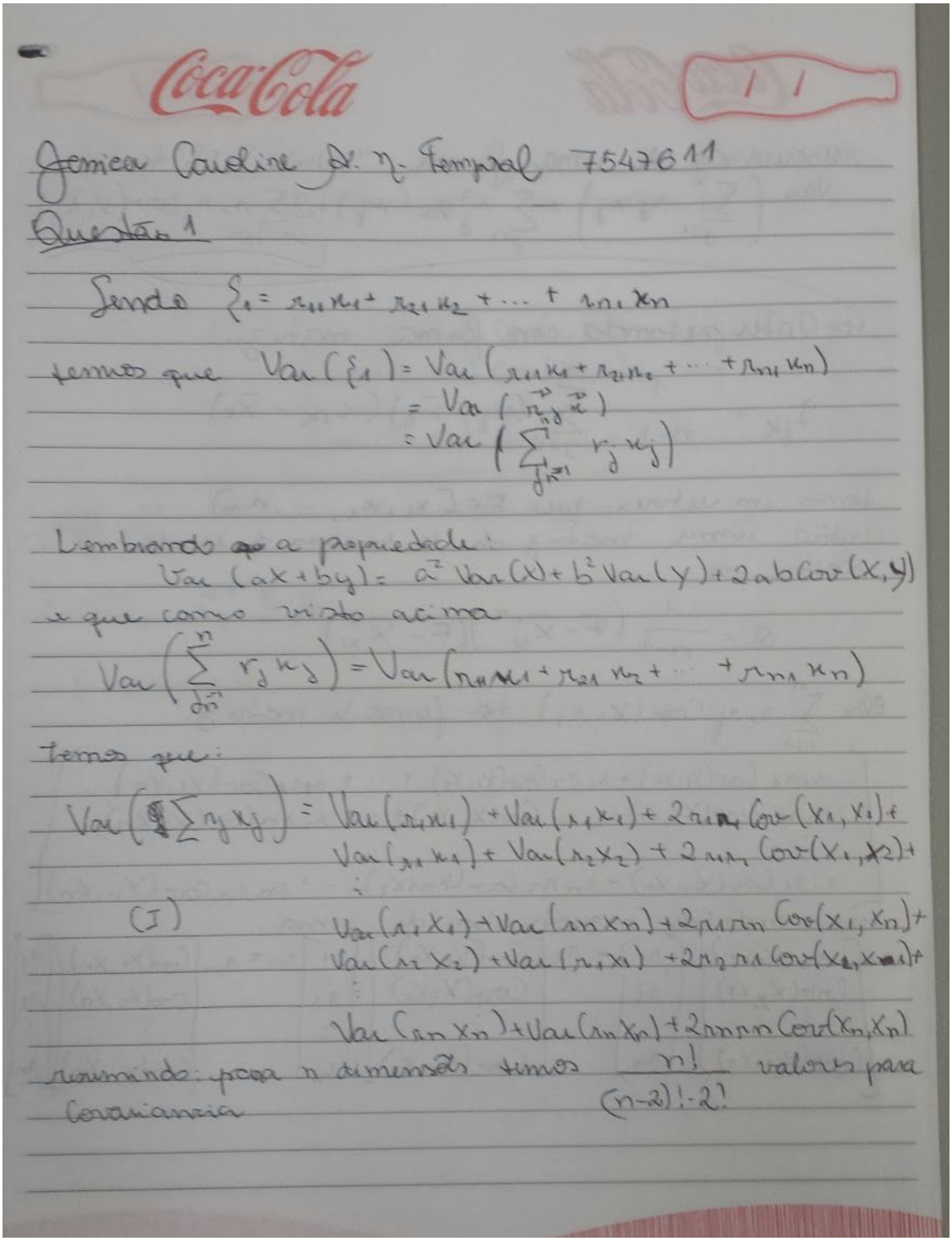
Jessica Caroline Alves Nunes Temporal 7547611
Questão 1



