Disciplina: Análise Multivariada I

Prof. Dr. Admir Antonio Betarelli Junior

AULA 1

1 ANÁLISE MULTIVARIADA

A investigação científica é um processo interativo, que, muitas vezes, envolve a coleta e análise de um grande conjunto de dados para explicar o fenômeno de interesse, ou até mesmo sugere modificações nas interpretações do fenômeno. Contudo, devido à complexidade destes fenômenos, o pesquisador adiciona ou suprimi variáveis em seus estudos, algumas vezes por causa dos problemas ou complexidades gerados pela simultaneidade de múltiplas variáveis. Diante dessas complexidades, esta disciplina está preocupada com os métodos estatísticos designados para extrair informações a partir desses tipos de conjuntos de dados. Como os dados incluem medições simultâneas de muitas variáveis, este corpo de metodologia é chamado de análise multivariada. A necessidade de compreender as relações entre muitas variáveis faz com que as análises multivariadas sejam um assunto complexo ou inerentemente difícil.

Por conceito, a Análise Multivariada refere-se a um conjunto de métodos estatísticos que torna possível a análise simultânea de medidas múltiplas para cada indivíduo, objeto ou fenômeno observado. Por realizar análise simultânea de mais de duas variáveis para cada observação da amostra, os métodos podem ser considerados como integrantes da Análise Multivariada. Em geral, as observações são correlacionadas e quanto maior o número de variáveis, mais complexa é a análise univariada. Ademais, as variáveis selecionadas para cada observação podem ser quantitativas (discretas ou contínuas) ou qualitativas (ordinais ou nominais). O truque na da estatística multivariada consiste em escolher o método apropriado ao tipo de dados, e usá-lo corretamente, bem como saber interpretar os resultados e retirar deles as conclusões corretas (REIS, 2001).

Na disciplina serão discutidas técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados, algumas vezes em aplicações na economia.

Os objetivos mais gerais do emprego de técnicas multivariadas são:

a) redução de dados ou simplificação estrutural: a partir de correlação ou associação das variáveis originais, busca-se construir índices ou variáveis alternativas que sintetizam as informações originais, sem sacrificar informações valiosas e que tornam as interpretações mais simples. Por exemplo: Análise de Componentes Principais (ACP), Análise Fatorial (AF), Análise de Correlação Canônica (ACC) ou Análise de Correspondência Múltipla (ACM);

b) classificação e discriminação: criam-se grupos de objetos ou variáveis similares, baseados em dados amostrais ou experimentais. Para tanto, utilizam-se as técnicas de análise de cluster (AA), quando a divisão da população não é conhecida a *priori*, ou análise discriminante (AD), quando já se detém conhecimento prévio sobre os possíveis grupos a fim de classificar um elemento amostral;

c) investigação de relação entre as variáveis: com o auxílio de técnicas multivariadas busca-se investigar a natureza da relação ente as variáveis, ou seja, se as mesmas são mutuamente independentes ou uma ou mais são dependentes de outras. Técnicas como regressão múltipla, regressão logística, modelagem de equações estruturais, dentre outras, são úteis para atingir esse objetivo.

A utilização adequada da análise multivariada depende do bom conhecimento das técnicas e das suas limitações. Como afirma Marriot (1974): "se os resultados divergirem com a opinião formada, impedirem uma simples interpretação lógica, não estiverem claramente em uma apresentação gráfica, logo os mesmos estariam provavelmente errados. [...] Os métodos não devem ser utilizados como máquinas automáticas de encher linguiça, transformando massas numéricas em pacotes de fatos científicos".

Feitas essas considerações iniciais, tona-se oportuno incialmente apresentar os conceitos e propriedades mais tradicionais da Análise Multivariada.

2 CONCEITOS BÁSICOS

2.1 Matriz de informação

É representada por uma matriz $\underset{nxp}{\mathbf{X}}$, com n elementos amostrais (observações) e $p > 1$ variáveis aleatórias ou características:

$$\underset{nxp}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} \forall \quad k = 1,2,..,p; \quad j = 1,2,...,n.$$

em que usa-se a notação x_{jk} para indicar o valor da *k-ésima* variável observada no *j-ésimo* elemento (item, objeto, indivíduo, fenômeno, ...). A partir desta matriz de informação, $\underset{nxp}{\mathbf{X}}$, pode-se simplificar, definindo o vetor aleatório, cujos elementos são as variáveis aleatórias:

$$\mathbf{X}' = \begin{bmatrix} X_1 & X_2 & \cdots & X_k & \cdots & X_p \end{bmatrix}$$

Nos extremos, o vetor $\mathbf{X}$ pode consistir em *n* observações em apenas uma variável, ou de uma observação multivariada em *p* variáveis. Aliás, quando se tem um vetor aleatório, cada variável pode ser analisada separadamente. Contudo, vale a pena analisá-lo como um todo, pois nele pode ter associações entre as *p*-variáveis.

2.2 Estatísticas descritivas

As estatísticas descritivas fornecem um valor central, a variabilidade e associação linear para o conjunto de dados.

2.2.1 Vetor de médias (ou esperança):

Sendo $\mathbf{X}$ um vetor aleatório, pode-se calcular a média μ_k para sintetizar a informação de tendência central da distribuição de x_k.

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix} = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

$$E(X_k) = \overline{X}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} .$$

sendo $\overline{X}_k$ também a média amostral, cujo vetor é $\overline{\mathbf{X}}' = \begin{bmatrix} \overline{X}_1 & \cdots & \overline{X}_p \end{bmatrix}$.

Lembre-se que $\mu_k = \int_{-\infty}^{\infty} x_k f_k(x_k) dx_k$, se for variável contínua com função densidade de probabilidade $f_k(x_k)$; e $\mu_k = \sum_{\forall k} x_k p_k(x_k)$ se for variável discreta com função de probabilidade $p_k(x_k)$. Essa diferença vale para as demais medidas estatísticas, porém elas não serão apresentadas.

2.2.2 Matriz de variância-covariância

As *p* variâncias e *p(p-1)/2* covariâncias são contidas em uma matriz simétrica:

a) <u>Populacional:</u>

$$\underset{\mathbf{pxp}}{\boldsymbol{\Sigma}} = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' = E\left[\begin{pmatrix} X_1-\mu_1 \\ \vdots \\ X_P-\mu_P \end{pmatrix}\begin{pmatrix} X_1-\mu_1 & \cdots & X_P-\mu_P \end{pmatrix}\right] = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$Var(X_k) = \sigma_k^2 = \sigma_{kk} = E(X_k - \mu_k)^2 ;$$

$\underset{k \neq i}{COV}(X_i X_k) = \sigma_{ik} = E[(X_i - \mu_i)(X_k - \mu_k)]$. É difícil julgar se a relação é forte ou não, bem como é sensível à escala.

b) <u>*Amostral*</u> (estimativa de $\underset{pxp}{\Sigma}$): representa uma amostra de $\underset{nxp}{\mathbf{X}}$, logo as matrizes precisam ser estimadas.

$$\underset{\mathbf{pxp}}{\mathbf{S}} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})' = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}$$

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

Veja a demonstração para o estimador não enviesado de σ_{ik} em Johnson e Wichern (2002, p.122-123).

c) *Propriedades gerais:*

i. Simétrica: $\mathbf{\Sigma} = \mathbf{\Sigma}'$ ou $\sigma_{ik} = \sigma_{ki}$, necessariamente quadrática. Logo, tem-se uma consequência direta para composição espectral. Para verificar isso, calculam-se os autovalores e correspondentes autovetores;

ii. Pode ser não negativa definida (n.n.d.), i.e., $\mathbf{a}'\mathbf{\Sigma}\mathbf{a} \geq 0, \forall \underset{px1}{a} \neq 0$. Todos os menores principais são não negativos. Seus p autovalores são não negativos ($\lambda_k \geq 0, \quad \forall k = 1,2,...,p$).

iii. Pode ser positiva definida (p.d.), i.e., $\mathbf{a}'\mathbf{\Sigma}\mathbf{a} > 0, \forall \underset{px1}{a} > 0$. Todos os menores principais são positivos. Seus p autovalores são positivos. Veja Simon e Blume (2004, p.389-395).

d) *Exemplo 1: testando as Propriedades gerais para* $\underset{2\mathbf{x}2}{\mathbf{\Sigma}} = \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix}$:

- i. É simétrica, pois $\mathbf{\Sigma} = \mathbf{\Sigma}'$ ou $\sigma_{ik} = \sigma_{ki} = -2$.
- ii ou iii.

Autovalores (λ_k): resumem as propriedades essenciais e são valores característicos da matriz: $\det(\mathbf{\Sigma} - \lambda I) = 0$

$$\det\begin{bmatrix} 8-\lambda & -2 \\ -2 & 5-\lambda \end{bmatrix} = 0 \Rightarrow (8-\lambda)(5-\lambda) - 4 = 0 \Rightarrow (\lambda_1 = 9, \lambda_2 = 4)$$

Todos os autovalores são positivos. Para maiores detalhes, veja Johnson e Wichern (2002, p. 63-65).

Autovetores (e_k): para cada autovalor, tem-se um respectivo vetor positivo se: $\boldsymbol{\Sigma} e_k = \lambda_k e_k \;\Rightarrow\; (\boldsymbol{\Sigma} - \lambda I) e_k = 0$.

Para $\lambda_1 = 9$: $\begin{bmatrix} 8-9 & -2 \\ -2 & 5-9 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{cases} -a-2b=0 \\ -2a+b=0 \end{cases} \Rightarrow a = -2b \Rightarrow e_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$

Para $\lambda_2 = 4$: $\begin{bmatrix} 8-4 & -2 \\ -2 & 5-4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow a = \frac{b}{2} \Rightarrow e_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

Ambos autovetores não estão normalizados (comprimento unitário). Por sua vez, os menores principais são:

$\det(\boldsymbol{\Sigma}_1) = 8$ e $\det(\boldsymbol{\Sigma}_2) = 36$. Portanto, $\boldsymbol{\Sigma}$ é uma matriz positiva definida (iii).

