ME - 731 Métodos em Análise Multivariada
Prof. Caio Azevedo
Segundo semestre de 2009, Data: 10/12/2009
Prova III

Leia atentamente as instruções abaixo:

- **Tenha em mãos somente: lápis, borracha e caneta.**
- **Leia atentamente cada uma das questões.**
- **Enuncie, claramente, todos os resultados que você utilizar.**
- **Coloque seu nome e RA em cada uma das folhas que você recebeu, inclusive nesta.**
- **Entregue todas as folhas que você recebeu, inclusive os rascunhos e a prova propriamente, informando o que deve ser corrigido.**
- **Faça a prova, preferencialmente, à caneta, e procure ser organizado. Se fizer à lápis, destaque, à caneta, sua resposta.**
- **Não proceda de maneira indevida como: conversar durante a prova, utilizar-se de material que não permitido, emprestar material à colegas, sem autorização do professor e atender ao telefone celular (a não ser em casos de EXTREMA URGÊNCIA). Isso acarretará em nota 0 na prova.**
- **A prova terá duração de 2 horas, improrrogáveis, das 14h às 16h.**

Faça uma excelente prova!!

Questões

1. Sejam $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ dois vetores aleatórios independentes tais que $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Defina $Y_i = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i, i = 1, 2$ (na análise discriminante, esta seria a função discriminante sob normalidade e homocedasticidade). Responda os itens:

   a) Qual a distribuição de $Y_i, i = 1, 2$? Justifique, adequadamente, sua resposta (0,5 pontos).

   b) Qual a distribuição conjunta do vetor $(Y_1, Y_2)'$? As componentes são independentes? Justifique, adequadamente, seus desenvolvimentos (0,5 pontos).

   c) Prove que $P(Y_1 \leq \frac{\mathcal{E}(Y_1)+\mathcal{E}(Y_2)}{2}) = P(Z \leq -\frac{\Delta}{2})$, em que $\Delta^2 = Var(Y_1), Z \sim N_1(0, 1)$ e $E(Y_i)$ é o valor esperado de $Y_i, i = 1, 2$. Na análise discriminante, qual seria a utilidade da probabilidade que você calculou? Justifique, adequadamente, seus desenvolvimentos (1,0 ponto).

2. Seja a tabela de contingência abaixo, que representa a distribuição do número de espécies de animais consideradas extintas (a chamaremos de "espécie") por continente. Considere que o total de espécies (499) foi fixado previamente. Utilizou-se a metodologia de análise de correspondência para avaliar a dependência entre as variáveis: continente e espécie. Abaixo, também se encontram alguns resultados dessa análise.

Continente	Espécie de animais consideradas extintas						
	Moluscos	**Insetos**	**Peixes**	**Répteis**	**Aves**	**Mamíferos**	**Total**
Ásia	0	1	0	0	14	5	20
Europa	2	1	0	0	4	71	10
América	40	9	31	9	18	30	137
Oceania	81	47	1	2	51	23	205
África	68	3	0	12	39	5	127
Total	191	61	32	23	126	66	499

Tabela 1: Inércia e percentual da variância explicada pelas componentes da análise de correspondência

Inércia principal	**Percentual**	**Percentual acumulado**
0,2207	57,2	57,2
0,0980	25,4	82,6
0,0652	16,9	99,5
0,0020	0,5	100,0

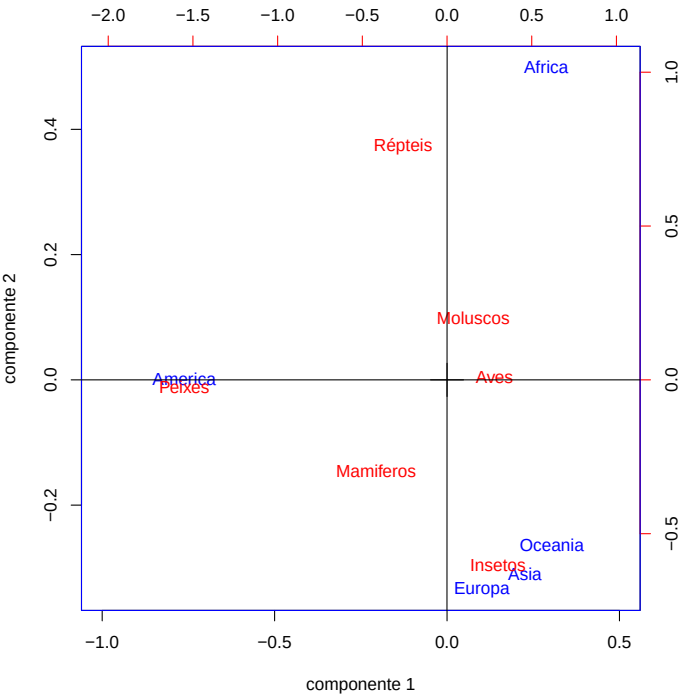

Figura 1: Bi-plot para a análise de correspondência

A estatística de qui-quadrado, para testrar H_0 : as variáveis continente e espécies de animais são indepedentes vs H_1 : tais variáveis são dependentes, forneceu p-valor $<$ 0,0001. Responda os itens:

a) O que você pode dizer sobre as variáveis continente e espécie, com relação a independência? Justifique, adequadamente, sua resposta, à um nível de significância $\alpha = 5\%$ (0,5 pontos).

b) Utilizar a técnica de análise de correspondência, para estudar o comportamento das variáveis (com relação à dependência) faz sentido neste caso? Justifique, adequadamente, sua resposta (0,5 pontos).

c) Com base na Tabela 1, o que você pode dizer sobre utilizar duas componentes para realizar a análise de correspondência? Justifique, adequadamente, sua resposta (1,0 ponto).

d) Através da Figura 1, analise, do modo mais amplo possível, o relacionamento entre as variáveis: espécie e continente. Utilize o máximo de informação possível (1,0 ponto).

3. A fronteira do Canadá com o Alasca (região que pertence aos Estados Unidos) é uma área de pesca de salmão. Sabidamente, os salmões nascem em água doce mas migram para o mar retornando, posteriormente, para o local onde nasceram, para fins de reprodução. Existem basicamente duas espécies de salmão nessa região: uma que nasce no Alasca (doravante grupo 1) e outra que nasce no Canadá (doravante grupo 2). Existe um acordo entre Estados Unidos e Canadá, de modo a proibir a pesca de um tipo de salmão por pescadores do outro país. Ou seja, pescadores canadenses devem pescar somente salmões provenientes do Canadá e o mesmo vale para pescadores do Alasca. Para se ter algum controle sobre isso, pretende-se criar uma regra de classificação utilizando duas variáveis medidas em 50 salmões provenientes do Alasca e em 50 salmões provenientes do Canadá. Tal regra visa poder identificar mais facilmente a origem de salmões pescados. As variáveis medidas são : X_1 = diâmetro das guelras durante a fase em água doce (em mm) e X_2 = diâmetro das guelras durante a fase no mar (em mm). Consideramos, para construir tal regra, distribuição normal bivariada para os dois grupos, homocedasticidade, custos iguais de classificação errada e probabilidades a priori de classificação iguais, com base numa amostra-treino de 25 salmões de cada grupo. Alguns resultados encontram-se abaixo:

$\overline{\boldsymbol{x}}_1 = (98,30; 429,66)'$, $\overline{\boldsymbol{x}}_2 = (137,38; 366,62)'$

$\widehat{y} = -0.15x_1 + 0.05x_2$ (função discriminante estimada) , $\widehat{m} = 2.72$.

Tabela 2: Classificação dos salmões da amostra complementar à amostra-treino

	Classificado	
Verdadeiro	**Alasca**	**Canadá**
Alasca	22	6
Canadá	1	21

Taxa aparente de erro = 14%.

Taxa ótima de erro = 7%.

Lembre-se de que a regra diz que se $\boldsymbol{x} = (x_1, x_2)'$ for tal que $\widehat{y} \geq \widehat{m}$ o salmão é considerado oriundo do Alasca e, caso contrário, é considerado oriundo do Canadá. Responda os itens:

a) Com base nas suposições feitas e nos valores dos vetores de médias, você diria que os grupos de salmões diferem entre si? Justifique, adequadamente, sua resposta (0,5 pontos).

b) Interprete, adequadamente, a função discriminante estimada (0,5 pontos).

c) Com base nos resultados, o que você diria sobre a regra de classificação obtida? Utilize o maior número possível de informações e justifique, adequadamente, seus comentários (1,0 ponto).

d) Se um salmão recém-pescado fosse tal que $\boldsymbol{x} = (90, 350)'$, de onde, provavelmente, ele teria vindo? Apresente todas os cálculos e jusitifique, adequadamente, sua resposta (1,0 ponto).

4. Considere a metodologia da análise de correlações canônicas. Quais são os principais objetivos, aspectos positivos, aspectos negativos e utilidades? Considere o máximo possível de informações em sua explanação (2,0 pontos).

Formulário

- Se $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ então $\mathcal{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ e $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$.