DCC639: Álgebra Linear Computacional **(Prazo para submissão: 01/03/21 23:59)**

Lista de Exercícios 04

Professores: Erickson e Fabricio

Política da Disciplina: Leia todas as instruções abaixo cuidadosamente antes de começar a resolver a lista, e antes de fazer a submissão.

- As questões podem ser discutidas entre até três alunos (conjuntos disjuntos). Os nomes dos colegas precisam ser incluídos na submissão. Contudo, a escrita das soluções e submissão deve ser feita individualmente.
- A submissão deve ser feita em formato PDF através do Moodle, mesmo que tenham sido resolvidas a mão e escaneadas.
- Todas as soluções devem ser justificadas.
- Todas as fontes de material precisam ser citadas. O código de conduta da UFMG será seguido à risca.

Problema 1: Considere o conjunto de dados abaixo com as notas de 5 estudantes em 4 disciplinas. Calcule a matriz de covariância.
(Dica: Para saber se devemos usar $\frac{X^\top X}{n-1}$ ou $\frac{XX^\top}{n-1}$, lembre-se de que a matriz final deve ter ordem igual ao número de atributos.)

Estudante	GAAL	PDS1	Cálculo 1	ALC
1	90	80	60	95
2	65	75	90	70
3	40	90	60	55
4	80	60	59	75
5	60	100	80	80

(Solução)
Primeiro, devemos achar a matriz centralizada X_c. Para isso, achar o vetor das médias das colunas de X, obtendo $u_i = \begin{bmatrix}67 & 81 & 69.8 & 75\end{bmatrix}$. Em seguida, calcular X_c subtraindo u_i de cada linha da matriz X, obtendo:

$$X_c = \begin{bmatrix} 23 & -1 & -9.8 & 20 \\ -2 & -6 & 20.2 & -5 \\ -27 & 9 & -9.8 & -20 \\ 13 & -21 & -10.8 & 0 \\ -7 & 19 & 10.2 & 5 \end{bmatrix}$$

A matriz de covariância de X é dada por $C_x = \frac{X_c^T X_c}{n-1}$. Fazendo-se esse cálculo, obtemos:

$$C_x = \begin{bmatrix} 370 & -165 & -53.25 & 243.75 \\ -165 & 230 & 55.25 & -18.75 \\ -53.25 & 55.25 & 205.2 & -12.5 \\ 243.75 & -18.75 & -12.5 & 212.5 \end{bmatrix}$$

Problema 2: Seja X_c um conjunto de dados centralizado onde as linhas representam as observações. Sabendo que Cov(X) é a matriz de covariância de X, mostre, algebricamente, como o PCA de X_c pode ser obtido a partir de seu SVD.

(Solução)

Da decomposição SVD tem-se que $X_c = U\Sigma V^T$, e da matriz de covariância usada no PCA, tem-se que $\text{Cov}(X_c) = \frac{X_c^\top X_c}{n-1}$. Dado que a matriz de covariância $\text{Cov}(X_c)$ é simétrica, ela admite a decomposição espectral $\text{Cov}(X_c) = PDP^T$, onde P é uma matriz ortogonal com os autovetores de $\text{Cov}(X_c)$ e D uma matriz diagonal com os autovalores de $\text{Cov}(X_c)$. Dessa forma, fazendo-se as devidas substituições, e sabendo-se que $U^TU = I$ e $\Sigma^T = \Sigma$ temos:

$$\begin{aligned}
\text{Cov}(X_c) &= \frac{X^TX}{n-1} \\
&= PDP^T \\
&= \frac{(U\Sigma V^T)^T U\Sigma V^T}{n-1} \\
PDP^T &= \frac{V\Sigma^T U^T U\Sigma V^T}{n-1} \\
&= V\Sigma^2 V^T \frac{1}{n-1}
\end{aligned}$$

Como $\frac{1}{n-1}$ é um escalar, podemos colocar a multiplicação apenas em Σ^2, obtendo:

$$PDP^T = V\frac{\Sigma^2}{n-1}V^T$$

Assim, estabelecemos uma relação entre o SVD e o PCA de X_c, onde $P = V$ e cada autovalor em D é equivalente a cada autovalor em Σ elevado ao quadrado e dividido por $n-1$.