A condição de $\boldsymbol{\Sigma}$ como n.n.d. implica que as combinações lineares construídos do vetor $\underset{nxp}{\mathbf{X}}$ são sempre não negativas. Isso permite que se construam novas variáveis definidas em termos estatísticos.

2.3 Particionamento da matriz de Covariância

Uma abordagem para medir as características de grupos distintos é considerá-lo como subconjunto no total de coleções de características:

$$\underset{1xp}{\mathbf{X}'} = \left[\underset{1xq}{\mathbf{X}^{(1)}} \vdots \underset{1x(p-q)}{\mathbf{X}^{(2)}} \right], \underset{1xp}{\boldsymbol{\mu}'} = \left[\underset{1xq}{\boldsymbol{\mu}^{(1)}} \vdots \underset{1x(p-q)}{\boldsymbol{\mu}^{(2)}} \right], \underset{\mathbf{pxp}}{\boldsymbol{\Sigma}} = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' = \begin{bmatrix} \underset{qxq}{\boldsymbol{\Sigma}_{11}} & \vdots & \underset{qx(p-q)}{\boldsymbol{\Sigma}_{12}} \\ \cdots & \cdots & \cdots \\ \underset{(p-q)xq}{\boldsymbol{\Sigma}_{21}} & \vdots & \underset{(p-q)x(p-q)}{\boldsymbol{\Sigma}_{22}} \end{bmatrix}$$

em que $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21}$. A matriz de covariâncias de $\mathbf{X}^{(1)}$ é $\boldsymbol{\Sigma}_{11}$, de $\mathbf{X}^{(2)}$ é $\boldsymbol{\Sigma}_{22}$, e entre os elementos de $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$ é $\boldsymbol{\Sigma}_{12}$. Esta matriz, $\boldsymbol{\Sigma}_{12}$, não necessariamente é simétrica ou até quadrática.

2.4 Variância total e generalizada

a) <u>*Variância total:*</u> é uma forma de sintetização da variância global da distribuição multivariada. Não considera as associações entre as *p* variáveis:

$$traço\left(\underset{pxp}{\mathbf{\Sigma}}\right) = \sigma_{11} + \sigma_{22} + ... + \sigma_{kk} + ... + \sigma_{pp}$$

b) <u>*Variância generalizada:*</u> é uma forma de sintetização da variância global da distribuição multivariada. Ou melhor, é desejável atribuir um único valor numérico para expressar a variação de $\underset{pxp}{\mathbf{\Sigma}}$ ou $\underset{\mathbf{pxp}}{\mathbf{S}}$. Assim, uma escolha para esse valor é o determinante de ambas as matrizes, que reduz para uma única característica – fornece um modo de escrever as informações sobre todas as variâncias e covariâncias como um único valor:

$$\det\left(\underset{pxp}{\mathbf{\Sigma}}\right) = \left|\underset{pxp}{\mathbf{\Sigma}}\right|$$

Por ser determinante, a mesma é influenciada pelas associações entre as *p* variáveis. Para maiores detalhes das propriedades de determinante e traço, veja Johnson e Wichern (2002, p.98).

2.4.1 Matriz de correlação

Para retirar a influência de escala, é possível normalizar os elementos das matrizes $\underset{pxp}{\Sigma}$ e $\underset{pxp}{S}$, como:

– <u>*Populacional:*</u>

$$\underset{\mathbf{pxp}}{\mathbf{P}} = \begin{bmatrix} 1 & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & 1 \end{bmatrix} \text{ em que } \underset{i \neq k}{\rho_{ik}} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}\sigma_{kk}}}$$

– <u>*Amostral:*</u>

$$\underset{\mathbf{pxp}}{\mathbf{R}} = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix} \text{ em que } \underset{i \neq k}{r_{ik}} = \frac{s_{ik}}{\sqrt{s_{ii}s_{kk}}}$$

Estas matrizes são adequadas para avaliar o grau de relacionamento linear entre as variáveis (muitas), pois $-1 \leq \rho_{ik} \leq 1$ e $-1 \leq r_{ik} \leq 1$. Cabe lembrar que capta somente a relação linear entre as variáveis. Relações não lineares geram covariância e correlação nulas.

Ademais, se for definida uma matriz de desvio-padrão, como por exemplo:

$$\underset{pxp}{\mathbf{V}^{1/2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Então, é possível alcançar as seguintes relações:

$$\mathbf{V}^{1/2}\mathrm{P}\mathbf{V}^{1/2} = \Sigma \qquad \text{e} \qquad \left(\mathbf{V}^{-1/2}\right)\Sigma\left(\mathbf{V}^{-1/2}\right) = \mathrm{P}$$

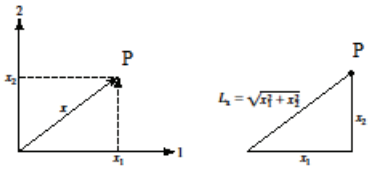
Veja exemplo 2.14 em Johnson e Wichern (2002, p.73).

Caso os dados, por exemplo, sejam padronizados: $z_k = \dfrac{(X_k - \mu_k)}{\sqrt{\sigma_{kk}}}$, ou em forma matricial, $\underset{pxp}{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, a matriz de covariâncias resulta na própria matriz de correlação.

2.4.2 Distâncias

A maioria das técnicas multivariadas é baseada no simples conceito de distância, sendo o mais comum à euclidiana. Pelo teorema de Pitágoras, a distância de um ponto $P = (x_1, x_2)$ em relação ao ponto $O = (0,0)$ é definida como:

$$d(O,P) = \sqrt{x_1^2 + x_2^2} = \sqrt{\mathbf{x}'\mathbf{x}} \qquad \Rightarrow \qquad \mathbf{x}'\mathbf{x} = x_1^2 + x_2^2 = c^2$$

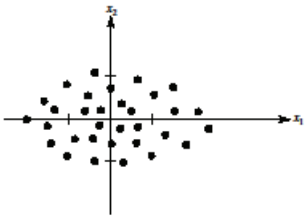


sendo L_x o comprimento do vetor de posição $\mathbf{x}' = [x_1, x_2]$. Por generalização, a distância entre dois pontos com suas respectivas coordenadas, $P = (x_1, x_2, ..., x_p)$ e $Q = (y_1, y_2, ..., y_p)$, é definida por:

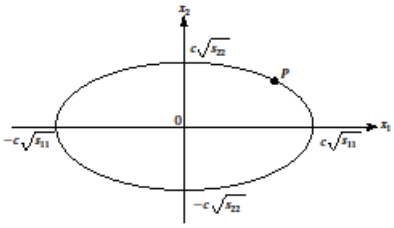
$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} \quad \Rightarrow \quad (\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y}) = c^2$$

em que c^2 representa uma hiperesfera (um círculo se *p*=2), e os pontos equidistantes da origem pertencem a mesma. Quando estas coordenadas representam medidas sujeitas às flutuações aleatórias de diferentes magnitudes, é desejável ponderar as coordenadas com grande variabilidade por menores pesos do que aquelas com baixa variabilidade. Nesse sentido, adota-se a "distância estatística", na qual a distância dependerá das

variâncias e covariâncias (amostrais). Na figura abaixo, parece mais razoável ponderar x_2 com mais peso do que x_1 no cálculo da distância, dividindo pelo desvio padrão (amostral):



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

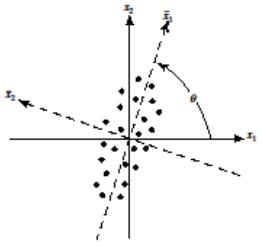


Forma-se uma elipse de distância estatística constante, em que figura acima é de *p=2*. Veja exemplo 1.14 em Johnson e Wichern (2002, p.33). Generalizando a equação para as coordenadas dos pontos, P e Q (supondo este fixo), tem-se:

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

Quando a variabilidade é diferente entre as coordenadas e ao mesmo tempo as mesmas estão correlacionadas, pode-se <u>rotacionar</u> o sistema de coordenadas originais por um anglo de θ mantendo a dispersão fixa. Na figura abaixo, a nova distância a partir de $\tilde{x}_1 = x_1 \cos(\theta) + x_2 sen(\theta)$ e $\tilde{x}_2 = -x_1 sen(\theta) + x_2 \cos(\theta)$, seria:

$$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}} = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2} \Rightarrow \mathbf{x}' \underset{2x2}{A}\, \mathbf{x} = c^2$$



em que $d(O,P) > 0$, os elementos positivos da matriz quadrática e simétrica A são determinados pelo anglo θ e s_{kk} são calculados pelos dados originais. A forma particular dos elementos de A não é importante, mas sim o produto cruzado $2a_{12}x_1x_2$, necessário para uma correlação r_{12} não nula. Generalizando para *p* variáveis aleatórias correlacionadas como coordenadas de um ponto no espaço *p*-dimensional:

$$\underset{1xp}{\mathbf{x}'}\ \underset{pxp}{A}\ \underset{px1}{\mathbf{x}} = c^2$$

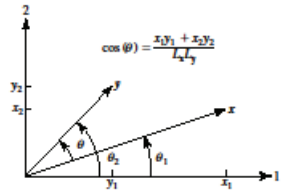
Considerando variáveis correlacionadas, a distância estatística do $P=(x_1,x_2,...,x_p)$ a partir do ponto fixado $Q=(y_1,y_2,...,y_p)$ é:

$$(\mathbf{x}-\mathbf{y})'\underset{pxp}{A}(\mathbf{x}-\mathbf{y})=c^2$$

Todos os pontos (*P*) situados a uma distância quadrática constante de *Q*, pertencem a uma elipse centrada em *Q*, em que seus eixos são paralelos as coordenadas rotacionadas.

