

# Lista de Exercícios - Normal Multivariada

Renato Assunção - DCC, UFMG

2017

Estes exercícios exploram as distribuições marginais e condicionais associadas com uma normal multivariada. Lembre-se: se  $\mathbf{A}$  é uma matriz de constantes e  $\mathbf{Y}$  um vetor aleatório com  $\mathbb{E}(\mathbf{Y}) = \boldsymbol{\mu}$  e matriz de covariância  $\text{Cov}(\mathbf{Y}) = \boldsymbol{\Sigma}$ , então

$$\mathbb{E}(\mathbf{AY}) = \mathbf{A}\boldsymbol{\mu} = \mathbf{A} \mathbb{E}(\mathbf{Y})$$

e

$$\text{Cov}(\mathbf{AY}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \mathbf{A} \text{Cov}(\mathbf{Y}) \mathbf{A}'$$

1. Seja  $\mathbf{Z} = (Z_1, Z_2, Z_3)$  um vetor de variáveis i.i.d. (independentes e identicamente distribuídas)  $N(0, 1)$ . Isto é,  $\mathbf{Z}$  segue uma distribuição normal multivariada com valor esperado  $\boldsymbol{\mu} = (0, 0, 0)$  e matriz  $3 \times 3$  de covariância igual à identidade  $\mathbf{I}$ . Para gerar uma instância do vetor  $\mathbf{Z}$  em R, basta usarmos o comando `z = rnorm(3)`. Para criarmos uma amostra de tamanho  $n = 235$  de instâncias independentes do vetor  $\mathbf{Z}$ , organizados numa matriz  $235 \times 3$ , basta digitar `z = matrix(rnorm(235*3), ncol=3)`.

Queremos agora gerar um vetor aleatório  $\mathbf{X} = (X_1, X_2, X_3)$  seguindo uma normal multivariada com valor esperado  $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3) = (10, 20, -50)$  e com matriz de covariância

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 9 & -14 \\ 9 & 30 & -44 \\ -14 & -44 & 94 \end{bmatrix}$$

Seja  $\rho$  a matriz  $3 \times 3$  de correlação com elemento  $\rho_{ij} = \text{Corr}(X_i, X_j) = \text{Cov}(X_i, X_j) / \sqrt{\sigma_{ii}\sigma_{jj}}$ . Observe que  $\rho_{ii} = 1$ . Obtenha  $\rho$  a partir da matriz  $\boldsymbol{\Sigma}$

Para gerar a amostra de  $\mathbf{X}$ , siga os seguintes passos em R:

- Encontre uma matriz  $L$  tal que  $\mathbf{LL}^t = \boldsymbol{\Sigma}$ . Uma matriz com esta propriedade é aquela obtida pela decomposição de Cholesky de matrizes simétricas e definidas positivas. Em R, isto é obtido pelo comando `L = t(chol(Sigma))`.
- Gere  $\mathbf{z}$ , um vetor 3-dim com v.a.'s iid  $N(0, 1)$ .
- A seguir, faça

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L} \%*\% \mathbf{z}$$

Gere uma amostra de tamanho 200 dos vetores  $\mathbf{x}$  3-dim e armazene numa matriz `amostra` de dimensão  $200 \times 3$ . A seguir, calcule a média aritmética dos 200 valores de cada coordenada de  $\mathbf{x}$  e compare com os três valores do vetor  $\boldsymbol{\mu}$ . Eles devem ser parecidos.

Usando a amostra, estime os 9 valores da matriz de covariância  $\boldsymbol{\Sigma}$ . Chame esta matriz estimada de  $S$ . Verifique que as estimativas são próximas dos valores verdadeiros que você usou para gerar seus dados. Por exemplo, estime o elemento  $\sigma_{12}$  da matriz  $\boldsymbol{\Sigma}$  por

$$s_{12} = \frac{1}{200} \sum_{i=1}^{200} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

onde  $\bar{x}_1$  e  $\bar{x}_2$  são as médias aritméticas dos 200 valores observados das v.a.'s 1 e 2. Os termos  $\sigma_{jj}$  da diagonal principal são estimados por

$$s_{jj} = \frac{1}{200} \sum_{i=1}^{200} (x_{ij} - \bar{x}_j)^2$$

O comando `cov(x)` calcula a matriz  $\mathbf{S}$  diretamente (usando 199 no denominador, ao invés de 200). Porcure calcular você os termos da matriz  $S$  para ter certeza de que você está entendendo o que estamos fazendo.

A matriz  $\Sigma$  é estimada a partir dos dados substituindo o operador teórico e probabilístico  $\mathbb{E}$  pela média aritmética dos números específicos da amostra. Assim,  $\sigma_{ij}$  é estimado por sua versão empírica  $s_{ij}$ . Qual a diferença entre  $\sigma_{ij}$  e  $s_{ij}$ ? Uma maneira de responder a isto é notar que  $s_{ij}$  vai ter um valor ligeiramente diferente cada vez que uma nova amostra for gerada, mesmo que a distribuição de probabilidade permaneça a mesma. Já  $\sigma_{ij}$  não vai mudar nunca, é fixo e determinado pela distribuição de probabilidade.

Seja  $\rho$  a matriz  $3 \times 3$  de correlação com elemento  $\rho_{ij} = \text{Corr}(X_i, X_j) = \text{Cov}(X_i, X_j) / \sqrt{\sigma_{ii}\sigma_{jj}}$ . Observe que  $\rho_{ii} = 1$ . Esta matriz  $\rho$  é estimada pela matriz  $\mathbf{R}$ , cujos elementos são obtidos a partir dos dados da amostra. Assim,  $\rho_{ij}$  é estimado por

$$r_{ij} = s_{ij} / \sqrt{s_{ii}s_{jj}}$$

Calcule as matrizes  $\rho$  e  $\mathbf{R}$  e compare-as.

Este é um dos sentidos em que empregamos à expressão *aprendizagem*: usamos os dados observados para aprender (ou inferir) sobre o mecanismo aleatório que gerou estes mesmos dados. Isto é, aprendemos sobre  $\mu$  e  $\Sigma$  através de  $(\bar{x}_1, \bar{x}_2)$  e de  $\mathbf{S}$ .

2. *Entendendo a variabilidade de  $\mathbf{R}$ .* Você viu no exercício anterior que  $\rho \neq \mathbf{R}$ . A matriz  $\rho$  não muda enquanto sua estimativa  $\mathbf{R}$  dependen da amostra instanciada da distribuição. Até onde  $\mathbf{R}$  pode ir? Quão diferentes podem ser  $\rho$  e  $\mathbf{R}$ ?

Simule a matriz de dados `amostra` de dimensão  $200 \times 3$  um grande número de vezes. Digamos, simule `amostra` 5000 vezes. Em cada uma dessas simulações de `amostra`, calcule a matriz de correlação empírica  $\mathbf{R}$ . Faça um histograma dos 5000 valores obtidos para  $R_{ij}$ , um gráfico-histograma separado para cada par  $(i, j)$ .

Os valores de  $R_{ij}$  oscilam em torno do correspondente  $\rho_{ij}$ ? Qual o desvio-padrão aproximado de cada  $R_{ij}$ ? Pode obter este DP no olhómetro.

