INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2018-19)

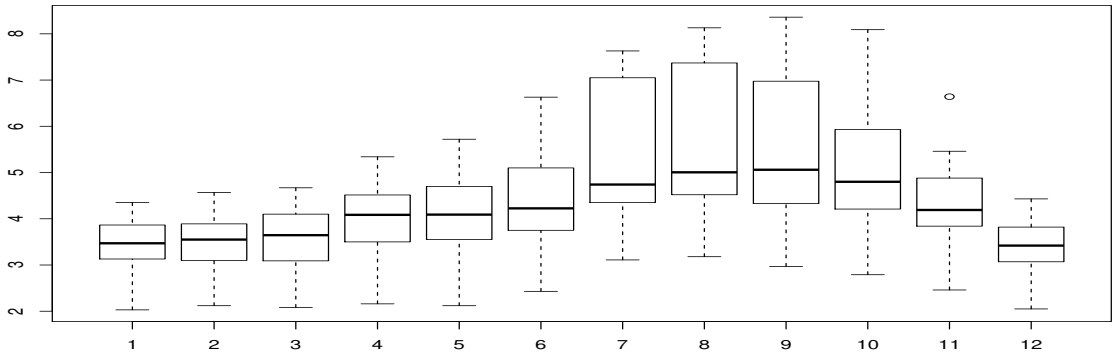
5 de Julho, 2019 **Exame Final** Duração: 3h30

A Floresta Experimental H.J. Andrews, no Estado norte-americano do Oregon, disponibiliza numerosos conjuntos de dados florestais, um dos quais tem a designação `CF002` e diz respeito a medições de várias variáveis em águas de pequenas bacias hidrográficas. As variáveis numéricas são médias mensais e os seus nomes terminam com a indicação `MO`. Na maioria dos casos são concentrações (em mg/l) de elementos ou compostos químicos, cujos símbolos ou fórmulas constituem a parte inicial do nome das variáveis. Há ainda medições de pH (variável `PH_MO`), de concentração de condutância específica (variável `COND_MO`, em micro-siemens por cm, $\mu S\, cm^{-1}$), de concentração de sedimentos suspensos (variável `SSED_MO`, em mg/l) e de caudal a dividir pela área da bacia hidrográfica (variável `Q_AREA_MO`, em cm). Existem ainda três factores: `YEAR`, indicando o ano de observação; `MONTH`, indicando a qual dos 12 meses do ano é que cada observação se reporta; e `SITECODE`, com o código de cada uma das oito bacias hidrográficas analisadas.

I [3,5 valores]

Os dados tratados apresentados no Anexo I são concentrações de potássio (K), sódio (NA), cálcio (CA) e cloro(CL) registadas ao longo de vários anos. Tenha-os em consideração na resposta **às perguntas 1 e 2**.

1. (a) Classifique, justificando, as variáveis em estudo.
   (b) Na Figura em baixo estão desenhados vários *boxplot* para a variável CA. Explique a que cada um se refere, compare-os e apresente os cálculos necessários à construção do *boxplot* correspondente a 11 (eixo das abcissas). Interprete este *boxplot*.
   (c) Indique, justificando convenientemente, um intervalo de confiança a 95% para o valor médio da concentração média mensal de cálcio. Poder-se-á dizer que aquele valor médio é 4 mg/l ? Justifique.
   (d) Pretende-se comparar a concentração média mensal de cálcio em Maio e em Outubro. Poder-se-á afirmar que em média a concentração média mensal de cálcio em Outubro é superior à de Maio? Justifique convenientemente, explicando que resultados usou do Anexo.



2. A distribuição gama parece ser um modelo adequado para caracterizar a variável *concentação média mensal de cálcio* que, por simplicidade, vamos designar por X. Continuando a simplificar, considere que há apenas um parâmetro desconhecido $\beta > 0$, sendo a função densidade definida como:

$$f(x|\beta) = \frac{1}{6\beta^4}\, x^3\, \mathrm{e}^{-\frac{x}{\beta}}, \quad \text{se} \;\; x > 0, \quad \text{e nula nos restantes valores de } x.$$

Nota: Sabe-se que $E[X] = 4\beta$ e $Var[X] = 4\beta^2$.

Considere que dispõe de uma amostra aleatória de dimensão n, $(X_1, X_2, .., X_n)$, associada a X.

(a) Obtenha o estimador de β pelo método dos momentos.