2.5 Ortogonalidade e Teorema de decomposição espectral

Sejam dois vetores, $\mathbf{x}'=\begin{bmatrix}x_1 & x_2\end{bmatrix}$ e $\mathbf{y}'=\begin{bmatrix}y_1 & y_2\end{bmatrix}$, com respectivos comprimentos $L_x=\sqrt{\mathbf{x}'\mathbf{x}}$ e $L_y=\sqrt{\mathbf{y}'\mathbf{y}}$, ambos plotados como segue:



Logo,

$$\cos(\theta)=\cos(\theta_2-\theta_1)=\underbrace{\left(\frac{y_1}{L_y}\right)}_{\cos(\theta_2)}\underbrace{\left(\frac{x_1}{L_x}\right)}_{\cos(\theta_1)}+\underbrace{\left(\frac{y_2}{L_y}\right)}_{sen(\theta_2)}\underbrace{\left(\frac{x_2}{L_x}\right)}_{sen(\theta_1)}$$

$$=\frac{x_1y_1+x_2y_2}{L_xL_y}=\frac{\mathbf{x}'\mathbf{y}}{L_xL_y}$$

Desde que o $\cos(90^\circ)=\cos(270^\circ)=0$ e $\cos(\theta)=0$, somente se, $\mathbf{x}'\mathbf{y}=0$, então $\mathbf{x}\perp\mathbf{y}$ (perpendiculares). Os referidos vetores são linearmente dependentes se existir $a_1,a_2\neq 0$, tal que $a_1\mathbf{x}+a_2\mathbf{y}=0$, caso contrário o conjunto de vetores são linearmente independentes. **Importante:** vetores mutuamente perpendiculares são linearmente independentes. Vetores $L=1$ são mutuamente perpendiculares e linearmente independentes. Para tanto, se necessário, divida os elementos de um vetor pelo seu comprimento, tornando-o de $L=1$. Matrizes com vetores de comprimento unitário são conhecidas como ortogonais. Uma matriz ortogonal $\underset{pxp}{O}$ com vetores de comprimento

unitário ($L = 1$) deve satisfazer: $O'O = OO' = \underset{pxp}{I}$ ou $O = O^{-1}$. Por exemplo, $O = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$ é ortogonal.

O uso de vetores perpendiculares ou de matriz ortogonal é fundamental em análise de estatística multivariada, uma vez que matrizes simétricas e de formas quadráticas, como $\underset{pxp}{\boldsymbol{\Sigma}}$ ou $\underset{\mathbf{pxp}}{\mathbf{S}}$, são consequências diretas de uma expansão por decomposição espectral:

$$\mathbf{O'\Sigma O = \Lambda}$$

$$\boldsymbol{\Sigma} = \mathbf{O\Lambda O'} == \sum_{k=1}^{p} \lambda_k \mathbf{e_k e_k'}$$

sendo $\underset{pxp}{\boldsymbol{\Lambda}} = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_p \end{bmatrix}$ $\therefore$ $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$; $\underset{px1}{\mathbf{e_k}} = \begin{bmatrix} e_{k1} \\ \vdots \\ e_{kp} \end{bmatrix}$ um vetor normalizado;

e $\underset{pxp}{\mathrm{O}} = [\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_p}]$, tal que $\mathbf{e}_k'\mathbf{e}_k = 1 (L_{e_k} = 1)$ e $\mathbf{e}_i'\mathbf{e}_k = 0$ (mutualmente perpendiculares e linearmente independentes). No exemplo 1 anteriormente mencionado, conforme Mingoti (2005, p.37), após normalizar os autovetores para que tenham, tem-se:

$$\underset{2x2}{\boldsymbol{\Sigma}} = \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix} = \sum_{k=1}^{p} \lambda_k \mathbf{e_k e_k'} = 9 \begin{bmatrix} 4/5 & -2/5 \\ -2/5 & 1/5 \end{bmatrix} + 4 \begin{bmatrix} 1/5 & 2/5 \\ 2/5 & 4/5 \end{bmatrix} = \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix}$$

$$\det(\boldsymbol{\Sigma}) = \lambda_1 x \lambda_2 = 9x4 = 36 \qquad traço(\boldsymbol{\Sigma}) = \lambda_1 + \lambda_2 = 9 + 4 = 13$$

Logo, como $\underset{pxp}{\boldsymbol{\Sigma}}$ é similar à $\underset{pxp}{\boldsymbol{\Lambda}}$ pelo teorema espectral, os seguintes resultados são alcançados:

a) $traço(\boldsymbol{\Sigma}) = traço(\boldsymbol{\Lambda}) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$ (variância total);

b) $\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Lambda}) = \prod_{k=1}^{p} \lambda_k$ (variância generalizada);

c) $\boldsymbol{\Sigma}^{-1} = \mathbf{O\Lambda}^{-1}\mathbf{O'} == \sum_{k=1}^{p} \frac{1}{\lambda_k} \mathbf{e_k e_k'}$;

d) $\boldsymbol{\Sigma}^{1/2} = \mathbf{O\Lambda}^{1/2}\mathbf{O'} == \sum_{k=1}^{p} \sqrt{\lambda_k}\, \mathbf{e_k e_k'}$.

2.6 Interpretação geométrica da matriz quadrática

A expressão da distância como raiz quadrada de uma forma quadrática positiva definida (p.d) possibilita a interpretação geométrica baseada nos autovalores e autovetores de uma matriz. Dada a matriz $\underset{pxp}{\mathbf{A}}$, e suponha que *p=2*, os pontos $\mathbf{x}' = [x_1, x_2]$ de distância constante *c* da origem satisfazem a:

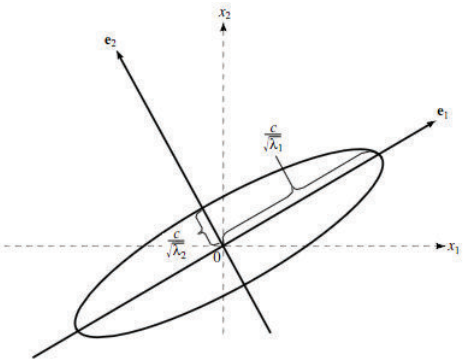
$$\mathbf{x}'A\mathbf{x} = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = c^2$$

Pela decomposição espectral:

$$A = \lambda_1\mathbf{e}_1\mathbf{e}_1' + \lambda_2\mathbf{e}_2\mathbf{e}_2' \quad \Rightarrow \quad \mathbf{x}'A\mathbf{x} = \mathbf{x}'(\lambda_1\mathbf{e}_1\mathbf{e}_1' + \lambda_2\mathbf{e}_2\mathbf{e}_2')\mathbf{x} =$$

$$= \mathbf{x}'A\mathbf{x} = c^2 = \lambda_1(\underbrace{\mathbf{x}'\mathbf{e}_1}_{y_1})^2 + \lambda_2(\underbrace{\mathbf{x}'\mathbf{e}_2}_{y_2})^2 \quad \Rightarrow \quad c^2 = \lambda_1(y_1)^2 + \lambda_2(y_2)^2$$

em que c^2 é um elipse, pois $\lambda_1, \lambda_2 > 0$ quando $\mathbf{A}$ é positiva definida (p.d.). Verifica-se que $\mathbf{x} = c\lambda_1^{-1/2}\mathbf{e}_1$ satisfaz $\mathbf{x}'A\mathbf{x} = \lambda_1(c\lambda_1^{-1/2}\mathbf{e}_1'\mathbf{e}_1)^2 = c^2$ e $\mathbf{x} = c\lambda_2^{-1/2}\mathbf{e}_2$ dá a apropriada distância na direção $\mathbf{e}_2$. Portanto, os pontos de distância constante *c* pertencem a uma elipse cujos eixos são dados pelos autovetores de $\mathbf{A}$ com tamanhos proporcionais ao recíproco da raiz quadrada dos autovalores. O semi-eixo na direção $\mathbf{e}_k$ tem $L_{\mathbf{e}_k} = c\lambda_1^{-1/2}$.



Exclusivamente neste caso, $\lambda_1 < \lambda_2$. Se *p>2*, os pontos $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ de uma distância constante $c = \sqrt{\mathbf{x}'A\mathbf{x}}$ da origem formam um hiperelipsóide, cujos eixos são dados pelos autovetores de $\mathbf{A}$.

2.7 Maximização de formas quadráticas

Na análise multivariada é geralmente necessária a maximização de uma forma quadrática.

2.7.1 Única forma quadrática

Como a forma quadrática $Q = \mathbf{x}'A\mathbf{x}$ pode ser aumentada quando se multiplica por $\mathbf{x}$ muito grande ($\mathbf{x}'\mathbf{x} > 1$), restringe-se o vetor $\mathbf{x}'\mathbf{x} = 1$ na maximização de Q**.** Assim, essa maximização se transforma na razão:

$$\lambda = \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'\mathbf{x}}$$

C.P.O (uso da regra do quociente)

$$\frac{\partial \lambda}{\partial \mathbf{x}} = 0 = \frac{2A\mathbf{x}(\mathbf{x}'\mathbf{x}) - 2(\mathbf{x}'A\mathbf{x})\mathbf{x}}{(\mathbf{x}'\mathbf{x})^2} = \frac{2}{\mathbf{x}'\mathbf{x}}\left(A - \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'\mathbf{x}} I\right)\mathbf{x} \Rightarrow \quad \div \frac{2}{\mathbf{x}'\mathbf{x}} \quad \Rightarrow \left(A - \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'\mathbf{x}} I\right)\mathbf{x} = 0$$

$$(A - \lambda_i I)\mathbf{x}_i = 0$$

Para a solução acima, é importante que a matriz $(A - \lambda_i I)$ seja singular, ou seja, que o $\det(A - \lambda_i I) = 0$ (ou que não tenha um posto completo). Também significa que $\mathbf{x}_i = e_i$, $\mathbf{e}_i'\mathbf{e}_i = 1$, $\mathbf{e}_i'\mathbf{e}_k = 0$ e λ_i é máximo valor da forma quadrática de $Q = \mathbf{x}'A\mathbf{x}$. Note que o problema de maximização forma o Lagrange:

$Max \quad \mathbf{x}_i'A\mathbf{x}_i \quad s.a. \quad \mathbf{x}_i'\mathbf{x}_i = 1$

$$L = \mathbf{x}_i'A\mathbf{x}_i - \lambda(\mathbf{x}_i'\mathbf{x}_i - 1) \quad \Rightarrow \quad C.P.O. \quad \Rightarrow A\mathbf{x}_i - \lambda\mathbf{x}_i = 0 \Rightarrow (A - \lambda_i I)\mathbf{x}_i = 0 \quad ou \quad \mathbf{x}_i'A\mathbf{x}_i = \lambda$$

