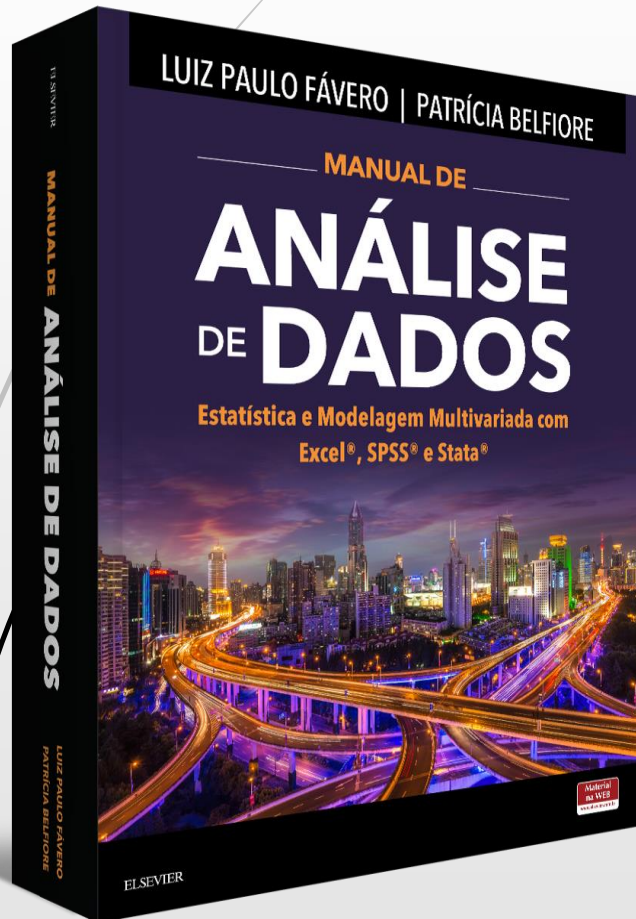


**MBA
USP
ESALQ**

***Unsupervised Machine
Learning: Análise Fatorial e
PCA I***

Rafael de Freitas Souza

Manual de Análise de Dados



https://www.grupogen.com.br/manual-de-analise-de-dados?gclid=CjwKCAjww-CGBhALEiwAQzWxOjpfZAJ9tTqzmcVUDQ5L8Wi-agPnlqOvgDFho0Xc_CWAooPSI1NMxBoCM-oQAvD_BwE

30% de desconto para os alunos do MBA USP/ESALQ

Cupom: E-FAVERO30

Validade: 10/07/2021



Conceitos Iniciais



A Análise Fatorial por Componentes Principais

- A Análise Fatorial por Componentes Principais (Principal Components Analysis – PCA), configura-se em técnica exploratória que lida com variáveis métricas que possuem, entre si, consideráveis valores de correlação, a fim de se estabelecer nova(s) variável(is) que capture(m) o comportamento conjunto das variáveis originais.
- A essas novas variáveis, dar-se-á o nome de fator.
- Portanto, pode-se entender a PCA como técnica que visa o agrupamento de variáveis a partir de critérios previamente estabelecidos.

Análise Fatorial por Componentes Principais

X

Análise de Clusters

	Clusterização	PCA
Variáveis Envolvidas	Métricas, em regra; há a possibilidade de se trabalhar com variáveis binárias.	Métricas.
Objetivo Principal	O agrupamento de observações em razão das diferenças ou semelhanças entre os indivíduos de uma base de dados, em razão de critérios previamente estabelecidos, a fim do estabelecimento de grupos de observações que sejam homogêneos internamente e heterogêneos entre si.	O agrupamento de variáveis, em razão de critérios previamente estabelecidos, levando-se em consideração suas correlações, evidenciando-se novas variáveis (fatores) que não eram observáveis antes da aplicação do método e que compartilhem a carga correlacional das variáveis anteriores e que sejam ortogonais entre si.
Output Principal	Agrupamentos, que são variáveis categóricas.	Fatores, que são variáveis métricas.