3. Seja  $\mathbf{X}' = (X_1, X_2, X_3, X_4)$  um vetor aleatório com vetor esperado  $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu} = (0, 1, 0, -1)'$  e matriz de covariância

$$\Sigma = \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 9 & -2 \\ 2 & 0 & -2 & 4 \end{bmatrix}$$

Particione  $\mathbf{X}$  como

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}.$$

Defina

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \end{bmatrix} \quad \text{e} \quad \mathbf{B} = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix}$$

e as combinações lineares  $\mathbf{AX}^{(1)}$  e  $\mathbf{BX}^{(2)}$ . Obtenha os seguintes elementos:

- A matriz de correlação  $\boldsymbol{\rho}$  de  $\mathbf{X}$ .
- $\mathbb{E}(\mathbf{X}^{(1)})$
- $\mathbb{E}(\mathbf{A}\mathbf{X}^{(1)})$
- $\text{Cov}(\mathbf{X}^{(1)})$
- $\text{Cov}(\mathbf{A}\mathbf{X}^{(1)})$
- $\mathbb{E}(\mathbf{X}^{(2)})$
- $\mathbb{E}(\mathbf{B}\mathbf{X}^{(2)})$
- $\text{Cov}(\mathbf{X}^{(2)})$
- $\text{Cov}(\mathbf{B}\mathbf{X}^{(2)})$

4. Seja  $\mathbf{X} = (X_1, X_2, X_3)'$  um vetor aleatório com distribuição normal multivariada com  $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)' = [-1, 0, 2]'$  e

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Obtenha a distribuição marginal de cada uma das v.a.'s

$$Y_1 = \frac{1}{4}X_1 - \frac{1}{4}X_2 + \frac{1}{2}X_3$$

e de

$$Y_2 = \frac{1}{4}X_1 + \frac{1}{4}X_2 - \frac{1}{2}X_3$$

Obtenha também a distribuição conjunta de  $(Y_1, Y_2)$ .

DICA: Escreva  $(Y_1, Y_2)$  como  $\mathbf{A}\mathbf{X}$  onde  $\mathbf{A}$  é uma matriz  $2 \times 3$  de constantes.

5. Seja  $\mathbf{X} = (X_1, X_2, X_3)'$  um vetor aleatório com distribuição normal multivariada com  $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)' = [-1, 0, 2]'$  e

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Quais das seguintes variáveis aleatórias são independentes?

- $X_1$  e  $X_2$
  - $X_2$  e  $X_3$
  - $(X_1, X_2)$  e  $X_3$
  - $(X_1 + X_2)/2$  e  $X_3$
  - $X_2$  e  $X_2 + 5X_1/2 - X_3$
6. Leia os slides 165 e seguintes do material `Top09-NormalMult.pdf`. Eles apresentam o uso da distância aleatória de Mahalanobis para detecção de anomalias. A distância de Mahalanobis entre um ponto aleatório gaussiano  $\mathbf{X}$  em  $\mathbb{R}^p$  e o seu perfil esperado  $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$  é dada por

$$D^2 = (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}).$$

onde  $\boldsymbol{\Sigma}$  é a matriz  $p \times p$  de covariância de  $\mathbf{X}$ . Lembre-se que a densidade da normal multivariada é baseada nesta medida de distância.

Como  $\mathbf{X}$  é um vetor aleatório gaussiano, a medida  $D^2$  é um número aleatório: possui uma faixa de valores possíveis e probabilidades associadas.

- A quantidade  $D^2$  tem um valor típico (ou valor esperado):  $\mathbb{E}(D^2) = ??$ .
- $D^2$  possui um afastamento típico de seu valor esperado, seu DP. O desvio-padrão de  $D^2$  é:  $\sqrt{\mathbb{V}(D^2)} = ??$ .
- Mais que isto, não somente estes dois resumos da distribuição de  $D^2$  são conhecidos mas a própria distribuição de  $D^2$  é conhecida.  $D^2 \sim ??$ .
- Fixando uma constante  $c$  qualquer, o conjunto de pontos  $\mathbf{x} \in \mathbb{R}^p$  que satisfazem  $D^2 = c$  formam um elipsóide em  $p$  dimensões. Isto é, os pontos  $\mathbf{x}$  que estão a uma distância  $D^2$  igual a  $c$  do seu perfil esperado formam um elipsóide. Quais são os eixos deste elipsóide e os seus tamanhos relativos?
- É possível mostrar que, com probabilidade  $1 - \alpha$ , o vetor aleatório  $\mathbf{X}$  deve cair dentro da elipse  $D^2 = c$  onde  $c = \chi_p^2(\alpha)$  é o quantil  $(1 - \alpha)100\%$  de uma distribuição qui-quadrado com  $p$  graus de liberdade onde  $p$  é a dimensão do vetor  $\mathbf{X}$ . No caso particular de um vetor bidimensional, o valor de  $c$  associado com a probabilidade  $1 - \alpha = 0.95$  é igual a  $c = 9.21$ . Assim, se  $\mathbf{X} = (X_1, X_2)$  estiver fora dessa elipse (isto é, se  $D^2 > 9.21$ ), o ponto pode ser consirado um tanto anômalo ou extremo.

O arquivo `stiffness.txt` contém quatro tipos de medições da rigidez de pranchas de madeira. A primeira é obtida aplicando-se uma onda de choque através da prancha, a segunda aplicando-se uma vibração à prancha e as outras duas são obtidas por meio de testes estáticos. Assuma que cada as 4 medições em uma prancha são instâncias de um vetor  $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Estime o vetor  $\boldsymbol{\mu}$  e a matriz  $4 \times 4$   $\boldsymbol{\Sigma}$  usando os dados do amostra. A seguir, usando estes valores estimados como se fossem os verdadeiros valores de  $\boldsymbol{\mu}$  e  $\boldsymbol{\Sigma}$ , calcule o valor de  $D^2$  para cada ponto da amostra. Quais pontos parecem extremos? Olhando as variáveis INDIVIDUALMENTE ou em pares através de scatterplots seria possível detectar estes pontos extremos? Faça scatterplot dos dados para entender sua resposta.

- 
7. Considere os dados do data frame `iris` do R. Este é um famoso conjunto de dados na comunidade de aprendizagem de máquina. Ele foi analisado inicialmente por Ronald Fisher quando desenvolveu em 1936 a técnica de análise de componentes principais (PCA). Ele contém medições de 150 flores *iris*. Em cada flor, foram feitas quatro medidas: o comprimento e a largura das pétalas e das sépalas em centímetros. Além disso, cada uma das 150 flores pertence a uma de três espécies distintas: *Iris setosa*, *Iris virginica* e *Iris versicolor*. São 50 flores de cada espécie. Veja detalhes em [https://en.wikipedia.org/wiki/Iris\\_flower\\_data\\_set](https://en.wikipedia.org/wiki/Iris_flower_data_set).