2.7.2 Pares de forma quadrática

Especialmente na análise canônica, maximiza-se a razão de duas formas quadráticas:

$$\lambda = \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'B\mathbf{x}}$$

C.P.O.:

$$\frac{\partial \lambda}{\partial \mathbf{x}} = 0 = \frac{2A\mathbf{x}(\mathbf{x}'B\mathbf{x}) - 2(\mathbf{x}'A\mathbf{x})B\mathbf{x}}{(\mathbf{x}'B\mathbf{x})^2} = \frac{2}{\mathbf{x}'B\mathbf{x}}\left(A - \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'B\mathbf{x}} B\right)\mathbf{x}$$

$$\Rightarrow \quad \times \frac{\mathbf{x}'B\mathbf{x}}{2} \quad \Rightarrow \left(A - \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'B\mathbf{x}} B\right)\mathbf{x} = 0$$

$$(A - \lambda_i B)\mathbf{x}_i = 0$$

2.8 Propriedades das combinações lineares de variáveis aleatórias

Seja Z uma variável de combinação linear como:

$$Z_1 = aX_1 + bX_2 \quad \text{(a e b constantes)}$$

$$E(Z_1) = (a\mathbf{X}_1 + b\mathbf{X}_2) = aE(\mathbf{X}_1) + bE(\mathbf{X}_2) = a\mu_1 + b\mu_2 = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_1 \end{bmatrix} = \mathbf{c}'\boldsymbol{\mu}.$$

$$Var(Z_1) = E[a(\mathbf{X}_1 - \mu_1) + b(\mathbf{X}_2 - \mu_2)]^2 = a^2\sigma_{11} + 2ab\sigma_{12} + b^2\sigma_{22} =$$

$$= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{c}'\Sigma\mathbf{c}$$

Portanto, uma combinação linear $\mathbf{c}'\mathbf{X} = c_1X_1 + ... + c_pX_p$ tem:

$E(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\boldsymbol{\mu}$	$Var(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\Sigma\mathbf{c}$

Assim, para *q* combinações lineares de *p* variáveis aleatórias:

$$\underset{\mathbf{qx1}}{\mathbf{Z}} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_q \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1p} \\ & & \\ c_{p1} & & c_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} = \underset{\mathbf{qxp\ px1}}{\mathbf{C\ X}}$$

$\boldsymbol{\mu}_z = E(\mathbf{Z}) = E(\mathbf{CX}) = C\boldsymbol{\mu}$	$\boldsymbol{\Sigma}_z = COV(\mathbf{CX}) = \mathbf{C}\Sigma\mathbf{C}'$

Em Johnson e Wichern (2002), veja o exemplo 2.15 (p.77) e o exercício 2.28 (p.107 e 108), que computa os elementos fora da diagonal em $\mathbf{C}\Sigma\mathbf{C}'$. Como o resultado final do exemplo 2.15 (p.77):

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \mathbf{CX}$$

$$\boldsymbol{\Sigma}_z = \mathbf{C}\Sigma\mathbf{C}' = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{11} - \sigma_{22} \\ \sigma_{11} - \sigma_{22} & \sigma_{11} + 2\sigma_{12} + \sigma_{22} \end{bmatrix}$$

Se X_1 e X_2 tivesse a mesma variância ($\sigma_{11} = \sigma_{22}$), os termos fora da diagonal em $\boldsymbol{\Sigma}_z$ desapareceriam. Tem-se um resultado conhecido: a soma e a diferença de duas variáveis aleatórias com idêntica variância não são correlacionáveis.

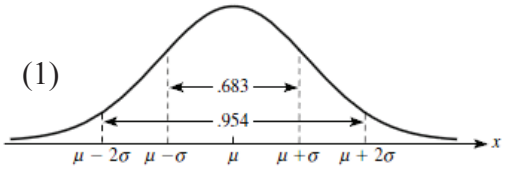
3 DISTRIBUIÇÃO NORMAL MULTIVARIADA

Algumas técnicas multivariadas parte do pressuposto de que os dados foram gerados de uma distribuição normal. Apesar dos dados não serem exatamente normal multivariados, a densidade normal constitui, algumas vezes, uma aproximação útil e adequada da real distribuição populacional. Além de facilitar o tratamento matemático, independentemente da distribuição populacional, as distribuições amostrais, tais como Poisson e binomial, podem ser próximas das normais devido ao efeito do limite central. Ou seja, é conhecido que a distribuição em várias estatísticas multivariadas torna-se tipicamente normal quando a amostra aumenta de tamanho (teorema do limite central). Do ponto de vista prático, existe consideráveis vantagens por trabalhar com grandes amostras.

3.1 Densidade normal multivariada

A densidade normal multivariada é uma generalização da distribuição normal univariada para $p \geq 2$. Com média μ e variância σ^2, tem-se a função de densidade de probabilidade:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2/\sigma^2}{2}} \qquad -\infty < x < \infty \qquad (1)$$



Na figura, as áreas entre aproximadamente 1 desvio padrão da média e 2 desvios padrões são respectivamente, 68,3% e 95,4%. O expoente da função, $(x-\mu)^2/\sigma^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$, mede o quadrado da distância entre x e μ em unidade de desvio padrão. Pra um vetor $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$, o termo pode ser reescrito como:

$$(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2 \text{ (distância constante)} \qquad (3)$$

que é a distância estatística (normalizada) com a matriz simétrica $\underset{pxp}{\boldsymbol{\Sigma}}$ positiva e definida (p.d.). O vetor $\underset{px1}{\boldsymbol{\mu}}$ é o valor esperado do vetor aleatório $\mathbf{X}$[1]. Além disso, no caso

[1] Mantendo $\mathbf{X}$ constante, representa-se $\mathbf{x}$. A distância estatística é também conhecida como matriz de Mahalanobis.

multivariado, as probabilidades são representadas por volumes sob a superfície da função $f(\mathbf{x})$ ao longo das regiões definidas pelos intervalos dos valores de $\mathbf{x}_k$: $(2\pi)^{-p/2}\det(\Sigma)^{-1/2}$ [sobre $\det(\Sigma)$ como área, veja Johnson e Wichern (2002, cap.3)]. Assim,

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-[(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})]/2} \qquad -\infty < \mathbf{x}_i < \infty \quad i = 1,2,...,p \qquad (4)$$

que é denotada por $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e a matriz Σ^{-1} é também (p.d.). A densidade normal multivariada é constante em superfícies, ou seja, c é constante, que leva $f(\mathbf{x})$ a ter o mesmo valor numérico.

Johnson e Wichern (2002, p. 151) apresenta o exemplo 4.1 da derivação de uma função densidade de probabilidade bivariada, como segue:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \qquad \Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix}$$

Sabe-se que $\underset{i\neq k}{\rho_{ik}} = \sigma_{ik}(\sigma_{ii}\sigma_{kik})^{-1/2} \Rightarrow \sigma_{ik} = \underset{i\neq k}{\rho_{ik}}(\sigma_{ii}\sigma_{kk})^{1/2}$, logo a variância generalizada no contexto bivariado seria: $|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22} - \rho_{12}^2\sigma_{11}\sigma_{22} = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$.

$$|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$$

Portanto, para σ_{11} e σ_{22} fixos, quanto maior o valor do coeficiente ρ, menor será a variância generalizada. Note que a presença de correlação faz om que as probabilidades se concentrem ao longo de uma linha. No caso em que $\rho_{12} = 0$, as variáveis seriam independentes (não correlacionadas) e a densidade conjunta poderia ser escrita como o produto de duas densidades normal univariada: $f(x_1, x_2) = f(x_1)f(x_2)$.

Desde que $|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$, o expoente de $f(\mathbf{x})$ da equação (4) pode ser reescrito como:

$$(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (x_1-\mu_1, x_2-\mu_2)\frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\begin{bmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{21} & \sigma_{11}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}$$

$$= \frac{1}{(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}^2\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]$$

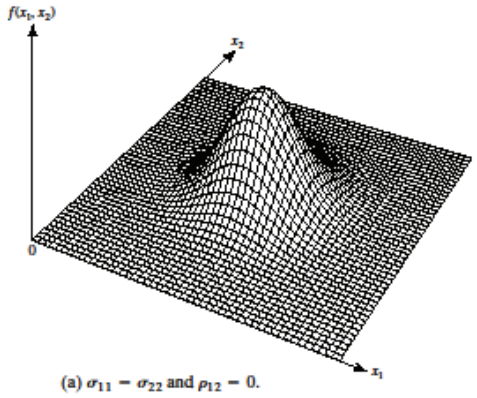
Logo, a função densidade de probabilidade bivariada seria:

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}}e^{\frac{-1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]} \qquad (5)$$

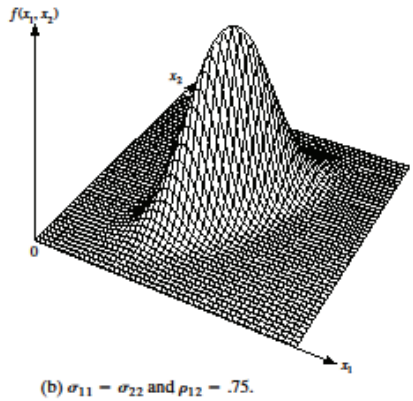
Portanto, das distribuições bivariadas com $\sigma_{11} = \sigma_{22}$, tem-se que:

a) x_1 e x_2 são independentes ($\rho_{12} = 0$);

b) $\rho_{12} = 0{,}75$, i.e., a correlação causa probabilidades que se concentram ao longo de uma linha.



(a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$.



(b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .75$.

