MAE5776 – Análise Multivariada LISTA 3 1º Semestre/2020

Considere o arquivo `congenitos.csv`, $Y_{35.129\times50}$ que corresponde a dados de expressão gênica ($\log_2$ transformados) de 35.129 fragmentos genéticos do cromossomo de ratos congênitos. No experimento realizado, 5 réplicas de 10 ratos foram selecionadas aleatoriamente das linhagens congênitas A, B, C, D e E, sendo 2 de cada. Então, dentro de cada réplica, metade foi aleatoriamente selecionada para receber uma dieta rica em sal e a outra metade recebeu dieta controle. A Tabela 1 ilustra o esquema do experimento realizado, o qual seguiu um Fatorial 2x5 com 5 réplicas, em que, em cada uma das 50 caselas (50=2x5x5), a expressão gênica dos 35.129 fragmentos foi avaliada. Note que se trata de uma matriz com *n*=50 unidades amostrais estruturadas de acordo com o esquema fatorial 2x5 (dois fatores: Dieta (em dois níveis, Sal e Controle) e Linhagem de rato (em cinco níveis, A, B, C, D e E)) e 5 réplicas, com *p*=35.129 variáveis sendo avaliadas. Além disso, a Tabela 1 indica uma estrutura de blocagem (nas réplicas) em que, materiais genéticos de cada grupo de 10 animais (um de cada dos 10 tratamentos), foram aleatoriamente colocados juntos em *plates* para a leitura das expressões.

Tabela 1. Esquema de delineamento Fatorial 2x5 aleatorizado em blocos (réplicas)

	Sal					**Controle**				
Réplica	**A**	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
R1										
R2										
R3										
R4										
R5										

1. Escolha as Top 10 variáveis dentre as *p*=35.129. Para tanto, defina algum critério, por exemplo, seleção aleatória, ou, de forma mais interessante, você pode realizar *p* ANOVAs e selecionar os genes com expressão mais significante. Note, que na ANOVA há 9 graus de liberdade para estudar o efeito de tratamento (efeito principal de Sal=1 g.l.; efeito principal de linhagem=4 g.l; efeito de interação=4 g.l), além de 4 graus de liberdade para bloco e o restante para o termo de erro.

Também, se quiser realizar uma solução mais sofisticada ainda, e ganhar até 1 ponto a mais nesta lista, represente os 35.129 p-valores da ANOVA em um Gráfico Vulcão. O que é isso? É um gráfico de dispersão com $-\log_{10}$(p-valor) na ordenada e uma medida de tamanho de efeito (*change-fold*) na abscissa, em geral, representada pela estimativa de um efeito ($\hat{\beta}$). Pesquise sobre esse gráfico (*Vulcano plot*) e implemente.

2. Com base na seleção dos Top 10 genes, será obtida uma matriz de trabalho $Y_{50\times10}$ (*n*=50, *p*=10). Com base na sua matriz de trabalho represente os dados por meio de dois eixos principais obtidos de:
a) Análise de Componentes Principais.
b) Escalonamento Multidimensional
c) Análise de Fatores (comuns e específicos)

d) Análise Discriminante: neste caso você pode adotar 2 grupos (Sal e Controle), 5 grupos (A, B, C, D e E), ou 10 grupos (do esquema fatorial 2x5). Faça sua escolha e justifique.

Em cada caso apresente as suposições envolvidas e interprete o padrão de variação mostrado nos gráficos.

3. Considere a análise discriminante linear de Fisher de sua matriz de trabalho, $Y_{50\times10}$, adotando G=2 grupos (Dieta de Sal e Controle). Além disso, considere também a análise discriminante via regressão logística (binária). Em cada caso obtenha a matriz de confusão via o método empírico e validação cruzada (*leave-one-out*). Compare os dois métodos de classificação dos grupos relativamente às suposições envolvidas e métricas de predição.

4. Com base na sua matriz de trabalho, $Y_{50\times10}$, obtenha a tabela de MANOVA. Há efeito significante de algum fator sob estudo? De acordo com a MANOVA é possível obter uma decomposição da matriz de soma de quadrados e produto cruzados total (SQPC total). De maneira correspondente, obtenha uma decomposição da matriz de trabalho $Y_{50\times10}$. Qual é a utilidade desse tipo de decomposição na análise de dados? Como exemplo, pense nas análises realizadas no item 2, as quais usaram a matriz de trabalho mas poderiam adotar componentes dos dados.

5. A partir de uma matriz de dados normalizados $Y^*_{n\times p}$, considere a matriz de covariâncias $nS_{p\times p} = Y^{*\prime}Y^* = V\Lambda V'$, tal que $V_{p\times p} = \left(V_1,...,V_p\right)$ e $\Lambda = diag\left(\lambda_j\right)$ são matrizes de autovetores (das colunas de $Y^*_{n\times p}$) e autovalores, respectivamente, e a matriz de distâncias $D_{n\times n}$, tal que seus elementos são função dos elementos de $B_{n\times n} = Y^*Y^{*\prime} = U\Lambda U'$, com $U_{n\times n} = \left(U_1,...,U_n\right)$ matriz de autovetores (das linhas de $Y^*_{n\times p}$). Três pesquisadores realizaram análises estatísticas e chegaram à seguinte conclusão:

Pesquisador 1: $Y^*_{n\times p} \cong Y^*\begin{pmatrix} V_1 & V_2 \end{pmatrix}$, $\Sigma \cong V_1V_1'\lambda_1 + V_2V_2'\lambda_2$

Pesquisador 2: $Y^*_{n\times p} \cong Y^*\begin{pmatrix} \frac{V_1}{\sqrt{\lambda_1}} & \frac{V_2}{\sqrt{\lambda_2}} \end{pmatrix}$

$$\Sigma \cong \Phi\Phi' + \Psi; \quad \Phi = \left(V_1\sqrt{\lambda_1} \quad V_2\sqrt{\lambda_2}\right), \quad \Psi = diag\left(1-(\phi_1^2+\phi_2^2)\right)$$

Pesquisador 3: $Y^*_{n\times p} \cong \left(U_1\sqrt{\lambda_1} \quad U_2\sqrt{\lambda_2}\right)$, $B \cong U_1U_1'\lambda_1 + U_2U_2'\lambda_2$

Qual análise estatística cada pesquisador realizou? Eles partiram do mesmo objetivo?
Se desejar ganhar até 2 pontos extras nesta lista, simule dados e realize as análises dos três pesquisadores.

Boa sorte ☺