(b) Verifique se o estimador dos momentos obtido na alínea a) é um estimador centrado de β e indique o seu erro quadrático médio.

(c) Obtenha o estimador de máxima verosimilhança para β.

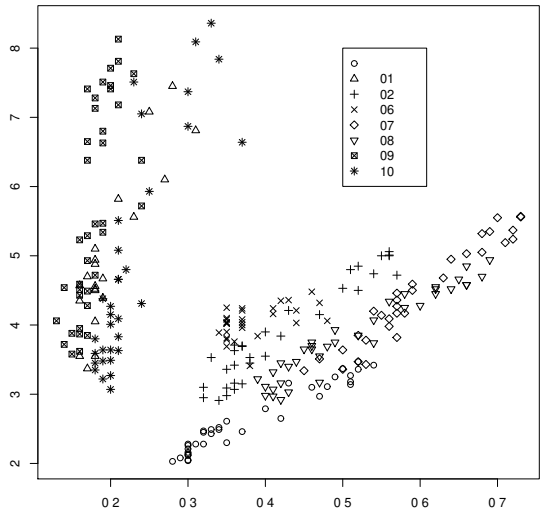
(d) Considerando os valores observados e os resultados apresentados no Anexo, relativos a uma amostra extraída daquela população determine uma estimativa para β.

II [7 valores]

Quer-se estudar a relação entre concentrações de cálcio (`CA_MO`) e potássio (`K_MO`). Dispõe-se da seguinte informação.

n	`K_MO`		`CA_MO`		correlação
	média	variância	média	variância	$r_{CA,K}$
260	0.3691538	0.02511667	4.265462	1.757796	-0.1208049

1. Calcule a declive da recta de regressão de `CA_MO` sobre `K_MO` e comente o seu sinal. Comente a qualidade da recta de regressão referida, mas tendo também em conta o gráfico em baixo, que é a nuvem de pontos das 260 observações dessas duas variáveis, usando símbolos diferentes consoante as oito localidades (bacias hidrográficas) de observação.



Tendo em conta semelhanças ambientais, os oito locais (bacias hidrográficas em `SITECODE`) foram divididos em três grupos, que definem um novo factor de nome `GRP3`. Eis alguma informação sobre os conjuntos resultantes.

Grupo	Locais	n_i	`K_MO`		`CA_MO`		correlação
			média	variância	média	variância	$r_{CA,K}$
1	GSWS01, GSWS09, GSWS10	92	0.2008	0.00204447	5.224	2.2483008	0.5657447
2	GSMACK, GSWS07, GSWS08	106	0.495	0.01575667	3.611	0.9244132	0.9698062
3	GSWS02, GSWS06	62	0.4039	0.00495854	3.962	0.2953915	0.8030870

2. Construa a tabela-resumo duma ANOVA que permita estudar eventuais efeitos dos grupos do factor GRP3 nas concentrações de cálcio (CA), indicando as suas contas. Comente os resultados do teste F resultante.

3. Foi ajustado um modelo ANCOVA, com os seguintes resultados:

```
Call:  lm(formula = CA_MO ~ K_MO * GRP3, data = dadosCF002)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.457318   0.368108   3.959 9.78e-05
K_MO         18.761058   1.789226  10.486  < 2e-16
GRP32        -1.523259   0.478885  -3.181  0.00165
GRP33         0.001557   0.682857   0.002  0.99818
K_MO:GRP32  -11.332819   1.887148  -6.005 6.58e-09
K_MO:GRP33  -12.562585   2.273859  -5.525 8.16e-08
---
Residual standard error: 0.7717 on 254 degrees of freedom
Multiple R-squared:  0.6677,  Adjusted R-squared:  0.6612
F-statistic: 102.1 on 5 and 254 DF,  p-value: < 2.2e-16

> vcov(dadosCF.ancova)
            (Intercept)      K_MO      GRP32      GRP33 K_MO:GRP32 K_MO:GRP33
(Intercept)   0.1355033 -0.642702 -0.1355033 -0.1355033  0.6427020   0.642702
K_MO         -0.6427020  3.201331  0.6427020  0.6427020 -3.2013312  -3.201331
GRP32        -0.1355033  0.642702  0.2293304  0.1355033 -0.8209006  -0.642702
GRP33        -0.1355033  0.642702  0.1355033  0.4662940 -0.6427020  -1.437967
K_MO:GRP32    0.6427020 -3.201331 -0.8209006 -0.6427020  3.5613283   3.201331
K_MO:GRP33    0.6427020 -3.201331 -0.6427020 -1.4379665  3.2013312   5.170437
```

   (a) Escreva o modelo ANCOVA, explicitando o significado de cada parcela da equação do modelo.