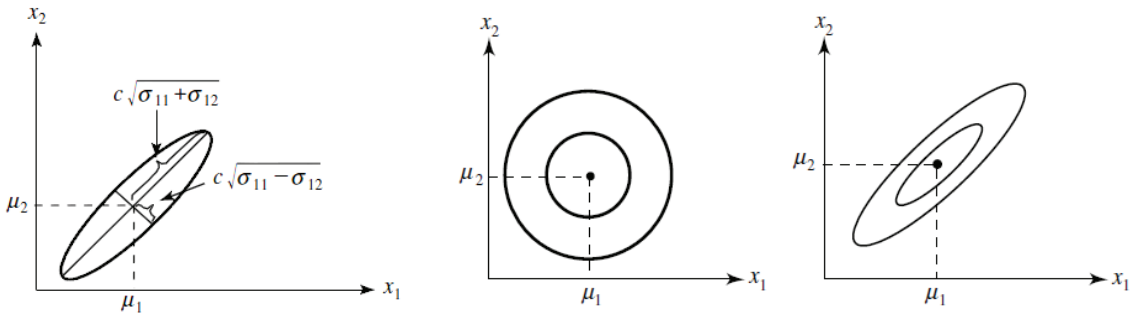
Para a densidade de uma variável normal de *p*-dimensões, os caminhos dos valores de $\mathbf{x}$ rendem uma altura constante. Ou melhor, $f(\mathbf{x})$ em (4) apresenta pontos de igual densidade, que são chamados de contornos. Esses contornos forma elipsóides definidos por $\mathbf{x}$, tal que:

$$(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

Essas elipsóides são centradas em $\boldsymbol{\mu}$ e têm eixos $\pm c\lambda_k^{1/2}\mathbf{e}_k$, na qual $\Sigma^{-1} = \sum_{k=1}^{p} \lambda_k^{-1}\mathbf{e_k}\mathbf{e_k'}$ ou $\Sigma = \sum_{k=1}^{p} \lambda_k \mathbf{e_k}\mathbf{e_k'}$, sendo $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$. Veja a discussão na seção 2.5, em que $c = \sqrt{\mathbf{x}'A\mathbf{x}}$ contém eixos $\pm c\lambda_k^{-1/2}\mathbf{e}_k$. Como Σ^{-1} é uma matriz inversa com as mesmas propriedades de A, então só muda o sinal do expoente sobre seus os autovalores. Considerando o exemplo 4.2 de Johnson e Wichern (2002, p.154), em uma função de densidade bivariada com $\sigma_{11} = \sigma_{22}$ e $\sigma_{12} > 0$:

$$|\boldsymbol{\Sigma} - \lambda I| = 0 = \begin{vmatrix} \sigma_{11} - \lambda & \sigma_{12} \\ \sigma_{21} & \sigma_{22} - \lambda \end{vmatrix} \quad \Rightarrow \quad \begin{aligned} \lambda_1 = \sigma_{11} + \sigma_{12} : \mathbf{e}_1 &= \left[1/\sqrt{2}, 1/\sqrt{2} \right] \\ \lambda_2 = \sigma_{11} - \sigma_{12} : \mathbf{e}_2 &= \left[1/\sqrt{2}, -1/\sqrt{2} \right] \end{aligned}$$

Os eixos das elipses de densidade constante são $\pm c\lambda_k^{1/2}\mathbf{e}_k \quad (k = 1,2)$, de maneira que o seu eixo principal será de maior autovalor (λ_1) e seu autovetor ($\mathbf{e}_1$), que se posiciona a um linha de 45º do ponto $\boldsymbol{\mu} = [\mu_1, \mu_2]$. Nas figuras abaixo, os contornos de densidade constante contém 50% e 90% de probabilidade sob uma superfície normal bivariada.



A escolha de $c^2 = \chi_p^2(\alpha)$, em que $\chi_p^2(\alpha)$ é o percentil (100α) superior da distribuição de Qui-quadrado com p graus de liberdade, leva a contornos que contém $(1-\alpha)\times 100\%$ de probabilidade. Para a distribuição normal multivariada (p variada), a elipsóide dos valores de $\mathbf{x}$ satisfaz: $\Pr[(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi_p^2(\alpha)] = 1-\alpha$.

3.2 Propriedades da Distribuição Normal Multivariada

Considerando que o vetor $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então:

a) combinações lineares de $\mathbf{X}$ têm distribuição normal: $\mathbf{a}'\mathbf{X} \sim N_p(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$;

b) todos os subconjuntos de $\mathbf{X}$ têm distribuição normal multivariada, ou seja, se $\underset{1xp}{\mathbf{X}'} = \left[\underset{1xq}{\mathbf{X}_1} \vdots \underset{1x(p-q)}{\mathbf{X}_2} \right]$, então $\mathbf{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ e $\mathbf{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$;

c) covariância zero implica que os componentes correspondentes de $\mathbf{X}$ são independentemente distribuídos;

d) distribuições condicionais dos componentes de $\mathbf{X}$ são normais (multivariadas);

e) $(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi_p^2$, em que χ_p^2 denota uma distribuição qui-quadrada com p graus de liberdade;

f) $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avalia a probabilidade $(1-\alpha)$ para uma elipsoide sólida $\left\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi_p^2(\alpha)\right\}$, em que $\chi_p^2(\alpha)$ é o percentil (100α) superior da distribuição χ_p^2.

Os exemplos 4.4 (p.157), 4.5 (p.159), 4.6 (p.160), 4.7 (161) de Johnson e Wichern (2002) tratam das propriedades (a)-(d) e o resultado 4.7 dos mesmos autores discute as propriedades de χ_p^2 (e-f). Em suma, as propriedades $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotam que todas as combinações lineares da normal individual são normais; e os contornos de densidade normal multivariada são elipsóides concêntricos.

Por fim, cabe mencionar a interpretação da distância estatística. Nela, se um componente tem uma variância muito maior do que o outro, o mesmo contribuirá menos na distância estatística. Além disso, duas variáveis aleatórias altamente correlacionadas influenciarão menos do que duas variáveis pouco correlacionadas. Essencialmente, o uso da inversa da matriz de covariâncias (Σ^{-1}): a) padroniza todas as variáveis; e b) ameniza os efeitos de correlação. Formalmente, verifica-se que:

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= Z_1^2 + Z_2^2 + \ldots + Z_p^2 \\ &= (\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu})\end{aligned}$$

4 AVALIANDO A SUPOSIÇÃO DE NORMALIDADE

Muitas técnicas estatísticas assumem que cada vetor $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Em situações em que o tamanho amostral é grande e as técnicas dependem somente sobre o comportamento de $\overline{\mathbf{X}}$, ou distâncias envolvendo $\overline{\mathbf{X}}$ em distâncias estatísticas, a suposição de normalidade para as observações individuais é menos crucial. Contudo, para algum grau, a qualidade das inferências feitas por estes métodos dependem sobre quão próximo à população verdadeira aparente se assemelha a forma normal multivariada. Este imperativo permite executar procedimentos a fim de detectar casos em que os dados exibem padrões moderados ou até extremos a partir do que é esperado sob a hipótese de normalidade multivariada. Diante disso, três questões podem ser consideradas:

a) as distribuições marginais dos elementos de $\mathbf{X}$ parecem normais?
b) os gráficos de dispersão bivariados parecerem elipsoidais?
c) há observações discrepantes (*outliers*)?

Inicialmente a análise se concentra sobre o comportamento das observações em uma ou duas dimensões (e.g., distribuições marginais e gráficos de dispersão) [2]. Cabe mencionar previamente algumas observações práticas: a) é possível construir distribuições bivariadas não normais com normalidades marginais (e.g., veja o caso do exercício 4.8 de Johnson e Wichern (2002)); b) muitos tipos de não normalidade são refletidos nas distribuições marginais e gráficos de dispersão; e c) conjunto de dados patológicos, que são normais em representações de menores dimensões e não são normais em maiores dimensões, não frequentemente encontrados.

4.1 Avaliando a normalidade das distribuições marginais univariadas

Alguns instrumentos podem ser usados para verificar a normalidade univariada, quais sejam:

---

[2] Ainda assim, estes procedimentos têm fornecido dificuldades para construir um "bom" teste global de normalidade conjunta em mais de duas dimensões porque um número de grande de situações pode dar errado. Até certo ponto, pode-se pagar um preço por concentrar-se sobre o contexto univariado ou bivariado, até porque em grandes dimensões surgem algumas características latentes.

a) **<u>distribuição da proporção</u>**: diagramas de pontos (*n* pequeno) e histogramas (*n*>25) são aplicadas para verificar a distribuição univariada. Para pequenas amostras, o histograma pode ser irregular na aparência e a avaliação da normalidade é dificultada. Se o histograma para uma variável $\mathbf{X}_k$ aparece razoavelmente simétrica, pode-se checar o número de observações que está dentro de certos intervalos definidos[3]. Por definição, o histograma particiona intervalos de $\mathbf{X}_k$ de igual comprimento e a média é o centro da distribuição[4]. Além do histograma, calcula-se a distância generalizada do centróide, padronizando as variáveis (i.e., variável aleatória normal padrão). Espera-se que a proporção das observações seja:

$$P(\mu - 1\sigma \le x \le \mu + 1\sigma) = 0{,}68$$
$$P(\mu - 2\sigma \le x \le \mu + 2\sigma) = 0{,}95$$
$$P(\mu - 3\sigma \le x \le \mu + 3\sigma) = 0{,}997$$

Por exemplo, espera-se que a proporção das observações dentro de um desvio padrão em relação à média seja entorno de 0,68.

b) **<u>gráficos Q-Q</u>**: são obtidos da distribuição marginal das observações de cada variável. Cada gráfico consiste em plotar em um plano cartesiano os quantis amostrais *versus* os quantis esperados pelo ajuste de uma distribuição normal. Se <u>os pontos pertencem quase a uma linha reta</u>, o pressuposto de normalidade deve se confirmar. Para tanto, considere $x_1, x_2, \ldots, x_n$ como observações de qualquer característica X_i. Ordene os valores de tais observações de forma crescente, por exemplo, suponha que $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. A proporção amostral j/n é aproximada por $(j - 0{,}5)/n$, em que o valor 0,5 é usado para correção de

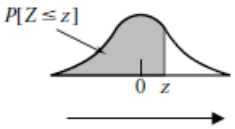
---

[3] Cabe mencionar que a distribuição pode ser simétrica e não ser ainda normal, porém frequentemente distribuições simétricas são próximas de uma normal.

