correlação amostral r_{ik}. Dessa forma, o |**R**| será grande quando todos os r_{ik} forem próximos de zero e será pequeno quando um ou mais dos r_{ik} for próximo de -1 ou de +1.

Utilizando os mesmos argumentos que conduziram a (3.9) pode-se verificar que:

$$|\mathbf{R}|=(n-1)^{-p}(\text{volume})^2 \qquad (3.12)$$

O volume gerado pelos vetores desvios de p=3 variáveis padronizadas está ilustrado na Figura 3.7. Estes vetores desvios padronizados são correspondentes aos vetores desvios da Figura 3.5, cuja comparação revela que a influência do vetor $\underset{\sim}{e}_2$ (com grande variabilidade na direção de x_2) no volume quadrado de $|\mathbf{S}|$ é maior do que sua influência no volume quadrado de $|\mathbf{R}|$.

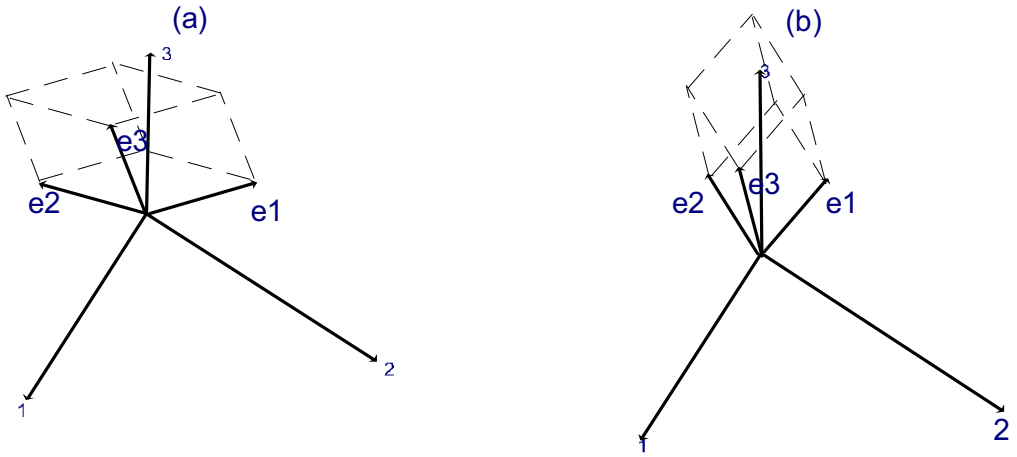


Figura 3.7. Volume gerado por três variáveis padronizadas: (a) grande variância e (b) pequena variância generalizada.

As quantidades $|\mathbf{S}|$ e $|\mathbf{R}|$ são relacionadas por:

$$|\mathbf{S}| = (S_{11}\, S_{22} \ldots S_{pp})\, |\mathbf{R}| \qquad (3.13)$$

Exemplo 3.6

É ilustrada através deste exemplo a relação (3.13) entre |**S**| e |**R**| para p=3 caracteres de milho (x_1: diâmetro do colmo; x_2: número de folhas; e x_3: comprimento de folhas). A matriz **R** e **S** obtidas são:

$$\mathbf{S} = \begin{bmatrix} 4{,}935 & 0{,}552 & 2{,}921 \\ 0{,}552 & 0{,}686 & 1{,}932 \\ 2{,}921 & 1{,}932 & 17{,}993 \end{bmatrix} \quad \mathbf{e} \quad \mathbf{R} = \begin{bmatrix} 1{,}00 & 0{,}30 & 0{,}31 \\ 0{,}30 & 1{,}00 & 0{,}55 \\ 0{,}31 & 0{,}55 & 1{,}00 \end{bmatrix}$$

Usando-se a definição de determinante (2.9), tem-se:

|**S**|=37,3878

|**R**|=0,6137

Usando (3.13) e os resultados obtidos:

$|\mathbf{S}| = (S_{11}\, S_{22} \ldots S_{pp})\, |\mathbf{R}|$

37,3878 = (4,935 x 0,686 x 17,993) x 0,6137

$37,3878 \approx 37,3828$ (verificado, apesar da pequena diferença devido às aproximações nos cálculos)

3.6. Outra generalização da variância

Uma outra medida capaz de sintetizar a informação contida na matriz de covariância que é utilizada em componentes principais é definida pela soma dos elementos da diagonal da matriz de covariância **S** e é denominada de variância amostral total. Portanto,

Variância amostral total = Traço de **S=** $Tr(S) = S_{11}+S_{22}+...+S_{pp}$ (3.14)

Exemplo 3.7

Calcular a variância amostral total da matriz **S** do exemplo (3.6)

$Tr(\mathbf{S}) = S_{11}+S_{22}+S_{33}=4,935+0,686+17,993=23,614$

Geometricamente a variância amostral total representa a soma dos comprimentos ao quadrado dos vetores residuais $\underset{\sim}{e}_i$ (i=1, 2, ...,p) dividido por n-1. Ela não considera as orientações dos vetores residuais, sendo portanto limitada para ser

utilizada com variáveis padronizadas, pois seu valor será sempre o mesmo para distintos conjuntos de dados desde que o número de variáveis destes seja igual.

3.7. Exercícios

3.7.1. Plote os n=4 pontos no diagrama bidimensional e localize $\underset{\sim}{\bar{X}}$ no diagrama resultante.

$$X = \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}$$

3.7.2. Encontre o ângulo entre os vetores $\underset{\sim}{y}_1$ e $\underset{\sim}{y}_2$ do exemplo 3.1. Calcule o coseno do mesmo e discuta sobre o significado deste resultado.

3.7.3. Obtenha a decomposição dos vetores $\underset{\sim}{y}_1$ e $\underset{\sim}{y}_2$ do exemplo 3.1 em componente de média e componente de desvio. Comprove a ortogonalidade dos componentes de média com os vetores de desvios ou residuais.

3.7.4. Calcule usando (3.3) o coseno do ângulo entre os vetores residuais $\underset{\sim}{e}_1$ e $\underset{\sim}{e}_2$ obtidos em 3.3. Calcule o coeficiente de correlação usando (1.4) entre as variáveis 1 e 2, e compare os resultados obtidos.

3.7.5. Obtenha as matrizes de covariância amostral para o conjunto de dados do exercício 3.7.1, e calcule as variâncias amostrais generalizadas das variáveis originais e padronizadas. Calcule também a variância amostral total.

3.7.6. Qual é a área do trapezóide gerado pelos p=2 vetores desvios, do exercício 3.7.1.

3.8. Referências

BOCK, R.D. **Multivariate statistical methods in behavioral research.** McGrawn Hill, 1975.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 4th edition. Prentice Hall, New Jersey, 1998. 816p.