   (b) Qual a equação da recta ajustada para as observações do grupo 2? Comente o declive estimado.

   (c) Teste se é admissível considerar que as rectas populacionais dos grupos 2 e 3 são paralelas.

4. Considere uma Regressão Linear Múltipla envolvendo p variáveis preditoras e ajustada com base em n observações.

   (a) Descreva em pormenor o Modelo de Regressão Linear Múltipla, usando notação vectorial/matricial.

   (b) Deduza, justificando, a distribuição de probabilidades do vector $\vec{\hat{\mathbf{Y}}}$ dos valores ajustados da variável resposta, ao abrigo do Modelo. Com base no resultado obtido, mostre que a variância associada a um valor ajustado individual, $\hat{Y}_i$, está compreendida entre $\frac{\sigma^2}{n}$ e σ^2.

   (c) Considere a matriz de projecção ortogonal $\mathbf{H}$. Calcule o seu traço. Use esse resultado para mostrar que o efeito alavanca médio é dado por $\overline{h} = \frac{p+1}{n}$.

   (d) Mostre que a matriz $\mathbf{I}_n - \mathbf{H}$, onde $\mathbf{I}_n$ indica a matriz identidade de dimensão $n \times n$, é também uma matriz de projecção ortogonal. Diga, justificando, quais as quantidades importantes na regressão que estão contidas no subespaço de $\mathbb{R}^n$ sobre o qual esta matriz projecta.

III [6 valores]

1. Decidiu-se ajustar um Modelo Linear Generalizado que visa classificar, com base nas concentrações medidas, se uma observação provém, ou não, do local GSWS02. Após algum trabalho preliminar, optou-se por um Modelo Probit com 4 preditores, que produziu os seguintes resultados:

```
> summary(CFex1SITEprbt.glm)
```

```
Call:   glm(formula = I(dadosCF002$SITECODE == "GSWS02") ~ COND_MO + SI_MO + NA_MO + SO4S_MO,
    family = binomial(probit), data = dadosCF002)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -43.0437    10.7539  -4.003 6.26e-05 ***
COND_MO      -1.4847     0.3382  -4.390 1.13e-05 ***
SI_MO         8.0231     1.9046   4.213 2.52e-05 ***
NA_MO         7.3254     1.7378   4.215 2.49e-05 ***
SO4S_MO      56.6871    14.8145   3.826  0.00013 ***
---
    Null deviance: 193.966  on 259  degrees of freedom
Residual deviance:  34.797  on 255  degrees of freedom
AIC: 44.797
```

(a) Descreva em detalhe o modelo e comente brevemente a qualidade do seu ajustamento aos dados.

(b) Comente a seguinte afirmação: "*tendo em conta o valor* $b_4 = 56.6871$*, a variável* ***SO4S_MO*** *é a que mais contribui para a identificação das observações do local* ***GSWS02***".

(c) Os valores medianos dos quatro preditores usados no modelo, para a totalidade das 260 observações, são 39.74 (para `COND_MO`); 8.775 (para `SI_MO`); 2.47 (para `NA_MO`); e 0.12 (para `SO4S_MO`). Qual a probabilidade estimada pelo modelo para que uma observação com estes valores seja proveniente do local `GSWS02`? Essa probabilidade estimada aumenta ou diminui quando, mantendo-se tudo o resto igual, aumentar a conductância específica?

(d) Um modelo análogo, mas com a função de ligação canónica para variáveis resposta dicotómicas produziu um desvio residual de valor 35.30. Identifique esse modelo e diga, com base na informação disponível, por qual dos dois modelos optaria.

2. Com o objectivo de avaliar a variabilidade da concentração de cálcio entre anos e entre locais, foram usadas medições desta variável realizadas no mês de Janeiro, em 29 anos e 5 locais. Neste sub-conjunto de dados, existe apenas uma observação por combinação ano-local. Admita que, tanto `ano` como `local`, são factores de efeitos aleatórios.

(a) Descreva, em pormenor, o modelo que lhe parece ser adequado para o estudo acima descrito.