Problema 3: Considere os pontos a seguir:

x	2.0	3.5	4.0	5.1	7.0
y	2.2	2.0	3.0	6.0	5.0

Usando o método dos quadrados mínimos, encontre os parâmetros da regressão linear simples $f(x) = \beta_0 + \beta_1 x$. Atenção: você não pode resolver esta questão usando uma função de biblioteca que retorne os coeficientes da regressão diretamente a partir de x e y.

(Solução)
Resolver o sistema $A\beta = C$, onde:

$$A = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},\ e\ C = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

$$A\beta = C \rightarrow A^{-1}A\beta = A^{-1}C \rightarrow \beta = A^{-1}C$$

$$\beta_0 = 0.426\ e\ \beta_1 = 0.744$$

Problema 4: Considere a tabela de pontos a seguir:

x	1	2	3	4	5	6
y	-4.501	83.453	112.953	123.824	170.335	183.008

Suponha que a relação entre x e y seja dada por $y = \beta_1 x + \beta_2 \ln x + \epsilon$. Obtenha os valores de β_1 e β_2 através do método dos quadrados mínimos. (Dica: a função y pode ser vista como uma regressão linear múltipla em x, onde $x_1 = x$ e $x_2 = \ln x$.)

(Solução)
Solução 1: Assumindo-se que o erro ϵ é aproximadamente zero, e a partir da dica, constrói-se uma nova tabela onde $x_1 = x$ e $x_2 = lnx$. Nesse caso, a regressão linear múltipla pode ser obtida através da solução do sistema

$$\begin{bmatrix} x_1 \cdot x_1 & x_1 \cdot x_2 \\ x_2 \cdot x_1 & x_2 \cdot x_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} y \cdot x_1 \\ y \cdot x_2 \end{bmatrix}$$

Onde $x_i \cdot x_j$ denota o produto interno dos vetores x_i e x_j. Fazendo-se os cálculos teremos o seguinte sistema:

$$\begin{bmatrix} 91.000 & 29.025 \\ 29.025 & 9.410 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 2946.28 \\ 955.64 \end{bmatrix}$$

Resolvendo-se o sistema usando a matriz inversa de X, obtem-se:

$$\beta_1 = -0.964 \; e \; \beta_2 = 104.531$$

Solução 2: Assumindo-se ϵ como também uma variável do sistema, podemos resolvê-lo da seguinte forma:

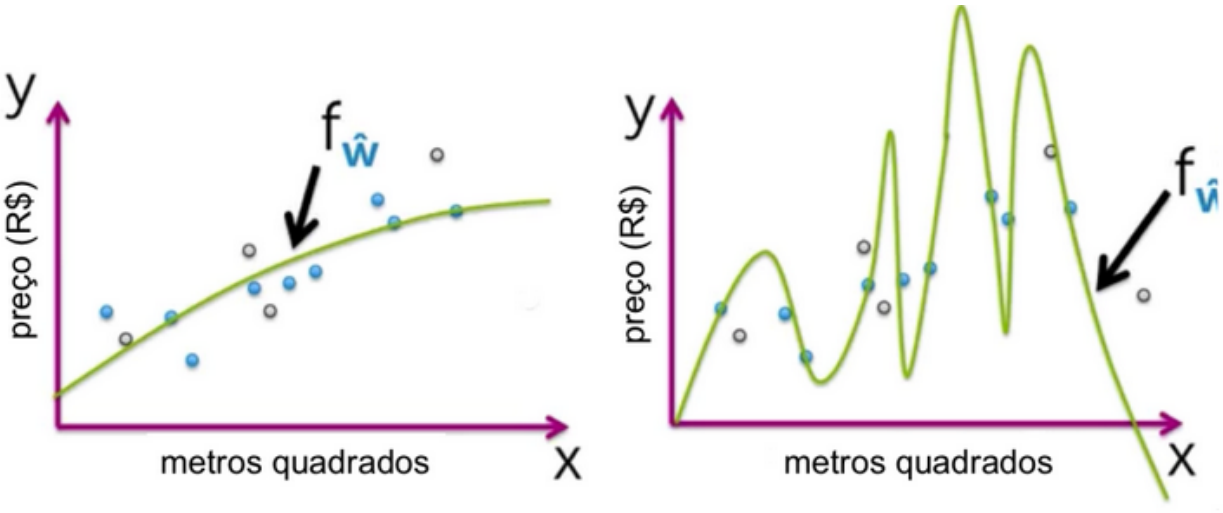
$$\begin{bmatrix} 1 & ln(1) & 1 \\ 2 & ln(2) & 1 \\ 3 & ln(3) & 1 \\ 4 & ln(4) & 1 \\ 5 & ln(5) & 1 \\ 6 & ln(6) & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \epsilon \end{bmatrix} = \begin{bmatrix} -4.501 \\ 83.453 \\ 112.953 \\ 123.824 \\ 170.335 \\ 183.008 \end{bmatrix}$$

Resolvendo-se usando a pseudo-inversa de X, $\beta = (X^T X)^{-1} X^T Y$, obtem-se:

$$\beta_1 = -1.680, \; \beta_2 = 105.696 \; e \; \epsilon = 1.492$$

Problema 5: Deseja-se usar a regressão polinomial $f(x_i) = w_0 + w_1 x_i + w_2 x_i^2 + ... + w_p x_i^p$ para estimar a relação entre a metragem (em m^2) de um imóvel e o seu preço em um bairro de Belo Horizonte. As figuras abaixo ilustram (não são uma representação exata) os resultados obtidos para $p = 3$ e $p = 8$, respectivamente, a partir dos pontos em azul.
Responda:

- Qual das regressões possui o menor desvio considerando-se apenas os pontos em azul?
- Qual das regressões possui o menor desvio considerando-se os pontos azuis e os pontos brancos?
- Com base nesta última resposta, qual dos valores de p é mais adequado e por quê?



(Solução)

- Considerando-se apenas os pontos em azul, a regressão com p = 8 possui menor desvio.
- Considerando-se todos os pontos, a regressão com p = 3 possui o menor desvio.

- O valor mais adequado é p = 3, pois é uma regressão que melhor representa todas as amostras do gráfico, visto que o preço cresce monotonicamente com a metragem (o que corresponde à intuição o sobre a relação entre essas variáveis), além de ser uma representação mais simples.

Problema 6: Assinale V para verdadeiro ou F para falso e justifique:
( ) Dado um conjunto de dados centralizado X, para obter uma representação de X em k dimensões via PCA, podemos utilizar o SVD truncado de X de posto k.
( ) A direção da PC_1 de um conjunto de dados bidimensional, em que cada ponto i tem coordenadas (x_i, y_i), é a mesma da reta $f(x)$ obtida pela regressão linear dos dados.

(Solução)
(V) Como visto na questão 2, podemos obter as componentes principais de X utilizando os vetores singulares direitos (V) de X obtidos com seu SVD. Se queremos representar X em k dimensões, precisamos dos k primeiros vetores singulares direitos, correspondentes aos k maiores valores singulares de X. Assim, precisamos apenas do SVD truncado de posto k.
(F) A reta obtida pela regressão linear minimizaria o erro médio quadrático (entre y e ŷ), enquanto o PCA minimizaria a distância ortogonal entre os dados e a reta (direção da PC_1).