Os seguintes comandos geram um primeira visão dos dados:

```
head(iris) # 1as linhas do data frame iris
dim(iris) # dimensao do data frame
?iris     # help sobre o data frame
pairs(iris[,1:4]) # matriz de pares de scatterplots com as 150 flores
```

Notamos que em cada gráfico existem dois grupamentos de dados. Provavelmente, estes grupamentos correspondem a diferentes espécies de flores. Medidas de diferentes espécies costumam ter diferentes distribuições de probabilidade, com seus valores concentrados em diferentes intervalos. Para verificar esta afirmação, vamos colorir cada flor de acordo com sua espécie:

```
titulo = "Iris Data (red=setosa,green=versicolor,blue=virginica)"
pairs(iris[,1:4],main=titulo, pch=21, bg = iris$Species)
```

A espécie *virginica* é bem diferente das outras duas. Embora menos discrepantes, vemos claramente que cada uma dessas duas, *setosa* e *versicolor*, possuem medidas ocupando regiões diferentes em cada plot. Vamos analisar apenas uma das espécies, *setosa*, colocando os seus dados num novo data frame.

```

setosa <- iris[iris$Species == "setosa", 1:4]
pairs(setosa)           # plots de pares das 4 variaveis
apply(setosa, 2, mean)  # media aritmetica de cada variavel
cov(setosa)             # estimativa da matriz de covariancia
cor(setosa)             # estimativa da matriz de correlacao
round(cor(setosa), 2)   # valores arredondados em duas casas decimais

```

Estes dados são uma amostra de 50 exemplos do vetor aleatório  $\mathbf{X} = (X_1, X_2, X_3, X_4)$  onde  $X_1$  é o comprimento da sépala,  $X_2$  é a largura da sépala,  $X_3$  é o comprimento da pétala e  $X_4$  é a largura da pétala. Assuma que a distribuição conjunta do vetor  $\mathbf{X}$  é uma normal multivariada de dimensão 4 com parâmetros  $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$  e matriz de covariância  $\Sigma$  de dimensão  $4 \times 4$ . Use os resultados obtidos no R como estimativas para os valores desconhecidos do vetor  $\boldsymbol{\mu}$  e da matriz de covariância  $\Sigma$  e da matriz de correlação  $\mathbf{\rho}$ . A seguir, responda às seguintes questões:

- Forneça uma estimativa para o vetor  $\boldsymbol{\mu}$  e para a matriz  $\Sigma$  e  $\rho$ .
- A partir da matriz de correlação entre os pares de v.a.'s (e do plot de dispersão dos pontos), quais as variáveis que são mais correlacionadas? E quais são menos correlacionadas?
- Obtenha a distribuição MARGINAL do sub-vetor  $\mathbf{X}^* = (X_1, X_2)$ , o comprimento e largura da sépala.
- Obtenha a distribuição CONDICIONAL do sub-vetor  $\mathbf{X}^* = (X_1, X_2)$  quando são conhecidos os valores  $x_3$  e  $x_4$  das v.a.'s  $(X_3, X_4)$ . Obtenha esta distribuição para dois valores genéricos  $x_3$  e  $x_4$ . A seguir use dois valores específicos:  $x_3 = 1.8$  e  $x_4 = 0.6$ , dois valores relativamente altos para estas variáveis. Compare  $DP_1 = \sqrt{V(X_1)}$  com  $\sqrt{V(X_1|X_3 = 1.8, X_4 = 0.6)}$ , o desvio padrão da variável  $X_1$  condicionada nos valores de  $X_3$  e  $X_4$ .
- Obtenha agora a distribuição CONDICIONAL do sub-vetor  $\mathbf{X}^* = (X_1, X_2)$  quando é conhecido apenas o valor de  $X_3$ .
- Obtenha também distribuição CONDICIONAL do sub-vetor  $\mathbf{X}^* = (X_1, X_2)$  quando é conhecido apenas o valor de  $X_4$ .
- Comparando as três últimas respostas que você forneceu, qual das duas variáveis isoladamente,  $X_3$  ou  $X_4$ , diminui a incerteza acerca de  $X_2$  mais fortemente? Isto é, se você tivesse de escolher apenas uma delas,  $X_3$  ou  $X_4$ , qual você iria preferir se seu objetivo fosse prever o valor de  $X_2$ ?
- Considere a melhor preditora para  $X_2$  que você escolheu, dentre  $X_3$  ou  $X_4$ , na questão anterior. Digamos que tenha sido  $X_4$ . Avalie quanto conhecer a outra variável (neste caso,  $X_3$ ) reduz ADICIONALMENTE a incerteza acerca de  $X_2$ . Isto é, compare  $V(X_2|X_4)$  com  $V(X_2|X_3, X_4)$ .

- 
8. Considere um vetor  $\mathbf{X} = (X_1, X_2)$  com distribuição normal bivariada com vetor esperado  $\boldsymbol{\mu} = (\mu_1, \mu_2)$  e matriz de covariância

$$\Sigma = \begin{bmatrix} \sigma_{11} & \rho\sqrt{\sigma_{11}\sigma_{22}} \\ \rho\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} \end{bmatrix}$$

Usando o resultado dos slides, mostre que a distribuição condicional de  $(X_2|X_1 = x_1)$  é  $N(\mu_c, \sigma_c^2)$  onde

$$\mu_c = \mu_2 + \rho\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}(x_1 - \mu_1) = \mu_2 + \rho\sqrt{\sigma_{22}}\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}$$

e

$$\sigma_c^2 = \sigma_{22}(1 - \rho^2)$$

A partir desses resultados, verifique se as afirmações abaixo são V ou F:

- Saber que o valor  $X_1 = x_1$  está dois desvios-padrão acima de seu valor esperado (isto é,  $(x_1 - \mu_1)/\sqrt{\sigma_{11}} = 2$ ) implica que devemos esperar que  $X_2$  também fique dois desvios-padrão acima de seu valor esperado.
- Dado que  $X_1 = x_1$ , a variabilidade de  $X_2$  em torno de seu valor esperado é maior se  $x_1 < \mu_1$  do que se  $x_1 > \mu_1$ .
- Conhecer o valor de  $X_1$  (e assim eliminar parte da incerteza existente) sempre diminui a incerteza da parte aleatória permanece desconhecida (isto é, compare a variabilidade de  $X_2$  condicionada e não-condicionada no valor de  $X_1$ ).
- $\mu_c$  é uma função linear de  $x_1$ .

9. Neste exercício, você vai gerar alguns vetores gaussianos tri-dimensionais que, de fato vivem em duas dimensões.

```
require(MASS)
nsims=200
Sigma = matrix(c(3,2,2,4),2,2) # matriz de covariancia de (x1, x2)
pts = mvrnorm(nsims, c(1, 2), Sigma) # gerando nsims vetores (x1, x2)

pts = cbind(pts, 3*pts[,1]+4*pts[,2]) # criando x3
pairs(pts)

library(scatterplot3d)
scatterplot3d(pts)

library(rgl)
plot3d(pts, col="red", size=3)

A = matrix(c(1, 0, 0, 1, 3, 4), 3, 2, byrow=T)
var.pts = A %*% Sigma %*% t(A)

var.pts
round(cov(pts),2)

eigen(var.pts)
eigen(cov(pts))
```

- O comando `mvrnorm` gera uma amostra de tamanho `nsims` do vetor BI-DIMENSIONAL  $(X_1, X_2)$ . Qual a distribuição desse vetor?
- A seguir, é criada uma amostra do vetor TRI-dimensional  $\mathbf{pts} = \mathbf{X} = (X_1, X_2, X_3)$  que é obtido por meio de uma transformação linear do vetor  $(X_1, X_2)$ :

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = A \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 4 \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ 3X_1 + 4X_2 \end{bmatrix}$$

Quais os parâmetros da distribuição gaussiana do vetor  $\mathbf{pts}$ ?