Jessica Caroline A. N. Temporal 7547611

Questão 1

Sendo $\xi_i = r_{i1}x_1 + r_{i2}x_2 + \dots + r_{in}x_n$

temos que $Var(\xi_i) = Var(r_{i1}x_1 + r_{i2}x_2 + \dots + r_{in}x_n)$

$= Var(\vec{r_{ij}}\vec{x})$

$= Var\left(\sum_{j=1}^{n} r_j x_j\right)$

Lembrando a propriedade:

$Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\,Cov(X,Y)$

e que como visto acima

$$Var\left(\sum_{j=1}^{n} r_j x_j\right) = Var(r_{i1}x_1 + r_{i2}x_2 + \dots + r_{in}x_n)$$

Temos que:

$$Var\left(\sum r_j x_j\right) = Var(r_1x_1) + Var(r_1x_1) + 2r_1r_1\,Cov(x_1,x_1) +$$
$$Var(r_2x_2) + Var(r_2x_2) + 2r_2r_2\,Cov(x_1,x_2) +$$
$$\vdots$$
$$Var(r_1x_1) + Var(r_nx_n) + 2r_1r_n\,Cov(x_1,x_n) +$$
$$Var(r_2x_2) + Var(r_1x_1) + 2r_2r_1\,Cov(x_2,x_1) +$$
$$\vdots$$
$$Var(r_nx_n) + Var(r_nx_n) + 2r_nr_n\,Cov(x_n,x_n) \quad (I)$$

resumindo: para n dimensões temos $\frac{n!}{(n-2)!\cdot 2!}$ valores para covariancia

reescrevendo de uma forma "menor" (I) temos:

$$\text{Var}\left(\sum_{j=1}^{n} x_j x_j\right) = \sum_{j=1}^{n} x_j^2 \text{Var}(x_j) + 2\sum_{i<j\leq n} x_i x_j \text{Cov}(X_i, X_j)$$

↳ Interpretando em forma matriz:

$$q_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

temos em vetores que $F = [x_1, x_2, \cdots, x_n]$ então uma matriz de k linhas e N colunas é dada por:

$$Q = \frac{1}{n-1}\left(F - \bar{X}_N^{1^T}\right)\left(F - \bar{X}_N^{1}\right)^T$$

Olha $\sum_{i,j=1}^{n} x_i x_j \text{Cov}(X_i, X_j)$ na forma de matriz:

$$\begin{bmatrix} x_1x_1\text{Cov}(X_1,X_1) + x_1x_2\text{Cov}(X_1,X_2) + \cdots + x_nx_n\text{Cov}(X_1,X_n) \\ x_2x_1\text{Cov}(X_2,X_1) + x_2x_2\text{Cov}(X_2,X_2) + \cdots + x_2x_n\text{Cov}(X_2,X_n) \\ \vdots \\ x_2x_1\text{Cov}(X_n,X_1) + x_nx_2\text{Cov}(X_n,X_2) + \cdots + x_nx_n\text{Cov}(X_n,X_n) \end{bmatrix}$$

reescrevendo isso mostrinho de uma:

$$x_1\begin{bmatrix}\text{Cov}(X_1,X_1)\\ \text{Cov}(X_2,X_1)\\ \vdots \\ \text{Cov}(X_n,X_1)\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix} + x_2\begin{bmatrix}\text{Cov}(X_1,X_2)\\ \text{Cov}(X_2,X_2)\\ \vdots\\ \text{Cov}(X_n,X_2)\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix} + \cdots + x_n\begin{bmatrix}\text{Cov}(X_1,X_n)\\ \text{Cov}(X_2,X_n)\\ \vdots\\ \text{Cov}(X_n,X_n)\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix}$$

que continuando com a manipulação matricial temos:

