MÉTODOS ESTATíSTICOS EM DATA MINING

- Folha 1: Distribuição Multinormal e Análise em Componentes Principais -

1. Considere a distribuição $N_2(\mu, \Sigma)$, em que $\mu = (\mu_1, \mu_2)^T$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$. Mostre que

   (a) $p(x_1) \sim N(\mu_1, \sigma_1^2)$ (densidade marginal),

   (b) $p(x_1/x_2) \sim N(\mu_1 + \rho\sigma_1(x_2 - \mu_2)/\sigma_2, \sigma_1^2(1 - \rho^2))$ (densidade condicional)

2. Desenhe as curvas de nível da distribuição $N_2(\mu, \Sigma)$, em que $\mu = (\mu_1, \mu_2)^T$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$,

   (a) $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, ρ = -0.5, $d^2(X, \mu)$ = 1,4,9

   (b) $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\rho = 0.5$, $d^2(X, \mu)$ = 1,4,9

   (c) $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\rho = 1$, $d^2(X, \mu)$ = 1,4,9

3. Gere duas amostras aleatórias de duas populações multinormais:
   $\Pi_1 \sim N_4(\mu_1, \Sigma)$ e $\Pi_2 \sim N_4(\mu_2, \Sigma)$ em que: $n_1 = n_2 = 100$;
   $\mu_1 = (0, 0, 0, 0)^T$; $\mu_2 = (1, 0.7, 2.8, 1)^T$;
   $$\Sigma = \begin{bmatrix} 0.1953 & 0.0922 & 0.0997 & 0.0331 \\ & 0.1211 & 0.0472 & 0.0253 \\ & & 0.1255 & 0.0396 \\ & & & 0.0251 \end{bmatrix}$$

4. Calcule as componentes principais da matriz de dados do exercicío anterior (200 linhas e 4 colunas). Desenhe um diagrama de dispersão nas duas primeiras componentes principais e um outro nas segunda e terceira componentes principais.

5. Para o conjunto de dados image-segmentation (ver página web da cadeira em http://www.fc.up.pt/pessoas/jpcosta/MEDM.html), calcule as componentes principais para as variáveis contínuas existentes. Escolha o número de componentes principais que achar adequado, e construa um novo ficheiro de dados resultante de image-seg.data por substituição das variáveis contínuas pelas componentes principais (guarde este ficheiro pois mais tarde vai precisar dele)

6. Considere a matriz de covariância $\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$. Calcule as comp. principais usando a matriz Σ e depois a matriz de correlação R e compare os resultados.

7. No R use o conjunto de dados *state* sobre iliteracia e esperança de vida de 50 estados americanos. Use o comando *biplot* para criar um gráfico dos dados.

   (a) De acordo com o gráfico, que variáveis são positivamente correlacionadas com a var. High School Grad? E negativamente correlacionadas? Dê uma possivel explicação em cada caso.

   (b) Dentro do gráfico deve encontrar áreas vazias com poucos estados. O que são essas áreas e o que representam?

   (c) Existem também algumas direções onde os estados aparecem muito compactados. O que são e o que significa?

(d) Se os valores próprios associados aos dois primeiros vectores próprios explicarem uma pequena percentagem da variância total, quais são os perigos de interpretar o gráfico como fez?

8. Sejam $x_1, x_2, \ldots, x_5$ as taxas semanais de retorno de cinco empresas cotadas na bolsa de valores (as três primeiras quimicas e as duas restantes do ramo do petróleo). Numa amostra de 100 semanas, os dois primeiros valores próprios e vectores próprios da matriz de correlações obervados foram:

   $\hat{\lambda}_1 = 2.857 \quad \hat{e}_1^T = (.464, .457, .470, .421, .421)$
   $\hat{\lambda}_2 = .809 \quad \hat{e}_2^T = (.240, .509, .260, -.526, -.582)$

   (a) Escreva e interprete as variáveis correspondentes às duas primeiras componentes principais.

   (b) Qual a percentagem de variância explicada por estas componentes?

9. Os recordes de atletismo de 55 países incluem medições efectuadas em oito provas: 100 m (s), 200 m (s), 400 m (s), 800 m (min), 1500 m (min), 5 000 m (min), 10 000 m (min), maratona (min).

   (a) Diga como a ACP pode ser usada para obter uma representação bidimensional dos dados.

   (b) Os resultados de uma ACP estão na tabela abaixo. Interprete as duas primeiras componentes principais.

	PC1xλ_1	PC2xλ_2
100 m	0.82	0.50
200 m	0.86	0.41
400 m	0.92	0.21
800 m	0.87	0.15
1500 m	0.94	-0.16
5000 m	0.93	-0.30
10 000 m	0.94	-0.31
Maratona	0.87	-0.42
Valor próprio	6.41	0.89

   (c) Qual a percentagem de variância explicada pelas duas primeiras componentes principais?

10. Prove que o sub-espaço afim E_k correspondente às componentes principais maximiza o traço da matriz de dispersão dos pontos projectados.

2) em R:
Há um comando no R específico para elipses e que precisa de uma package especial (ellipse); não encontrei nada que desenhassse o conjunto de pontos que satisfaz uma dada equação.
install.packages(”ellipse”) (PRIMEIRO TIVE DE INSTALAR O PACKAGE)
library(ellipse)
plot(ellipse(x=matrix(c(1,-0.5*1*sqrt(2),-0.5*1*sqrt(2),2),2,2),
centre=c(1,2),level=0.95), type = ’l’)

NOTA: As matrizes são preenchidas por coluna

---

3)

library(MASS)
Sigma = matrix(c(0.1953,0.0922,0.0997,0.0331,0.0922,0.1211,0.0472,
0.0253,0.0997,0.0472,0.1255,0.0396,0.0331,0.0253,00.0396,0.0251),4,4)
amostra1=mvrnorm(100,c(0,0,0,0),Sigma)
amostra2=mvrnorm(100,c(1,0.7,2.8,1),Sigma)
amostra=rbind(amostra1,amostra2)

---

4)

pc=princomp(amostra,cor=FALSE)
summary(pc)
biplot(pc,choice=c(1,2))
biplot(pc,choice=c(2,3))

NOTA: FAZER ?princomp e ?biplot ou ?biplot.princomp para ver todas as possibilidades

---

5)

dados=read.table(”image-seg.data”,sep=” ”) (primeiro fazer o download dos dados e lê-los para o R)
pca=princomp(dados[,-1],cor=T,scores=T)
summary(pca)
dados1=data.frame(dados[,1],pca$scores[,1:7]) (guardar as 7 primeiras pcs em dados1 e juntar a 1a coluna com o nome das classes)

---

7)

data(state) (load state data)
summary(state)
state = state.x77[, 2:7] (extract out the useful pieces of information)
state[1:5,] (lets have a look)
state.pca = princomp(state,cor=TRUE) (calculate the pc’s of the data)
biplot(state.pca, pc.biplot=TRUE, cex=0.8, font=2, expand=0.9)