- Qual a diferença entre os objetos `var.pts` e `cov(pts)`? O que cada um deles representa?
- Por que o comando `round(cov(pts),2)` não gera exatamente a matriz de covariância de  $\mathbf{X}$ ? Ignore os possíveis erros devido à precisão finita da representação numérica no computador.
- Qual o menor autovalor da matriz de covariância de  $\mathbf{X}$ ? E de `cov(pts)`?

Agora, um conjunto de dados simulado que não está exatamente mas sim, quase completamente contido num plano do  $\mathbb{R}^3$ .

```
x1 <- rnorm(1000)
x2 <- rnorm(1000)
x3 <- 3 + 1.2*x1 - 2.3*x2 + rnorm(1000, sd=0.3)
d2 <- data.frame(x1,x2,x3)
open3d()
plot3d(d2)

eigen(cov(d2))
```

Isto mostra que não precisamos realmente de 3 dimensões. Esta massa de pontos vive praticamente num espaço de dimensão 2, o das duas primeiras coordenadas. A terceira coordenada é praticamente uma combinação linear das duas primeiras. O menor autovalor é igual a 0.011 e é 87 vezes menor que o segundo autovalor. Isto é, o menor autovalor é um valor próximo de zero e muito pequeno relativamente aos demais autovalores.

- Qual a distribuição de  $(X_1, X_2)$ ?
- Qual a distribuição de  $(X_1, X_2, X_3)$ ?

Agora, um exemplo com dados reais. Comece lendo a descrição dos dados:

```
?trees # girth = circunferencia 'a altura do peito
pairs(trees)
scatterplot3d(trees) # requer pacote scatterplot3d
plot3d(trees, col="red", size=3) # requer pacote rgl
eigen(cov(trees))
```

O menor autovalor é 0.56, que não é tão próximo de zero quanto nos exemplos extremos anteriores. Entretanto, pelo plot tri-dimensional, notamos que esta nuvem de pontos em  $\mathbb{R}^3$  vive aproximadamente num plano, um sub-espaço de dimensão 2.

- 
10. Neste exercício, você vai trabalhar com os dados de uma análise química de vinhos. Você vai ler uma matriz com 178 amostras de diferentes vinhos. Haverá uma linha para cada vinho. A primeira coluna indica o cultivar do vinho (entenda como o tipo de uva usado na fabricação do vinho) tal como Sauvignon Blanc, Cabernet ou Chardonnay (rotulados como 1, 2 ou 3). As 13 colunas seguintes contêm as concentrações de 13 diferentes compostos químicos na amostra.

O objetivo é diferenciar entre os 3 tipos de vinho com base na sua composição química representada pelo vetor 13-dimensional  $\mathbf{X}$ . Você precisa criar uma regra para predizer o tipo de vinho (a primeira coluna) a partir das 13 variáveis da composição química. Vamos verificar que, ao invés de usarmos todas as 13 variáveis, poderemos nos basear em dois índices-resumo, os dois primeiros PCAs. Eles resumem quase toda a informação (ou variabilidade) simultânea das 13 variáveis.

Estude o script R abaixo. De propósito, ele tem uma quantidade *mínima* de comentários. Procure identificar o que cada linha está fazendo.

WARNING: o help da função `prcomp` é confuso, misturando PCA e análise fatorial nas explicações.

```
arq = "http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"
wine=read.table(arq, sep=",")

head(wine)
```

```

pairs(wine[,2:6])
round(100*cor(wine[,2:14]))
round(apply(wine[,2:14], 2, sd),2)

wine.pca = prcomp(wine[,2:14], scale. = TRUE)
summary(wine.pca)
wine.pca$sdev
sum((wine.pca$sdev)^2)
screeplot(wine.pca, type="lines")

# Barplot das variancias acumuladas
barplot(cumsum(wine.pca$sdev^2)/sum(wine.pca$sdev^2))
# os dois primeiros PCA's explicam aprox 60% da variancia total
# os 5 primeiros explicam aprox 80%

# Os autovetores
dim(wine.pca$rot)

# O 1o autovetor
wine.pca$rot[,1]

# O 2o autovetor
wine.pca$rot[,2]

# Coordenadas dos pontos ao longo do primeiro componente
fscore1 = wine.pca$x[,1]

# Coordenadas dos pontos ao longo do segundo componente
fscore2 = wine.pca$x[,2]

# plot dos pontos projetados
plot(fscore1, fscore2, pch="*", col=wine[,1]+8)

```

Seja  $\mathbf{X}_i = (X_{i1}, \dots, X_{i13})$  a *linha*  $i$  da matriz `wine`. Seja  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{i13})$  a *linha*  $i$  da matriz `wine` PADRONIZADA. Isto é,  $Z_{ij} = (X_{ij} - \bar{x}_j)/s_j$  onde  $\bar{x}_j$  é a média aritmética e  $s_j$  é o desvio-padrão da coluna  $j$  da matriz `wine`. Esta matriz padronizada é obtida com o comando `z = scale(wine[2:14])`:

```

z = scale(wine[2:14]) # colunas de wine padronizadas (desvios padronizados)
round(apply(z, 2, mean), 5) # media das colunas da matriz
round(apply(z, 2, sd), 5) # sd das colunas da matriz

```

Vamos considerar  $\mathbf{Z}_i$  como um *vetor-coluna* 13-dimensional. Ao invés de usarmos o vetor  $\mathbf{Z}_i$ , estamos usando apenas o vetor  $\mathbf{Y}_i$  composto pelos dois índices formados pelos dois primeiros componentes principais:

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} = \begin{bmatrix} \mathbf{v}'_1 \mathbf{Z}_i \\ \mathbf{v}'_2 \mathbf{Z}_i \end{bmatrix}$$

onde  $\mathbf{v}_1$  e  $\mathbf{v}_2$  são os dois primeiros autovetores da matriz de correlação de  $\mathbf{X}$ .

- Preencha os locais com (??) com os valores numéricos corretos (duas casa decimais apenas):

$$\begin{aligned} Y_{i1} &= (??)Z_{i1} + (??)Z_{i2} + (??)Z_{i3} + \dots + (??)Z_{i,13} \\ Y_{i2} &= (??)Z_{i1} + (??)Z_{i2} + (??)Z_{i3} + \dots + (??)Z_{i,13} \end{aligned}$$



- O último gráfico do script R acima é um plot dos pontos  $\mathbf{Y}_i$  dos 178 vinhos. Identifique três regiões do plano  $Y_1, Y_2$  que podem ser usadas para classificar futuras amostras de vinhos em uma das três categorias. Pode apenas esboçar grosseiramente no gráfico a mão livre ou descrever as regiões em palavras.
- Suponha que uma nova amostra de vinho tem sua composição química medida e encontra-se

$$\mathbf{x} = (13.95, 3.65, 2.25, 18.4, 90.18, 1.55, 0.48, 0.5, 1.34, 10.2, 0.71, 1.48, 587.14)$$

Obtenha seu vetor  $\mathbf{z}$ , as suas coordenadas  $(y_1, y_2)$  e prediga o seu tipo. Confira sua resposta no final desta lista.

- O que podemos fazer com isto? Imagine uma medida de qualidade de cada um dos vinhos. Ela poderia ser obtida tomando-se a nota média que um conjunto de experts ou apreciadores de vinhos poderia dar a um dos vinhos. A seguir, podemos tentar, no espaço reduzido das duas componentes, entender quais os valores das componentes levam a um bom vinho. Diferentes combinações das 13 variáveis químicas podem levar a uma mesma nota alta.