[4] Assim, como o histograma, o "box plot" é uma ferramenta para avaliar as simetrias de uma distribuição empírica por meio de percentis (ou quantis).

descontinuidade. Para uma distribuição normal padronizada, os quantis $q_{(j)}$ são definidos da relação:

$$P(Z \le q_{(j)}) = \int_{-\infty}^{q_{(j)}} \frac{1}{\sqrt{2\pi}} e^{-Z^2/2} dz = p(j) = \frac{j - 0{,}5}{n}$$



Os quantis $q_{(j)}$ podem ser obtidos, por exemplo, pelas tabelas de distribuição normal. Gráficos Q-Q não são particularmente informativos, ao menos que o tamanho amostral seja moderado ou grande $(n \ge 20)$. Ou seja, pode existir um pouco de linearidade do gráfico Q-Q para pequenas amostras, mesmo quando as observações são conhecidas de uma população normal. Veja abaixo o exemplo 4.10 de Johnson e Wichern (2002, p.180). A linearidade do gráfico Q-Q pode ser mensurada ao calcular o coeficiente de correlação dos pontos no gráfico.

c) **Coeficiente de correlação de Pearson**: refere-se a um teste complementar ao Gráfico Q-Q. Rejeita-se a hipótese de normalidade se o valor estiver abaixo do valor crítico (r_c).

$$r_Q = \frac{\sum_{j=1}^{n} (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n} (x_{(j)} - \bar{x})^2} \sqrt{\sum_{j=1}^{n} (q_{(j)} - \bar{q})^2}}$$

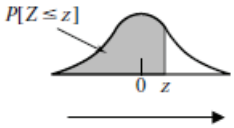
Table 4.2 Critical Points for the Q–Q Plot Correlation Coefficient Test for Normality

Sample size n	Significance levels α		
	.01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9749	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

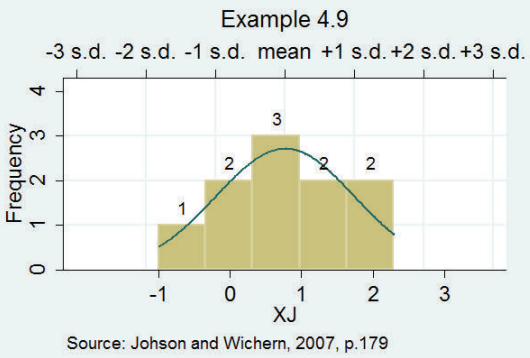
Alguns programas avaliam a estatística original, proposta por Shapiro e Wilk. Esta forma de correlação corresponde em substituir $q_{(j)}$ por uma função de valor esperado de ordem normal padrão e suas covariâncias. Johnson e Wichern (2002) preferem a correlação de Pearson porque a mesma corresponde diretamente os pontos de escores normais nos gráficos. Para grandes amostras,

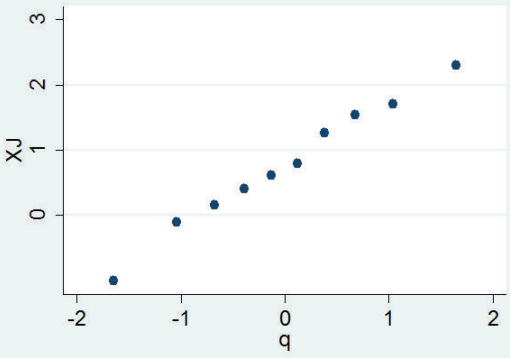
essas estatísticas são próximas, que podem ser usadas para jugar à falta de ajuste.

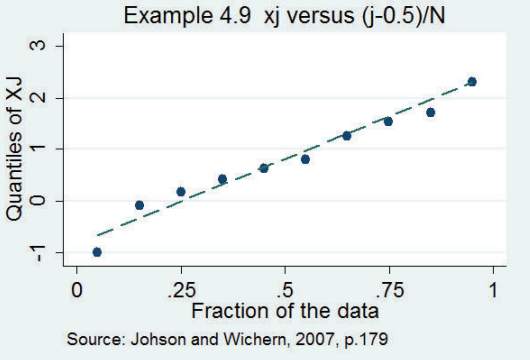
Execute os dois programas (*do-file*) com seus respectivos dados para avaliar os exemplos 4.9 e 4.10 de Johnson e Wichern (2002, p.179-180). Abaixo estão os resultados. No exemplo 4.9, 80% das observações estão dentro de 1 desvio padrão em relação à média, e todas as observações estão dentro de 2 desvios padrões. Os pontos indicam que existem pouco discrepantes e, pela linearidade, eles sugerem uma distribuição normal, apesar do tamanho amostral pequeno (n=10). Por exemplo, para a observação 1 tem-se:



$$\frac{1-0{,}5}{n} = \frac{1-0{,}5}{15} = 0.05 = P(Z \leq -1.65)$$







ordem	XJ	(J-0,50)/n	Quantil normal padrão (q)	Variável Normal Padrão z(j)
1	-1	0.05	-1.64	-1.82
2	-0.1	0.15	-1.04	-0.90
3	0.16	0.25	-0.67	-0.63
4	0.41	0.35	-0.39	-0.37
5	0.62	0.45	-0.13	-0.15
6	0.8	0.55	0.13	0.03
7	1.26	0.65	0.39	0.51
8	1.54	0.75	0.67	0.79
9	1.71	0.85	1.04	0.97
10	2.3	0.95	1.64	1.58

Fonte: Johnson e Wichern (2007, p.179)

Pela correlação de Pearson, o teste de normalidade em um nível de significância de 10% com n=10 seria 0,9351 (*rc*). Portanto, desde r>0,9351, não se rejeita a hipótese de normalidade. No exemplo numérico anterior, em que sempre $\bar{q} = 0$, tem-se:

Correlação de Pearson

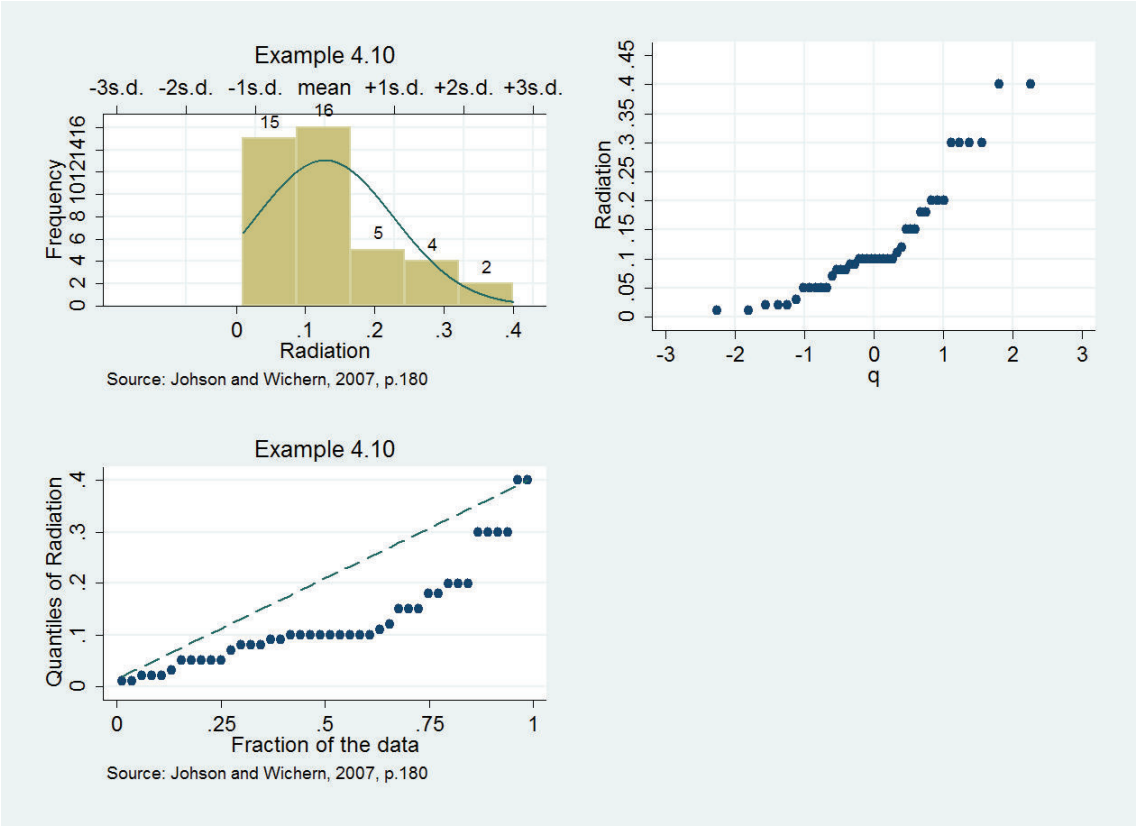
N.	$x_{(j)} - \bar{x}$	$(x_{(j)} - \bar{x})^2$	$q_{(j)} - \bar{q}$	$(q_{(j)} - \bar{q})^2$	$(x_{(j)} - \bar{x})(q_{(j)} - \bar{q})$
1	-1.8	3.1	-1.645	2.7	2.9
2	-0.9	0.8	-1.036	1.1	0.9
3	-0.6	0.4	-0.674	0.5	0.4
4	-0.4	0.1	-0.385	0.1	0.1
5	-0.2	0.0	-0.126	0.0	0.0
6	0.0	0.0	0.126	0.0	0.0
7	0.5	0.2	0.385	0.1	0.2
8	0.8	0.6	0.674	0.5	0.5
9	0.9	0.9	1.036	1.1	1.0
10	1.5	2.3	1.645	2.7	2.5
Total	**0.0**	**8.472**	**0.0**	**8.79787**	**8.585**

rj=	**0.994**

$$r_Q = \frac{\sum_{j=1}^{n}(x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n}(x_{(j)} - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(q_{(j)} - \bar{q})^2}} \quad \Rightarrow r_Q = \frac{8.585}{\sqrt{8.472}\sqrt{8.797}} = 0.994$$

Esse teste converge com o de Shapiro-Wilk (1965), pois não se rejeita a hipótese de distribuição normal da variável (Prob>z= 0.99676). O teste de assimetria/curtose para normalidade corrobora com tal análise (Prob>chi2= 0.9364).