Problema 8: Considerando o seguinte conjunto de dados que descreve preços de passagens aéreas de uma cidade O até cada uma das cidades (A a E), assim como a distância entre elas:

Destino	Distância (km)	Preço (R$)
A	572	177
B	371	138
C	612	192
D	409	158
E	946	260

(a) Utilizando o método de equações normais visto em aula, calcule a regressão linear para o conjunto de dados. (Considere o preço como a variável dependente e deixe explícito os passos intermediários para o método)
(b) De acordo com sua regressão, qual seria o preço de uma passagem aérea para a cidade F, sendo que sua distância para O é de 800km?

(Solução)
(a) De acordo com o enunciado, definimos nossas variáveis da seguinte forma:

$$X = \textbf{Distância} \quad \textbf{e} \quad Y = \textbf{Preço}$$

Calculando a média de cada atributo, temos:

$$\mu_X = 582 \quad \textbf{e} \quad \mu_Y = 185$$

Para o método, precisamos obter os elementos de $X - \mu_X$ e $Y - \mu_Y$:

$$X - \mu_X = [-10 \ -211 \ 30 \ -173 \ 364]$$

$$Y - \mu_Y = [-8 \ -47 \ 7 \ -27 \ 75]$$

Agora, vamos determinar os valores do numerador e denominador e encontrar β_1:

$$\sum_i (X_i - \mu_X)(Y_i - \mu_Y) = sum\{(X - \mu_X) \odot (Y - \mu_Y)\} = sum\{[80 \ 9917 \ 210 \ 4671 \ 27300]\}$$

$$\sum_i (X_i - \mu_X)^2 = sum\{(X - \mu_X) \odot (X - \mu_X)\} = sum\{[100 \ 44521 \ 900 \ 29929 \ 132496]\}$$

$$\frac{\sum_i (X_i - \mu_X)(Y_i - \mu_Y)}{\sum_i (X_i - \mu_X)^2} = \frac{42178}{207946} = 0.203$$

Por fim, temos $\beta_0 = \mu_Y - \beta_1\mu_X$:

$$\beta_0 = 185 - 0.203 * 582 = 66.95$$

(b) Calculamos o valor de Y para o valor desejado usando os parêmetros da reta obtidos pela regressão:

$$Y_F = \beta_0 + \beta_1 * 800$$

$$Y_F = 66.95 + 0.203 * 800 = R\$229.21$$

Problema 9: Suponha que temos um conjunto de dados X de dimensões $m \times n$ em que as linhas possuem variáveis, e as colunas possuem observações, e que a matriz já está *z-normalizada*, e portanto suas linhas possuem média 0 e desvio padrão 1. Suponha que executamos o PCA em X, e fazemos uma projeção dos dados de X nos 10 primeiros componentes principais, e assim obtemos uma matriz Y com os dados de dimensão reduzida $10 \times n$. Explique porque não é uma boa estratégia fazer uma nova execução do PCA, dessa vez em Y e projetar os dados nos 5 primeiros componentes principais de Y, com o objetivo de reduzir ainda mais a dimensionalidade. Explique as diferenças dessa estratégia do que simplesmente apenas projetar X nos seus primeiros 5 componentes principais.

(Solução)
O PCA envolve uma diagonalização da matriz $XX^\top$, e a geração de uma matriz de covariância diagonal, em que variáveis diferentes possuem covariância zero. Por esse motivo, quando calculamos a matriz de covariância de Y, ela ja é uma matriz diagonal, e portanto a sua diagonalização já é a própria matriz:

$$D = IDI^\top$$

Seus autovetores serão os vetores canônicos, já que os dados em Y já estão projetados nos vetores de maior variância dos dados originais. Portanto, ao projetar os dados novamente com esses componentes principais (autovetores), consistirá em apenas multiplicar os dados por uma matriz identidade e portanto os mesmos dados serão obtidos.

Portanto, essa estratégia adotada é a mesma do que simplesmente apenas projetar os dados originais X nos seus primeiros 5 componentes principais diretamente.