## Respostas

1. Direto no R
2. Direto no R
3. Seja  $\mathbf{V} = \text{diag}(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{44})$  a matriz diagonal  $4 \times 4$  formada pelas variâncias de cada uma das 4 variáveis de  $\mathbf{X}$ . Então a matriz de correlação  $\boldsymbol{\rho}$  de  $\mathbf{X}$  é dada por

$$\begin{aligned} \boldsymbol{\rho} &= \mathbf{V}^{-1/2} \boldsymbol{\Sigma} \mathbf{V}^{-1/2} \\ &= \begin{bmatrix} 1/\sqrt{3} & & & \\ & 1 & & \\ & & 1/\sqrt{9} & \\ & & & 1/\sqrt{4} \end{bmatrix} \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 9 & -2 \\ 2 & 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & & & \\ & 1 & & \\ & & 1/\sqrt{9} & \\ & & & 1/\sqrt{4} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0.38 & 0.58 \\ 0 & 1 & 0.33 & 0 \\ 0.38 & 0.33 & 1.0 & -0.33 \\ 0.58 & 0 & -0.33 & 1 \end{bmatrix} \end{aligned}$$

Em R:

```
mat = matrix(c(3, 0, 2, 2, 0, 1, 1, 0, 2, 1, 9, -2, 2, 0, -2, 4), ncol=4)
round(diag(1/sqrt(diag(mat)))) %*% (mat %*% diag(1/sqrt(diag(mat)))) , 2)
```

- $\mathbb{E}(\mathbf{X}^{(1)}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
- $\mathbb{E}(\mathbf{A}\mathbf{X}^{(1)}) = \mathbb{E}(X_1 - X_2) = 0 - 1 = -1$
- $\text{Cov}(\mathbf{X}^{(1)}) = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$
- Observe que  $\mathbf{A}\mathbf{X}^{(1)} = X_1 - X_2$  é um escalar, uma variável aleatória, um vetor de dimensão 1. Portanto, a sua matriz de covariância é de dimensão  $1 \times 1$  contendo simplesmente a variância da v.a.:  $\text{Cov}(\mathbf{A}\mathbf{X}^{(1)}) = \text{Cov}(X_1 - X_2) = \mathbb{V}(X_1 - X_2)$ . Podemos obter esta variância com nossa

fórmula geral para obter a matriz de covariância de uma transformação linear de um vetor aleatório:

$$\begin{aligned}\text{Cov}(\mathbf{A}\mathbf{X}^{(1)}) &= \mathbf{A}\text{Cov}(\mathbf{X}^{(1)})\mathbf{A}' \\ &= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= 4 = \mathbb{V}(X_1 - X_2)\end{aligned}$$

- $\mathbb{E}(\mathbf{X}^{(2)}) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$
- Temos

$$\mathbf{B}\mathbf{X}^{(2)} = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} X_3 - X_4 \\ X_3 + 2X_4 \end{pmatrix}$$

e

$$\mathbb{E}(\mathbf{B}\mathbf{X}^{(2)}) = \mathbf{B}\mathbb{E}(\mathbf{X}^{(2)}) = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} 0 - (-1) \\ 0 + 2(-1) \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

- $\text{Cov}(\mathbf{X}^{(2)}) = \begin{bmatrix} 9 & -2 \\ -2 & 4 \end{bmatrix}$
- Temos

$$\begin{aligned}\text{Cov}(\mathbf{B}\mathbf{X}^{(2)}) &= \mathbf{B}\text{Cov}(\mathbf{X}^{(2)})\mathbf{B}' \\ &= \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 9 & -2 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 17 & -1 \\ -1 & 17 \end{bmatrix}\end{aligned}$$

4.  $Y_1$  possui distribuição gaussiana. Se  $\mathbf{c}_1 = (1/4, -1/4, 1/2)'$  então o valor esperado é igual a

$$\mathbb{E}(Y_1) = \mathbb{E}(\mathbf{c}_1'\mathbf{X}) = \mathbf{c}_1'\boldsymbol{\mu} = \frac{1}{4}\mu_1 - \frac{1}{4}\mu_2 + \frac{1}{2}\mu_3 = -1/4 + 0 + 2/2 = 3/4$$

e a variância é

$$\mathbb{V}(Y_1) = \mathbf{c}_1'\boldsymbol{\Sigma}\mathbf{c}_1 = \begin{bmatrix} 1/4 & -1/4 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1/4 \\ -1/4 \\ 1/2 \end{bmatrix} = 7/8$$

Similarmente, fazendo  $\mathbf{c}_2 = (1/4, 1/4, -1/2)'$ , encontramos  $Y_2 \sim N(-5/4, 5/8)$ .

O vetor  $(Y_1, Y_2)$  possui distribuição normal bivariada com quase todos os seus parâmetros já calculados. Falta apenas a correlação (ou covariância) entre  $Y_1$  e  $Y_2$ :

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2 \left( \begin{bmatrix} 3/4 \\ -5/4 \end{bmatrix}, \begin{bmatrix} 7/8 & ?? \\ ?? & 5/8 \end{bmatrix} \right)$$

O valor faltante é obtido facilmente como o elemento 21 da matriz:

$$\begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 \end{bmatrix}$$

que é igual a  $-3/4$ .

Outra opção mais simples é usar a propriedade da bilinearidade do operador covariância:

$$\text{Cov}\left(\sum_i a_i X_i, \sum_j b_j X_j\right) = \sum_{i,j} a_i b_j \text{Cov}(X_i, X_j) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{b}$$

Assim,

$$\begin{aligned}
\text{Cov}(Y_1, Y_2) &= \text{Cov}(X_1/4 - X_2/4 + X_3/2, X_1/4 + X_2/4 - X_3/2) \\
&= \mathbf{c}_1' \mathbf{\Sigma} \mathbf{c}_2 \\
&= \begin{bmatrix} 1/4 & -1/4 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ -1/2 \end{bmatrix} \\
&= -3/4
\end{aligned}$$

Portanto,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2 \left( \begin{bmatrix} 3/4 \\ -5/4 \end{bmatrix}, \begin{bmatrix} 7/8 & -3/4 \\ -3/4 & 5/8 \end{bmatrix} \right)$$

5. Numa normal multivariada  $\mathbf{Y} = (Y_1, \dots, Y_p)'$ , duas de suas variáveis aleatórias  $i$  e  $j$  são independentes se, e somente se, o elemento  $(i, j)$  da matriz de covariâncias (ou de correlações) é igual a zero. Para sub-vetores de  $\mathbf{Y}$  o mesmo resultado vale olhando-se para a matriz  $\mathbf{\Sigma}$  particionada. Assim,

- $X_1$  e  $X_2$ : não são independentes
- $X_2$  e  $X_3$ : são independentes
- $(X_1, X_2)$  e  $X_3$ : são independentes pois o bloco formado por  $\mathbf{\Sigma}_{1,3}$  e  $\mathbf{\Sigma}_{2,3}$  é igual a zero.
- $(X_1 + X_2)/2$  e  $X_3$ : são independentes pois  $g(X_1, X_2) = (X_1 + X_2)/2$  é uma função apenas de  $X_1$  e  $X_2$ , que são independentes de  $X_3$ .
- $X_2$  e  $g(X_1, X_2, X_3) = X_2 + 5X_1/2 - X_3$ : são independentes. Usando a bilinearidade da covariância, calculamos

$$\text{Cov}(X_2, X_2 + 5X_1/2 - X_3) = \text{Cov}(X_2, X_2) + (5/2)\text{Cov}(X_2, X_1) - \text{Cov}(X_2, X_3) = 5 + (5/2)(-2) - 0 = 0$$