Já o exemplo 4.10 aponta que existem alguns pontos discrepantes, além dos mesmos não seguirem uma distribuição normal. Para estes dados, algumas observações são iguais, cujos valores são associados ao mesmo quantil normal. A correlação de Pearson registrou 0.9279, inferior aos valores críticos (entre n=40 e 45). Este resultado converge com os testes de Shapiro-Wilk e assimetria/curtose, que rejeitaram a hipótese de normalidade ao nível de significância de 1%. Ademais, aproximadamente 74% das observações encontram-se dentro de 1 desvio padrão em relação à média.



4.2 Avaliando a normalidade bivariada

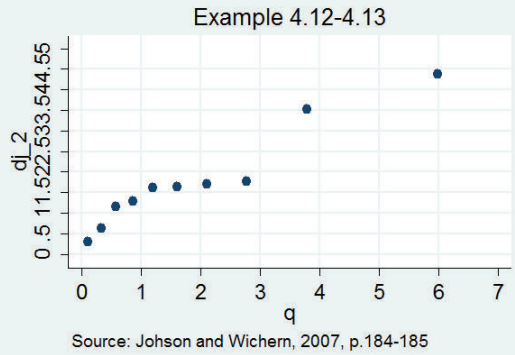
Busca-se também verificar de maneira prática a suposição de normalidade para todas as distribuições de várias dimensões $(p \geq 2)$. Para tanto, é suficiente investigar as distribuições bivariadas (cada uma deve ser normal se a **distribuição global conjunta** é normal multivariada). Se as observações foram geradas a partir de uma distribuição normal multivariada, cada distribuição bivariada seria normal, e os contornos da densidade constante seriam elipses. Além do gráfico de dispersão, que deve exibir um padrão quase elíptico, elabora-se um gráfico de probabilidade qui-quadrado, que relaciona os valores da distância quadrática generalizada entre o centróide e cada observação, $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$, com as respectivas ordenadas dos quantis da distribuição qui-quadrada, cujos passos de construção são:

Passo 1: calcule d_j^2 para todas as observações e ordene-as conforme $d^2_{(1)} \leq d^2_{(2)} \leq \ldots \leq d^2_{(n)}$ (ordem crescente).

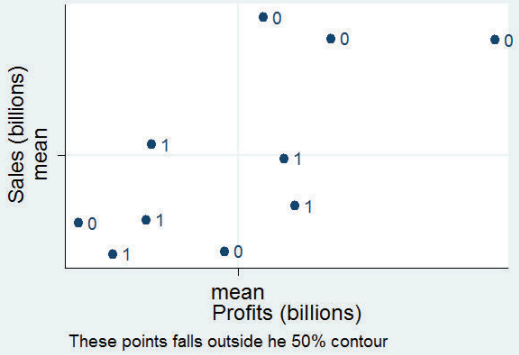
Passo 2: calcule $\chi_p^2((j-0{,}5)/n)$ de p graus de liberdade. Em seguida, construa um gráfico relacionando os valores de $\chi_p^2((j-0{,}5)/n)$ com os de d_j^2. Em dados de normalidade p-variada, espera-se algo próximo de uma reta no gráfico.

Passo 3: para amostras grandes**,** pelo menos 50% das observações devem residir na elipse: $(\mathbf{x}-\bar{\mathbf{x}})'S^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \le \chi_p^2(0.5)$. Ou melhor, se estão dentro do elipsóide que contém 50% de todas as observações.

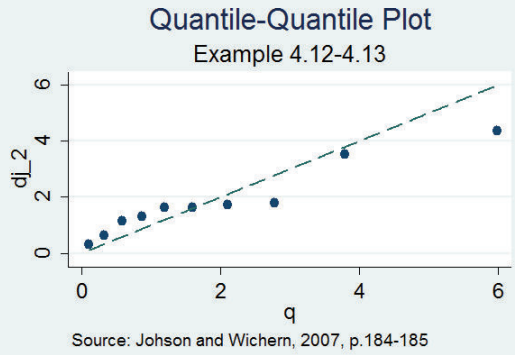
Para tamanhos amostrais pequenos, somente comportamentos atípicos serão identificados como falta de ajuste. Já as amostras grandes produzem invariavelmente estatísticas significativas da falta de ajuste. Os exemplos 4.12 e 4.13 de Johnson e Wichern (2002, p.183-184) estão reportados abaixo.



Source: Johson and Wichern, 2007, p.184-185



These points falls outside he 50% contour



Source: Johson and Wichern, 2007, p.184-185

company	dj_2	j	j_05	q
General Electric	0.30	1	0.05	0.10
Amaerican Intl Group	0.62	2	0.15	0.33
Toyota Motor	1.16	3	0.25	0.58
HSBC Group	1.29	4	0.35	0.86
Citigroup	1.61	5	0.45	1.20
Royal Dutch/Shell	1.63	6	0.55	1.60
ING Group	1.71	7	0.65	2.10
Bank of America	1.78	8	0.75	2.77
BP	3.53	9	0.85	3.79
ExxonMobil	4.37	10	0.95	5.99

O valor crítico de $\chi_{p=2}^2(0.5) = 1.39$ é 1,39 e existem 50% das observações que estão dentro do contorno com probabilidade de 50%. Essa proporção poderia fornecer

evidências para rejeitar a hipótese de normalidade bivariada. Entretanto, o tamanho da amostra de 10 é muito pequeno para alcançar esta conclusão.

5 DETECTANDO *OUTLIERS*

Muitos conjuntos de dados contêm uma ou algumas observações que são discrepantes com o padrão de variabilidade produzida por outras observações. Esta situação pode ser dificultada em contextos multivariados. Os *outliers*, algumas vezes, não são resultados errados. Os mesmos podem, inclusive, ajudar no entendimento do fenômeno em estudo. *Outliers* são melhores detectados se sua visualização for possível. Quando o número de observações é grande, o gráfico de pontos é inviável. Por outro lado, quando o número de variáveis é grande, é inviável construir gráficos de dispersão $(p \geq 4)$. Assim, existem alguns passos para detectá-los em um contexto multivariado:

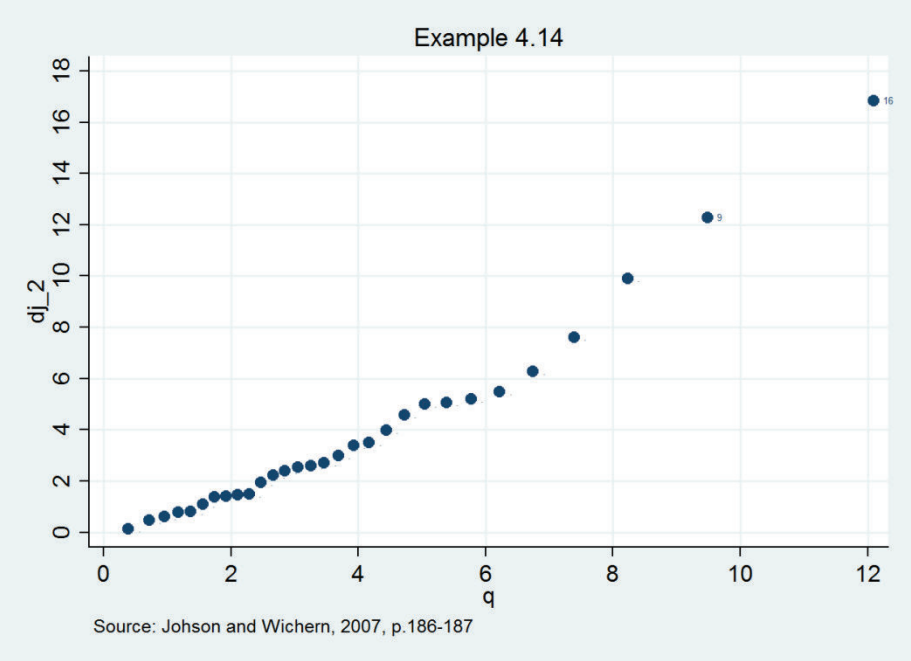
a) faça um gráfico de dispersão para cada par de variáveis;
b) padronize as variáveis e examine pequenos e grandes valores;
c) calcule a distância quadrática generalizada. Examine essas distâncias para valores grandes. Estes valores seriam os mais distantes da origem em um gráfico Q-Q.

Cabe mencionar que no passo (b) o "grande" deve ser interpretado relativamente ao tamanho da amostra e o número de variáveis. Esperam-se *outliers* mesmo se os dados atenderem uma distribuição normal. Por seu torno, no passo (c) o "grande" é medido por um percentil apropriado de uma distribuição χ_p^2. Se o tamanho da amostra para n=100, esperam-se 5 observações com valores de d_j^2 acima do quinto percentil superior da distribuição $\chi_p^2 = (0{,}005)$. O maior percentil deve servir para determinar observações que não se ajustam ao padrão das demais observações.