(b) No R, com a função `lmer` do pacote `lme4` executaram-se os seguintes comandos:

```
> dadoslmer1<-lmer(CA~1+(1|local)+(1|ano), data=dados)
> summary(dadoslmer1)
Linear mixed model fit by REML ['lmerMod']
Formula: CA ~ 1 + (1 | local) + (1 | ano)
 Random effects:
 Groups   Name        Variance Std.Dev.
 ano      (Intercept) 0.13669  0.3697
 local    (Intercept) 0.10618  0.3259
 Residual             0.07571  0.2752
Number of obs: 145, groups:  ano, 29; local, 5
Fixed effects:
            Estimate Std. Error t value
(Intercept)   3.4468     0.1627   21.18

> logLik(dadoslmer1)
'log Lik.' -60.72956

> dadoslmer2<-lmer(CA~1+(1|local), data=dados)
> logLik(dadoslmer2)
'log Lik.' -100.6637

> dadoslmer3<-lmer(CA~1+(1|ano), data=dados)
> logLik(dadoslmer3)
'log Lik.' -104.1064
```

i. O que conclui sobre a existência de variabilidade da concentração de cálcio entre anos e entre locais? Descreva detalhadamente um dos testes de hipóteses que efectuar.

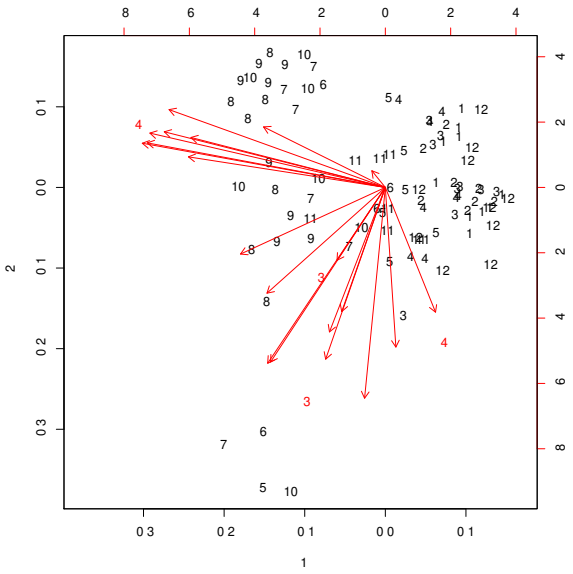
ii. Calcule o valor do Critério de Informação de Akaike (AIC) dos 3 modelos ajustados. Comente.

iii. Um investigagor defende que, dado o reduzido número de níveis do factor `local`, seria defensável admiti-lo como um factor de efeitos fixos. Calcule $\mathrm{Cov}[Y_{ij}, Y_{ij'}]$ para o modelo que admite o `local` como um factor de efeitos fixos e para o modelo que admite o `local` como um factor de efeitos aleatórios. Explique em que medida as covariâncias calculadas justificam a opção de se ter inicialmente admitido o `local` como um factor de efeitos aleatórios.

IV [3,5 valores]

1. Foi efectuada uma Análise em Componentes Principais das 92 observações de primeiro dos três grupos definidos no Grupo II (factor `GRP3`), sobre 20 variáveis numéricas *normalizadas* (sobretudo concentrações de elementos e compostos químicos). Obtiveram-se os seguintes resultados, em que no *biplot* as observações são representadas pelo número do mês a que cada observação corresponde.

```
> summary(prcomp(dadosnumCF002[dadosCF002$GRP3=="1",-c(1,3,5,14)], scale=TRUE))
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.6296 2.0228 1.7928 1.26189 1.18548 1.01075 0.75253
Proportion of Variance 0.3458 0.2046 0.1607 0.07962 0.07027 0.05108 0.02832
Cumulative Proportion  0.3458 0.5503 0.7110 0.79065 0.86092 0.91200 0.94032
```