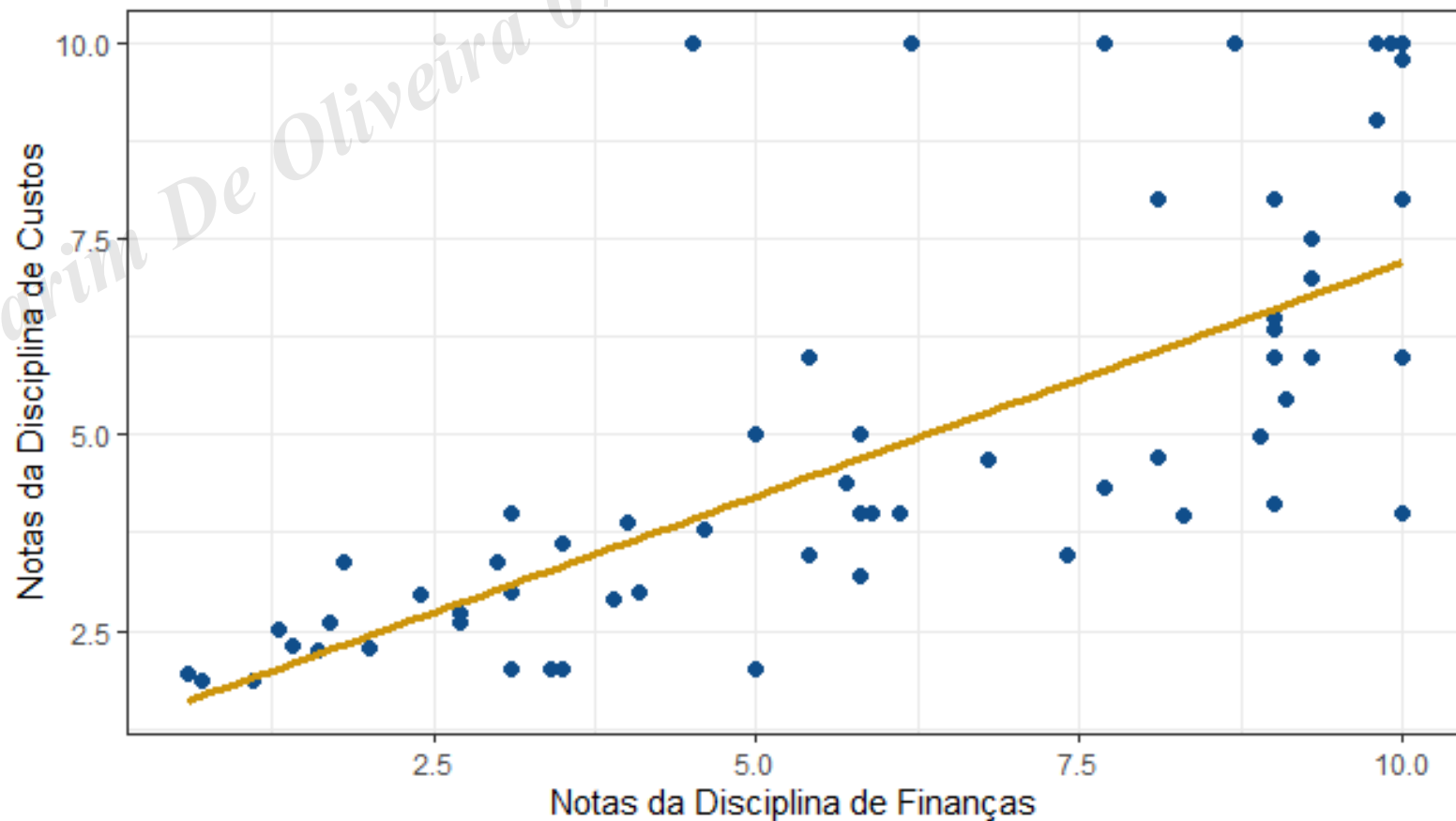
Outros Objetivos Relevantes da PCA

- A redução estrutural da dimensionalidade da base de dados;
- A evidenciação de variáveis (fatores) ortogonais entre si, isto é, que não guardem correlações entre si.
- A validação de constructos;
- A elaboração de rankings sem a utilização de ponderação arbitrária.