O exemplo 4.5 de Johnson e Wichern (2002, p.190) ilustra os passos supracitados.

obs	xj1	xj2	xj3	xj4	dj_2	Z1	Z2	Z3	Z4
1	1889	1651	1561	1778	0.60	-0.05	-0.31	0.17	0.16
2	2403	2048	2087	2197	5.48	1.53	0.94	1.91	1.46
3	2119	1700	1815	2222	7.62	0.66	-0.16	1.01	1.54
4	1645	1627	1110	1533	5.21	-0.80	-0.38	-1.32	-0.59
5	1976	1916	1614	1883	1.40	0.22	0.52	0.35	0.49
6	1712	1712	1439	1546	2.22	-0.60	-0.12	-0.23	-0.55
7	1943	1685	1271	1671	4.99	0.11	-0.20	-0.79	-0.17
8	2104	1820	1717	1874	1.49	0.61	0.22	0.69	0.46
9	2983	2794	2412	2581	12.26	3.31	3.28	2.98	2.65
10	1745	1600	1384	1508	0.77	-0.50	-0.47	-0.41	-0.67
11	1710	1591	1518	1667	1.93	-0.60	-0.50	0.03	-0.18
12	2046	1907	1627	1898	0.46	0.43	0.49	0.39	0.54
13	1840	1841	1595	1741	2.70	-0.20	0.29	0.28	0.05
14	1867	1685	1493	1678	0.13	-0.12	-0.20	-0.05	-0.15
15	1859	1649	1389	1714	1.08	-0.14	-0.32	-0.40	-0.03
16	1954	2149	1180	1281	16.85	0.15	1.25	-1.09	-1.38
17	1325	1170	1002	1176	3.50	-1.79	-1.82	-1.67	-1.70
18	1419	1371	1252	1308	3.99	-1.50	-1.19	-0.85	-1.29
19	1828	1634	1602	1755	1.36	-0.24	-0.36	0.31	0.09
20	1725	1594	1313	1646	1.46	-0.56	-0.49	-0.65	-0.24
21	2276	2189	1547	2111	9.90	1.14	1.38	0.12	1.20
22	1899	1614	1422	1477	5.06	-0.02	-0.43	-0.29	-0.77
23	1633	1513	1290	1516	0.80	-0.84	-0.74	-0.72	-0.65
24	2061	1867	1646	2037	2.54	0.48	0.37	0.45	0.97
25	1856	1493	1356	1533	4.58	-0.15	-0.81	-0.51	-0.59
26	1727	1412	1238	1469	3.40	-0.55	-1.06	-0.89	-0.79
27	2168	1896	1701	1834	2.38	0.81	0.46	0.63	0.34
28	1655	1675	1414	1597	3.00	-0.77	-0.23	-0.31	-0.40
29	2326	2301	2065	2234	6.28	1.29	1.73	1.83	1.58
30	1490	1382	1214	1284	2.58	-1.28	-1.15	-0.97	-1.37

Este exemplo revela que a observação "16" é um *outlier* multivariado, desde que $\chi^2_{p=2}(0.005) = 14{,}86$. Todas as observações estão bem dentro das suas respectivas dispersão univariada. A observação "9' também revela um grande valor de d_j^2. Assim, essas duas observações, "9" e "16", com grande distância quadrática, se destacam como diferentes do padrão, conforme a reta esperada. Uma vez que estas duas observações sejam removidas, o padrão restante segue conforme a reta esperada.



Nos gráficos de dispersão, a observação "16" situa fora de todos eles, enquanto que a observação "9" está escondida no gráfico (x3 *versus* x4) e no gráfico (x1 *versus* x3). Não obstante, a observação "9" é claramente identificada como um *outlier* multivariado quando quatro variáveis são consideradas. Os pesquisadores concluíram que para essas duas observações, houve um erro de digitação.

Dependendo da natureza dos *outliers* e dos objetivos da pesquisa, tais pontos podem ser removidos ou apropriadamente "ponderados" em uma subsequente análise. Existem duas regras básicas quanto ao tratamento dos *outliers*:

a) o investigador pode desejar eliminar esses *outliers* a partir de uma análise, porém reportá-los com análises estatísticas;
b) ou executar duas análises, com e sem *outliers,* para ver se os mesmos fazem diferença expressiva nos resultados.

Para uma revisão dos testes formais na identificação de *outliers*, veja Barnett e Lewis (2000).

6 TRANSFORMAÇÕES PARA APROXIMAR DE UMA NORMALIDADE

Se a normalidade não é uma suposição viável, uma alternativa seria ignorar os resultados da análise e prosseguir como se os dados fossem normalmente distribuídos. Esta prática não é recomendada, uma vez que pode levar a conclusões incorretas. Uma segunda alternativa é transformar os dados originais para se chegar aproximadamente a uma distribuição normal. Formalmente, transformações são nada mais que uma nova expressão dos dados em unidades diferentes. Por exemplo, quando um histograma de observações positivas exibe uma longa calda à direita, ou uma distribuição achatada, é possível transformar a variável tomando o logaritmo ou raiz quadrada. Talvez esse procedimento matemático possa melhorar a simetria sobre a média e se aproximar de uma distribuição normal. Ademais, essas novas unidades fornecem expressões mais "naturais" das características a serem estudadas.

Transformações apropriadas são sugeridas por (a) considerações teóricas e/ou (b) dados propriamente. As transformações de dados de contagem são frequentemente feitas por raiz quadrada. Transformações logísticas (logit) são aplicadas às proporções. Por sua vez, transformações-z de Fisher são feitas para produzir coeficientes de correlação, que podem aproximar os dados de distribuição normal.

Helpful Transformations To Near Normality

Original Scale	*Transformed Scale*	
1. Counts, y	$\sqrt{y}$	
2. Proportions, $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$	(4-33)
3. Correlations, r	Fisher's $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$	

Os casos mais comuns seriam: $\sqrt{x}$, x^{-1}, $\ln(x)$. Lembre-se que o logaritmo de qualquer número negativo ou nulo é indefinido. Neste caso, pode-se adicionar uma constante (k) para tornar todos os valores positivos, desde que $k > \min(x)$. Para x^{λ} com $\lambda = -1$ teria uma relação recíproca; com $\lambda = 1/2$ geraria $\sqrt{x}$; com $\lambda = 0$, definir-se-ia $x^0 = \ln(x)$.

Para selecionar um expoente de transformação, o pesquisador deve visualizar um histograma e decidir se grandes valores devem ser puxados ("*pulled in*") ou empurrados ("*pushed out*") para melhorar a simetria da distribuição. A escolha final seria examinar um gráfico Q-Q a fim de averiguar se a tentativa de normalidade é satisfatória.
Ademais, as transformações discutidas assumem que somente a aparência dos dados influencia a escolha de uma apropriada transformação. Dessa maneira, inexistem considerações externas envolvidas.

Um conveniente método analítico é disponível para escolher o expoente de transformação. O método Box e Cox considera uma leve modificação do expoente de transformação:

$$x^{\lambda} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

que é contínuo em λ para $x > 0$. Considerando as observações $x_1, x_2, x_3, \ldots, x_n$, a solução Box-Cox escolhe um valor apropriado de λ que maximiza a expressão:

$$l(\lambda) = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{j=1}^{n}\left(x_j^{(\lambda)} - \bar{x}_j^{(\lambda)}\right)^2\right] + (\lambda - 1)\sum_{j=1}^{n}\ln x_j$$

em que $\bar{x}_j^{(\lambda)}$ é a média aritmética das observações transformadas pelo expoente λ, ou seja:

$$\bar{x}_j^{(\lambda)} = \frac{1}{n}\sum_{j=1}^{n}x_j^{(\lambda)} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{x^{\lambda} - 1}{\lambda}\right)$$

O primeiro termo de $l(\lambda)$ é, além da constante, o logaritmo de uma função verossimilhança normal, maximizando com respeito à média populacional e os parâmetros da variância. Se $\widetilde{\lambda} \cong 0{,}5$, é mais simples aplicar a raiz quadrada na variável. O Stata cria uma nova variável, como recomendado pelos estatísticos.

Em suma a transformação pode melhorar a distribuição de uma variável para uma normal. Contudo, não existem garantias que o método Box-Cox produzirá um conjunto de valores normalmente distribuídos. Deve-se ser cuidadoso ao avaliar possíveis violações dessa suposição de normalidade. O Stata fornece o método Box-Cox, de Escada de potência (Ladder) para somente valores positivos, transformação log de assimetria zero para valores negativos ou nulos. Este último método, $\ln(\pm\exp - k)$, encontra o valor da constante (k) e o sinal do expoente de forma que a assimetria da nova variável seja zero. Com as observações multivariadas, o expoente de transformação deve ser feito para cada variável.

Por fim, cabe mencionar a questão dos valores *"missing"*, que podem ocorrer tanto para a observação quanto para uma determinada variável. Para tratá-los, a decisão deve ser feita sobre como obter um completo conjunto de dados para a análise multivariada. Existem duas regras básicas:

a) se uma variável está faltando em uma alta proporção de casos, então a variável deve ser deletada;
b) se um caso está faltando em muitas variáveis, que são cruciais para sua análise, então o caso deve ser excluído.

Valores faltantes (missing) podem ocorrer por vários motivos. Por exemplo, o entrevistado com renda alta pode se indispor a responder o valor do seu salário em uma pesquisa. A melhor maneira de lidar com observações incompletas, ou em falta valores, depende, em grande medida, do contexto da pesquisa. Se o padrão de valores faltantes está intimamente ligado ao valor da resposta, como no exemplo supracitado, as inferências subsequentes sobre os dados devem ser fortemente enviesadas. Para estes tipos casos, não há técnicas estatísticas desenvolvidas para trata-los. No entanto, é possível tratar de situações em que os dados são faltantes ao acaso (aleatório), isto é, casos em que a falta de informação não tinha sido influência pela característica da variável. Nesses casos, pode-se usar o algoritmo de máxima verossimilhança para dados incompletos, indicado por Dempster, Laird, e Rubin (1977). Essa técnica, denominada de algoritmo EM, consiste em um cálculo iterativo com dois passos: a) etapa preditiva e b) etapa de estimação. Na primeira etapa, preditiva, dada alguma estimativa dos parâmetros desconhecidos, prevê a contribuição de qualquer observação faltante para as estatísticas suficientes (de dados completo). Por sua vez, na segunda etapa, usam-se as estatísticas suficientes previstas para calcular e revisar as estimativas dos parâmetros. Para maiores detalhes sobre esse algoritmo, veja o exemplo 5.13 em de Johnson e Wichern (2002, p.253).

Cuidado. O algoritmo de predição-estimação é desenvolvido na base na hipótese que os valores faltantes correram por acaso (aleatório). Se os valores faltantes estão relacionados com os níveis de resposta, então manipulá-los, pode introduzir vieses graves nos procedimentos de estimação. Geralmente os valores faltantes estão relacionados com as respostas a serem medidas. Por conseguinte, é preciso ser sempre duvidoso com qualquer sistema computacional que preencham os valores como se os mesmo fossem perdidos de forma aleatória. Na existência de muitos valores faltantes, é imperativo que o pesquisador busque as causas sistêmicas que os criaram.