6. É possível deduzir que, se  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ , então  $D^2$  segue uma distribuição qui-quadrado com  $p$  graus de liberdade. Isto permite obter  $\mathbb{E}(D^2) = p$  e também  $\mathbb{V}(D^2) = p$ . Os eixos do elipsóide estão na direção dos autovetores da matriz  $\mathbf{\Sigma}$  e com tamanhos proporcionais aos seus autovalores.

```

stiffness = matrix(scan("stiffness.txt"), ncol=5, byrow=T)
x = stiffness[,1:4]
mu = apply(x, 2, mean)
sigma = cov(x)

n = nrow(x)
desvio = x - matrix(mu, nrow=nrow(x), ncol=ncol(x), byrow=T)
d2meu = diag(desvio %*% (solve(sigma) %*% t(desvio)))
# comando acima calcula D2
# R possui um comando proprio (e mais eficiente) para isto: mahalanobis
d2 = mahalanobis(x, mu, sigma)

# verificando que meu comando ineficiente calculou a mesma coisa
plot(d2, d2meu)

# identificando as anomalias
anomalias = d2 > qchisq(0.95, 4)

x[anomalias,]

```

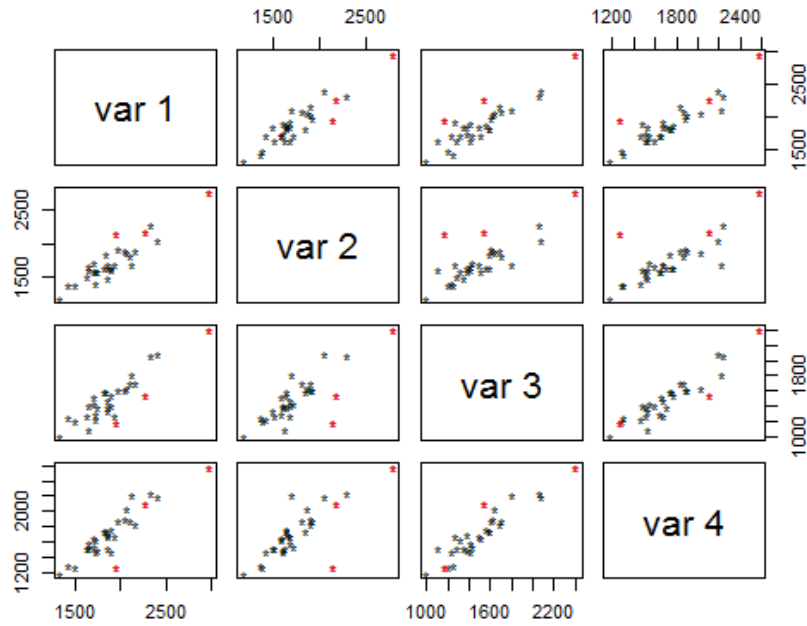


Figura 1: Scatterplot das variáveis de stiffness. Anomalias estão marcadas em vermelho.

```
nanom = sum(anomalias)
# plotando e marcando em vermelho as anomalias
pairs(rbind(x, x[anomalias,]), pch="*", col=rep(c("black", "red"), c(n, nanom)))
```

Scatterplot das 4 variáveis com as anomalias marcadas em vermelho estão na Figura 1.

#### 7. Para o problema das flores *setosa*:

- Estimativas de  $\mu$  e  $\Sigma$ :

```
> apply(setosa, 2, mean) # estimativa de mu
Sepal.Length Sepal.Width Petal.Length Petal.Width
      5.006      3.428      1.462      0.246
> round(cov(setosa),3) # estimativa de Sigma
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      0.124      0.099      0.016      0.010
Sepal.Width       0.099      0.144      0.012      0.009
Petal.Length      0.016      0.012      0.030      0.006
Petal.Width       0.010      0.009      0.006      0.011
```

- A partir da matriz de correlação  $\text{cor}(\text{setosa})$ , o comprimento  $X_1$  e a largura  $X_2$  das sépalas são as variáveis mais correlacionadas:  $\rho_{12} = 0.74$ . A largura da sépala  $X_2$  e o comprimento da pétala  $X_3$  são as menos correlacionadas, com  $\rho_{23} = 0.18$ .
- A distribuição do sub-vetor  $\mathbf{X}^* = (X_1, X_2)$ , o comprimento e largura da sépala, vem diretamente dos elementos 1 e 2 de  $\mu$  e do bloco da matriz  $\Sigma$ :

$$\mathbf{X}^* = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\mu[1:2], \Sigma[1:2, 1:2]) = N_2\left(\begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix}, \begin{bmatrix} 0.124 & 0.099 \\ 0.099 & 0.144 \end{bmatrix}\right)$$

- A distribuição condicional de  $\mathbf{X}^* = (X_1, X_2)$  quando são conhecidos os valores  $(x_3, x_4)$  das v.a.'s  $(X_3, X_4)$ .

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| \begin{pmatrix} X_3 = x_3 \\ X_4 = x_4 \end{pmatrix} \sim N_2(\mathbf{m}, \mathbf{V})$$

onde, usando a notação das notas de aula,

$$\begin{aligned} \mathbf{m} &= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \begin{pmatrix} x_3 - \mu_3 \\ x_4 - \mu_4 \end{pmatrix} \\ &= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{bmatrix} 0.016 & 0.010 \\ 0.012 & 0.009 \end{bmatrix} \begin{bmatrix} 0.030 & 0.006 \\ 0.006 & 0.011 \end{bmatrix}^{-1} \begin{pmatrix} x_3 - 1.462 \\ x_4 - 0.246 \end{pmatrix} \\ &= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{bmatrix} 0.399 & 0.712 \\ 0.247 & 0.702 \end{bmatrix} \begin{pmatrix} x_3 - 1.462 \\ x_4 - 0.246 \end{pmatrix} \end{aligned}$$

Para  $x_3 = 1.8$  e  $x_4 = 0.6$  temos

$$\mathbf{m} = \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{pmatrix} 0.387 \\ 0.332 \end{pmatrix} = \begin{pmatrix} 5.393 \\ 3.760 \end{pmatrix}$$

Quanto a matriz de covariância  $\mathbf{V}$  para a distribuição condicional, temos

$$\begin{aligned} \mathbf{V} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \\ &= \begin{bmatrix} 0.124 & 0.099 \\ 0.099 & 0.144 \end{bmatrix} - \begin{bmatrix} 0.016 & 0.010 \\ 0.012 & 0.009 \end{bmatrix} \begin{bmatrix} 0.030 & 0.006 \\ 0.006 & 0.011 \end{bmatrix}^{-1} \begin{bmatrix} 0.016 & 0.012 \\ 0.010 & 0.009 \end{bmatrix} \\ &= \begin{bmatrix} 0.110 & 0.088 \\ 0.088 & 0.134 \end{bmatrix}. \end{aligned}$$

Temos  $DP_1 = \sqrt{\mathbb{V}(X_1)} = \sqrt{0.124} = 0.352$  e  $\sqrt{\mathbb{V}(X_1|X_3 = 1.8, X_4 = 0.6)} = \sqrt{0.110} = 0.332$ .