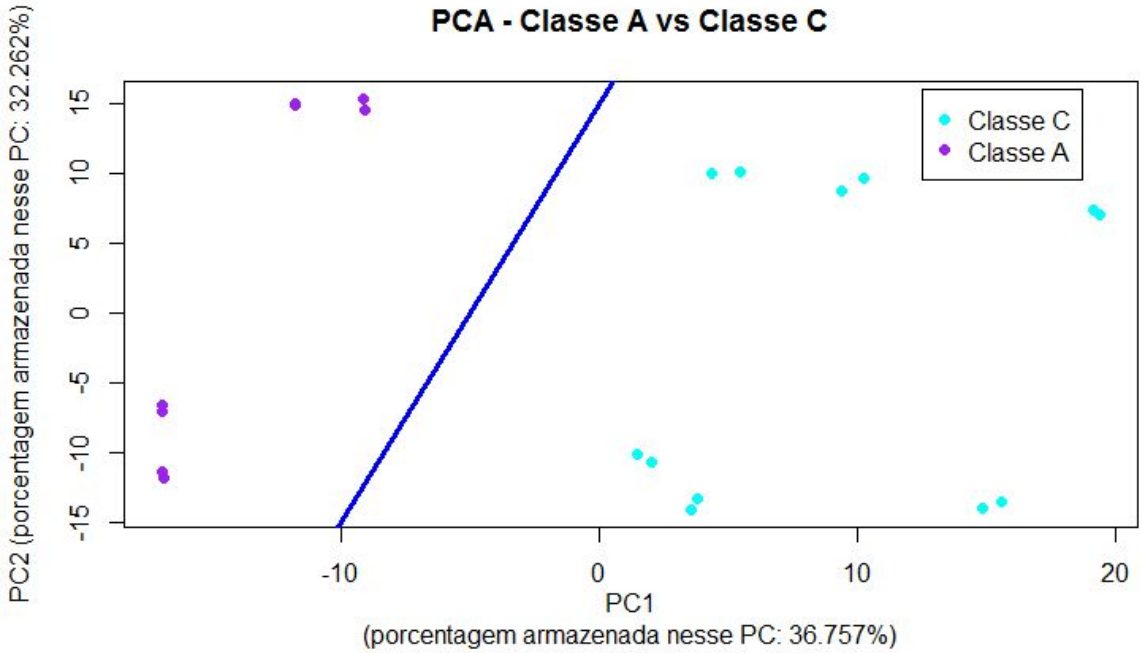
$$[r_1\ r_2 \cdots r_n] \cdot \begin{bmatrix} cov(X_1,X_1) & cov(X_1,X_2) & \cdots & cov(X_1,X_n) \\ cov(X_2,X_1) & cov(X_2,X_2) & \cdots & cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \\ cov(X_n,X_1) & cov(X_n,X_2) & & cov(X_n,X_n) \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}$$

$\vec{r}_j^T$ — mat de cov — $\vec{r}_j$

Então, $\vec{r}_j^T \sum \vec{r}_j$.

finalmente, $Var(\xi_1) = \vec{r}_j^T \sum \vec{r}_j$

Questão 2



De acordo com o gráfico mostrado pelo código no aquivo zip e considerando a linha hipotética definida no gráfico em azul, as amostras tem agrupamento claro que pode ser visualizado e o plot possibilita essa visualização, se olharmos pra linha desenhada em azul e supondo que ela fosse a melhor fronteira de decisão parar classificar uma nova amostra, fica claro que é possível usar os dois primeiros componentes principais para classificar as amostras nas classes descritas. As porcentagens das variâncias "armazenadas" no PC1 e no PC2 são 36.757% e 32.262% respectivamente.

Questão 3

O plot do PCA tridimensional pode ser uma enrascada tendo vista que o dependendo do referencial usado o gráfico não será informativo, por exemplo no caso da figura 4, é difícil observar as reais distâncias de uma amostra ou grupo de amostras para a outra amostra ou grupo de amostras. Além disso a imagem não mostra uma clara aproximação da as amostras de RNA amplificado de urina com as amostras tecido-especificas.
Uma questão é, que se o câncer analisado é de próstata, as amostras de urina vindas de indivíduos masculinos (provavelmente doentes) deveriam se aproximar de algum cubo azul (amostra característica) mostrando a similaridade, enquanto as amostras de urina de indivíduos femininos deveriam estar o mais longe possível dos cubos azuis mostrando que, de fato mulheres não poderiam ter câncer de próstata. Porem não fica claro essa diferenciação. Do ponto de vista de reconhecimento de padrões podemos ressaltar que não existe um plano que indique a separação dos clusters mencionados na descrição da figura.