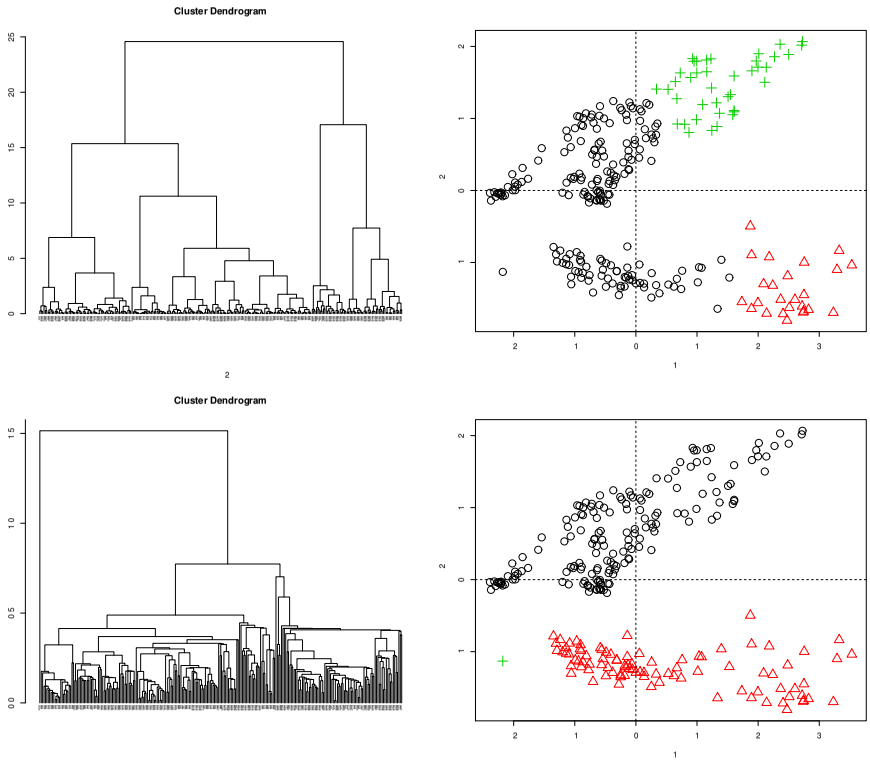


(a) Discuta as principais características dos dados salientadas no *biplot*. Em particular: discuta a qualidade do *biplot*; procure interpretar a primeira componente principal; identifique uma variável que esteja mal representada no *biplot*; identifique uma variável bem correlacionada com a segunda componente principal.

(b) Sabendo que a variável `CA_MO` tem um dos marcadores que mal se distingue, no canto superior esquerdo do *biplot*, identifique em que meses foram feitas as observações na metade superior da nuvem de pontos do ponto II.1.

2. Efectuaram-se duas classificações hierárquicas das concentrações normalizadas de potássio (K), sódio (NA) e cálcio (CA), usando os métodos da inércia mínima (*ward*) e do vizinho mais próximo (*single*),

ambos com a distância euclideana. Posteriormente efectuaram-se cortes nos respectivos dendrogramas de modo a obter 3 grupos em ambas as classificações. Os respectivos dendrogramas, bem como as projeções no plano definido pelas duas primeiras Componentes Principais dos indivíduos agrupados por classes (identificados por meio de símbolos distintos), encontram-se representados na figura abaixo. Sabe-se ainda que as percentagens de variância explicadas pelas duas primeiras Componentes Principais são respectivamente, 57% e 36%.



(a) Comente e explique os resultados obtidos à luz dos modelos de análise classificatória escolhidos. Qual das classificações lhe parece mais adequada?

As seguintes alíneas referem-se a resultados obtidos com método do vizinho mais próximo.

(b) Sabe-se que os 3 grupos contêm respectivamente 1, 168 e 91 indíviduos (seguindo a ordem definida pelo dendrograma). Indique o índice de RAND que obteria comparando esta partição em 3 grupos com a partição em 4 grupos dada pelo mesmo método.

(c) Sabendo que as distâncias entre os 4 grupos da partição referida na alínea anterior são dadas pela tabela abaixo (segundo a ordem definida pelo dendrograma), calcule as distâncias cofenéticas entre pares de indivíviduos representados por símbolos distintos no plano definido pelas duas primeiras Componentes Principais.

	C_2	C_3	C_4
C_1	1.620	1.515	1.662
C_2		1.060	0.773
C_3			0.702

3. Efectou-se uma classificação hierárquica de um conjunto finito de dados $X \subset \mathbb{R}^n$ com o método do vizinho mais afastado e a distância do máximo. Mostre que a distância cofenética $d_c(x, y)$ entre elementos $x = (x_1, \ldots, x_n)$ e $y = (y_1, \ldots, y_n)$ de X verifica $d_c(x, y) \geq |x_i - y_i|$, $\forall\, i = 1, \ldots, n$.