- Queremos a distribuição condicional do sub-vetor  $\mathbf{X}^* = (X_1, X_2)$  quando é conhecido apenas o valor de  $X_3 = 1.8$ . Neste caso, é como se a variável  $X_4$  não existisse: ela não está envolvida. Vamos obter a distribuição conjunta do vetor  $(X_1, X_2, X_3)$  e então usar a nossa fórmula de condicional da normal multivariada. Para obter a distribuição marginal de  $(X_1, X_2, X_3)$ , basta olhar  $\boldsymbol{\mu}$  e  $\boldsymbol{\Sigma}$  e ignorar as entradas associadas com  $X_4$ .

Temos

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3 \left( \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right) = N_3 \left( \begin{bmatrix} 5.006 \\ 3.428 \\ 1.462 \end{bmatrix}, \begin{bmatrix} 0.124 & 0.099 & 0.016 \\ 0.099 & 0.144 & 0.012 \\ 0.016 & 0.012 & 0.030 \end{bmatrix} \right)$$

Dividindo este vetor em dois blocos, representados por letras em negrito e com indexação ligada aos blocos (e não às variáveis), podemos usar as fórmulas derivadas em sala de aula:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3 \left( \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right) = N_3 \left( \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

Temos

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| X_3 = 1.8 \sim N_2(\mathbf{m}, \mathbf{V})$$

onde, usando a notação das notas de aula,

$$\begin{aligned}
\mathbf{m} &= \mathbf{m}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{x}_2 - \mathbf{m}_2) \\
&= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} [\sigma_{33}]^{-1} (1.8 - \mu_3) \\
&= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{bmatrix} 0.016 \\ 0.012 \end{bmatrix} [0.030]^{-1} (1.8 - 1.462) \\
&= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{bmatrix} 0.533 \\ 0.400 \end{bmatrix} (1.8 - 1.462) \\
&= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{pmatrix} 0.180 \\ 0.135 \end{pmatrix} = \begin{pmatrix} 5.186 \\ 3.563 \end{pmatrix}
\end{aligned}$$

e

$$\begin{aligned}
\mathbf{V} &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} - \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} [\sigma_{33}]^{-1} \begin{bmatrix} \sigma_{31} & \sigma_{32} \end{bmatrix} \\
&= \begin{bmatrix} 0.124 & 0.099 \\ 0.099 & 0.144 \end{bmatrix} - \begin{bmatrix} 0.016 \\ 0.012 \end{bmatrix} [0.030]^{-1} \begin{bmatrix} 0.016 & 0.012 \end{bmatrix} \\
&= \begin{bmatrix} 0.115 & 0.093 \\ 0.093 & 0.139 \end{bmatrix}
\end{aligned}$$

- A distribuição condicional do sub-vetor  $\mathbf{X}^* = (X_1, X_2)$  quando  $X_4 = 0.6$  é obtida de forma idêntica ao item anterior: Temos

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big|_{X_4 = 0.6} \sim N_2(\mathbf{m}, \mathbf{V})$$

onde

$$\begin{aligned}
\mathbf{m} &= \mathbf{m}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{x}_2 - \mathbf{m}_2) \\
&= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{bmatrix} \sigma_{14} \\ \sigma_{24} \end{bmatrix} [\sigma_{44}]^{-1} (0.6 - \mu_4) \\
&= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{bmatrix} 0.010 \\ 0.009 \end{bmatrix} [0.011]^{-1} (0.6 - 0.246) \\
&= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{bmatrix} 0.909 \\ 0.818 \end{bmatrix} (0.6 - 0.246) \\
&= \begin{pmatrix} 5.006 \\ 3.428 \end{pmatrix} + \begin{pmatrix} 0.322 \\ 0.290 \end{pmatrix} = \begin{pmatrix} 5.328 \\ 3.718 \end{pmatrix}
\end{aligned}$$

e

$$\begin{aligned}
\mathbf{V} &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} - \begin{bmatrix} \sigma_{14} \\ \sigma_{24} \end{bmatrix} [\sigma_{44}]^{-1} \begin{bmatrix} \sigma_{41} & \sigma_{42} \end{bmatrix} \\
&= \begin{bmatrix} 0.124 & 0.099 \\ 0.099 & 0.144 \end{bmatrix} - \begin{bmatrix} 0.010 \\ 0.009 \end{bmatrix} [0.011]^{-1} \begin{bmatrix} 0.010 & 0.009 \end{bmatrix} \\
&= \begin{bmatrix} 0.115 & 0.091 \\ 0.091 & 0.136 \end{bmatrix}
\end{aligned}$$

- Temos  $\mathbb{V}(X_2|X_4 = 0.6) = 0.136 < 0.139 = \mathbb{V}(X_2|X_3 = 1.8)$ . Assim, saber que  $X_4 = 0.6$  leva a uma menor incerteza acerca do valor de  $X_2$  que aquela que resta quando  $X_3 = 1.8$ .

Para prever  $X_2$ , saber o valor de  $X_4$  é melhor que saber o valor de  $X_3$ . Observe que  $0.139 = \mathbb{V}(X_2|X_3 = 1.8) = \mathbb{V}(X_2|X_3 = x)$  para todo  $x$ , bem como  $0.136 = \mathbb{V}(X_2|X_4 = 0.6) = \mathbb{V}(X_2|X_4 = x)$  para todo  $x$ . Portanto, a conclusão sobre a maior redução da incerteza de  $X_2$  alcançada pelo conhecimento do valor de  $X_3$ , não depende dos valores específicos  $x_3 = 1.8$  e  $x_4 = 0.6$  usados no exercício. Teríamos a mesma conclusão com quaisquer dois valores para  $x_3$  e  $x_4$  pois as variâncias condicionais não variam com  $x_3$  e  $x_4$ .

- Entre  $X_3$  e  $X_4$ , a melhor preditora de  $X_2$  é  $X_4$ . Acrescentar o conhecimento sobre o valor de  $X_3$  ao conhecimento de que  $X_4 = 0.6$  reduz pouco a variabilidade (ou incerteza) acerca de  $X_2$ :

$$0.134 = \mathbb{V}(X_2|X_3 = 1.8, X_4 = 0.6) < \mathbb{V}(X_2|X_4 = 0.6) = 0.136 < \mathbb{V}(X_2) = 0.144$$

8. Usando a fórmula matricial para a distribuição condicional no caso bivariado, temos  $(X_2|X_1 = x_1) \sim N(\mu_c, \sigma_c^2)$  onde

$$\begin{aligned}\mu_c &= \mu_2 + \Sigma_{12}\Sigma_{22}^{-1}(x_1 - \mu_1) \\ &= \mu_2 + \rho\sqrt{\sigma_{11}\sigma_{22}} (1/\sigma_{11}) (x_1 - \mu_1) \\ &= \mu_2 + \rho\sqrt{\sigma_{22}} \frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\end{aligned}$$

e

$$\begin{aligned}\sigma_c^2 &= \Sigma_{22} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \sigma_{22} - \rho\sqrt{\sigma_{11}\sigma_{22}} (1/\sigma_{11}) \rho\sqrt{\sigma_{11}\sigma_{22}} \\ &= \sigma_{22} - \rho^2\sigma_{22} \\ &= \sigma_{22}(1 - \rho^2)\end{aligned}$$