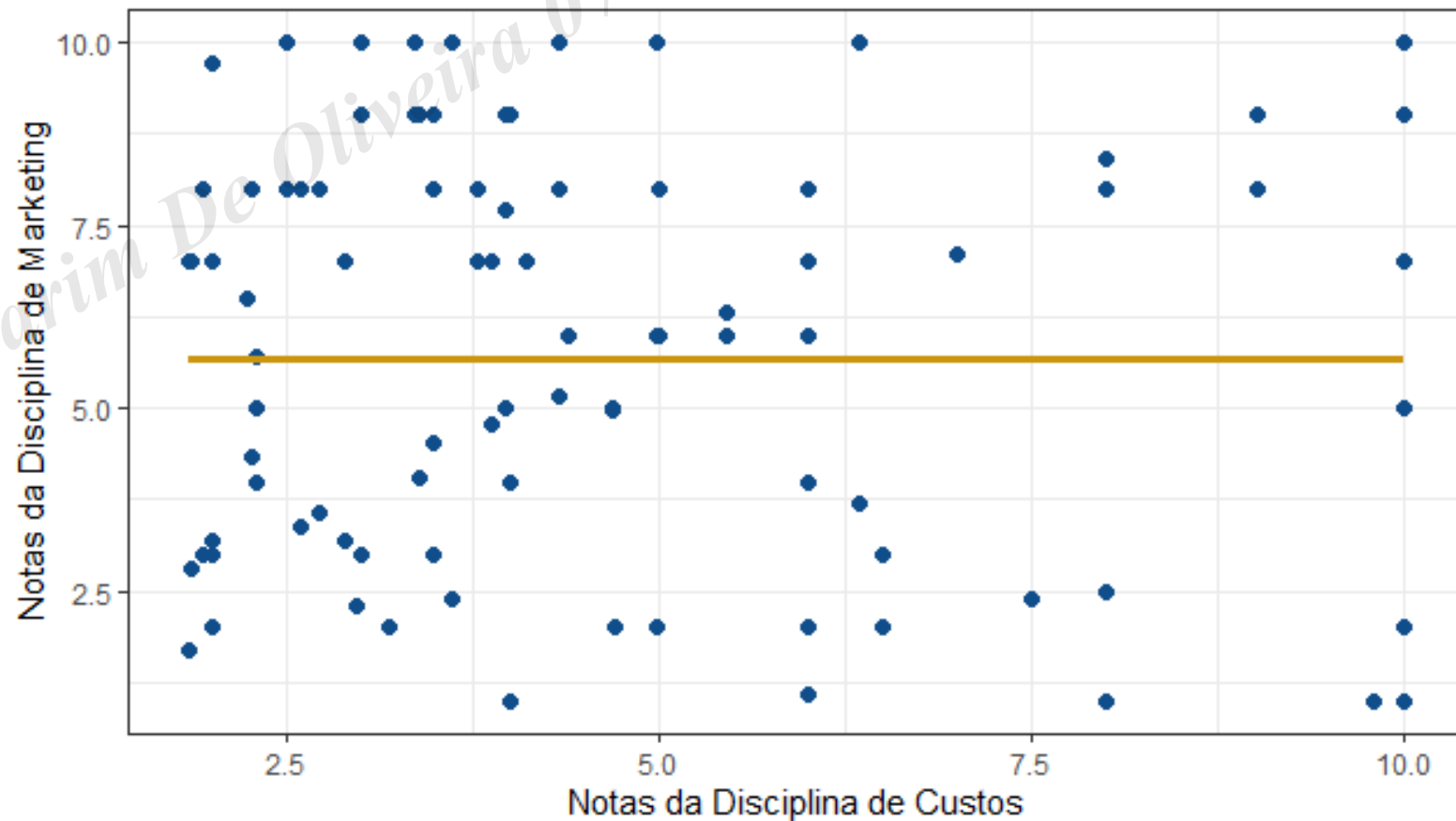
Redução Dimensional dos Dados