Quanto às afirmações:

- F: Se  $(x_1 - \mu_1)/\sqrt{\sigma_{11}} = 2$ , o valor de  $X_2$  vai oscilar em torno de seu valor esperado condicional que será  $\mu_c = \mu_2 + \rho 2\sqrt{\sigma_{22}}$ . Como  $|\rho| < 1$ , temos o incremento  $|\rho 2\sqrt{\sigma_{22}}| < 2\sqrt{\sigma_{22}}$ , ou seja, menor que 2 desvios-padrões.
- F: pois  $\mathbb{V}(X_2|X_1 = x) = \sigma_c^2 = \sigma_{22}(1 - \rho^2)$  não depende de  $x$ .
- V: pois  $\mathbb{V}(X_2|X_1 = x) = \sigma_{22}(1 - \rho^2) < \sigma_{22} = \mathbb{V}(X_2)$  já que  $\rho^2 < 1$ .
- V: pois

$$\mathbb{E}(X_2|X_1 = x_1) = \mu_c = \mu_2 + \rho\sqrt{\sigma_{22}} \frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} = a + b(x_1 - \mu_1),$$

uma função linear de  $x_1$ .

9. O vetor  $(X_1, X_2)$  possui distribuição normal bivariada com vetor esperado  $\boldsymbol{\mu} = (1, 2)$  e matriz de covariância

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$$

O vetor  $\mathbf{X} = (X_1, X_2, X_3)$  possui distribuição normal multivariada de dimensão 3 com vetor esperado

$$\mathbb{E}(\mathbf{X}) = \mathbb{E}\left(\mathbf{A}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) = \mathbf{A}\mathbb{E}\left(\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 4 \end{bmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{bmatrix} 1 \\ 2 \\ 11 \end{bmatrix}$$

e matriz de covariância dada por

$$\mathbf{A}\Sigma\mathbf{A}' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 17 \\ 2 & 4 & 22 \\ 17 & 22 & 139 \end{bmatrix} =$$

O objeto `cov(pts)` contém a matriz  $\mathbf{A}\Sigma\mathbf{A}'$ . O objeto `var.pts` contém uma estimativa empírica desta matriz, uma estimativa baseada nas 200 instâncias de dados que você gerou. Para a amostra de tamanho `nsims`, estas duas matrizes são similares.

O comando `round(cov(pts),2)` calcula a estimativa empírica da matriz `var.pts = AΣA'`. Esta última matriz é fixa. A estimativa `cov(pts)` varia de amostra para amostra. Se `nsims` não for muito pequeno, `cov(pts)` e  $\mathbf{A}\Sigma\mathbf{A}'$  devem ser parecidas, como é o caso neste exercício.

Com a amostra gerada por mim, obtive `min(eigen(var.pts)$values)` igual a  $9.841374 \times 10^{-15}$  e `min(eigen(cov(pts))$values)` igual a  $6.915267 \times 10^{-15}$ . Os valores são próximos, ambos próximos de zero. O menor autovalor de `cov.pts` é exatamente zero, e isto pode ser verificado se tentamos fazer uma decomposição de Chleky:

```
> chol(var.pts)
Error in chol.default(var.pts) :
  the leading minor of order 3 is not positive definite
```

O algoritmo implementado em *R* para obter os autovalores de `var.pts` obtém apenas uma aproximação numérica para os reais autovalores e autovetores. De acordo com a página de help da função `eigen`, temos: Computing the eigendecomposition of a matrix is subject to errors on a real-world computer: the definitive analysis is Wilkinson (1965). All you can hope for is a solution to a problem suitably close to x. So even though a real asymmetric x may have an algebraic solution with repeated real eigenvalues, the computed solution may be of a similar matrix with complex conjugate pairs of eigenvalues.

O segundo bloco de comandos gera também uma gaussiana tri-dimensional. Como `x1 = rnorm(1000)` e `x2 = rnorm(1000)` geram independentemente vetores gaussianos  $N(0, 1)$  então  $(X_1, X_2) \sim N_2(\mathbf{0}_2, bsI_2)$  onde  $bs\mathbf{0} = (0, 0)'$  e  $I_2$  é a matriz identidade  $2 \times 2$ .

O vetor  $(X_1, X_2, X_3)$  tem a distribuição de suas duas primeiras coordenadas já determinadas acima:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3 \left( \begin{bmatrix} 0 \\ 0 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \sigma_{13} \\ 0 & 1 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right)$$

A terceira coordenada  $X_3 = 3 + 1.2X_1 - 2.3X_2 + \epsilon$  onde  $\epsilon \sim N(0, 0.3^2)$ , independente de  $X_1$  e  $X_2$ . Assim, os elementos que faltam para determinar a distribuição de  $(X_1, X_2, X_3)$  são os seguintes:

$$\mathbb{E}(X_3) = \mathbb{E}(3 + 1.2X_1 - 2.3X_2 + \epsilon) = 3 + 1.2\mathbb{E}(X_1) + 2.3\mathbb{E}(X_2) + \mathbb{E}(\epsilon) = 3 + 0 + 0 + 0 = 3$$

e, pela independência entre  $X_1$ ,  $X_2$  e  $\epsilon$ ,

$$\begin{aligned} \sigma_{33} &= \mathbb{V}(X_3) = \mathbb{V}(3 + 1.2X_1 - 2.3X_2 + \epsilon) \\ &= (1.2)^2\mathbb{V}(X_1) + (-2.3)^2\mathbb{V}(X_2) + \mathbb{V}(\epsilon) \\ &= 1.44 + 5.29 + 1.0 = 7.73 \end{aligned}$$

enquanto que

$$\begin{aligned} \sigma_{13} &= \sigma_{31} = \text{Cov}(X_1, X_3) \\ &= \text{Cov}(X_1, 3 + 1.2X_1 - 2.3X_2 + \epsilon) \\ &= (1.2)\text{Cov}(X_1, X_1) - 2.3\text{Cov}(X_1, X_2) + \text{Cov}(X_1, \epsilon) \\ &= 1.2\mathbb{V}(X_1) - 2.3 \times (0) + 0 = 1.2 \end{aligned}$$



e

$$\begin{aligned}\sigma_{23} &= \sigma_{32} = \text{Cov}(X_2, X_3) \\ &= \text{Cov}(X_2, 3 + 1.2X_1 - 2.3X_2 + \epsilon) \\ &= 1.2\text{Cov}(X_2, X_1) - 2.3\text{Cov}(X_2, X_2) + \text{Cov}(X_2, \epsilon) \\ &= 1.2 \times 0 - 2.3\mathbb{V}(X_2) + 0 = -2.3\end{aligned}$$

Assim,

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3 \left( \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1.2 \\ 0 & 1 & -2.3 \\ 1.2 & -2.3 & 7.73 \end{bmatrix} \right)$$

---

10. Para o problema do vinho:

```
x = c(13.95, 3.65, 2.25, 18.4, 90.18, 1.55, 0.48, 0.5, 1.34, 10.2, 0.71, 1.48, 587.14)
z = (x - apply(wine[,2:14], 2, mean))/apply(wine[,2:14], 2, sd)
y1 = sum( wine.pca$rot[,1] * z)
y2 = sum( wine.pca$rot[,2] * z)
plot(fscore1, fscore2, pch="*", col=wine[,1]+8)
points(y1, y2, pch="*", cex=4)
```