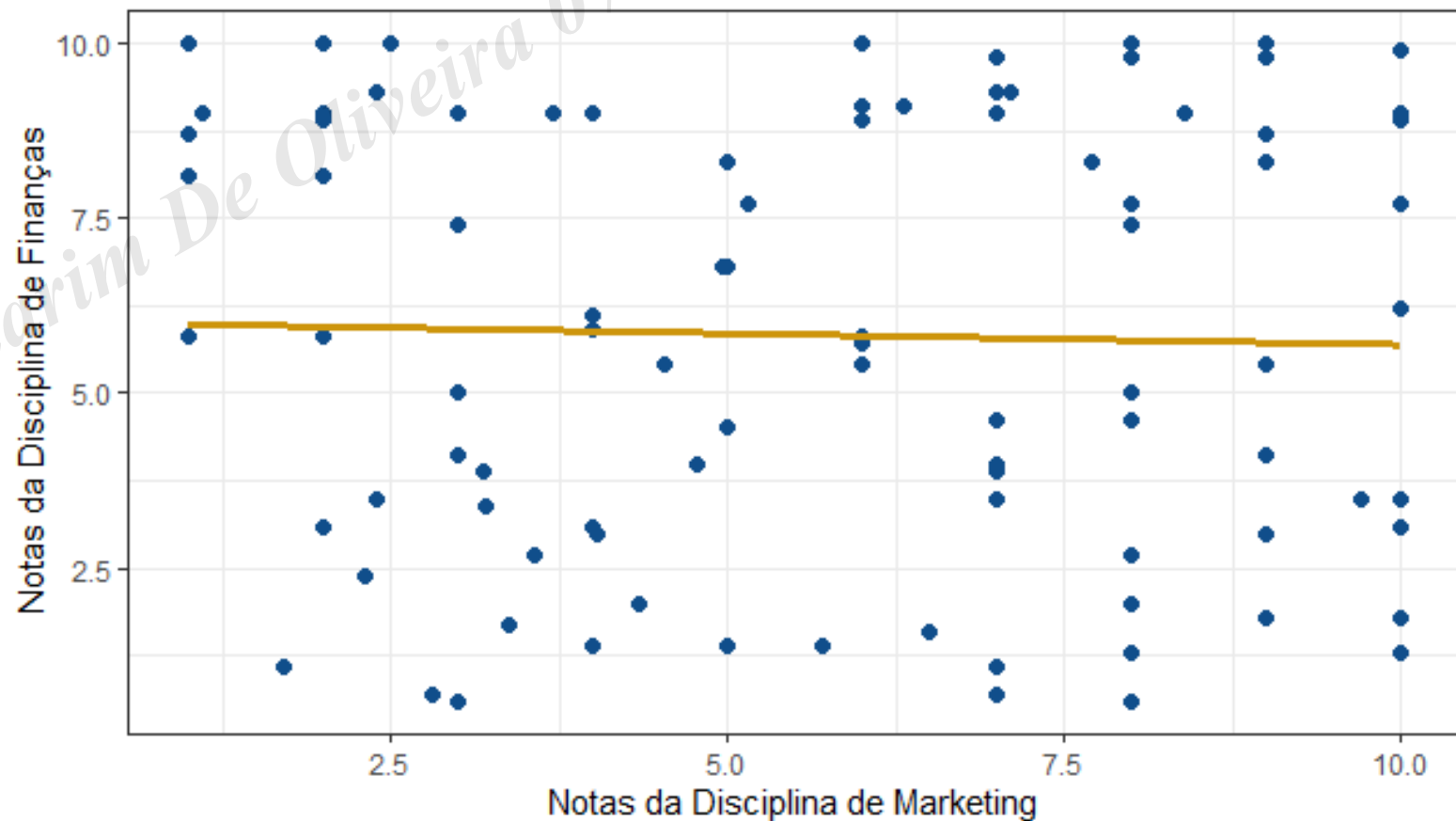
Redução Dimensional dos Dados

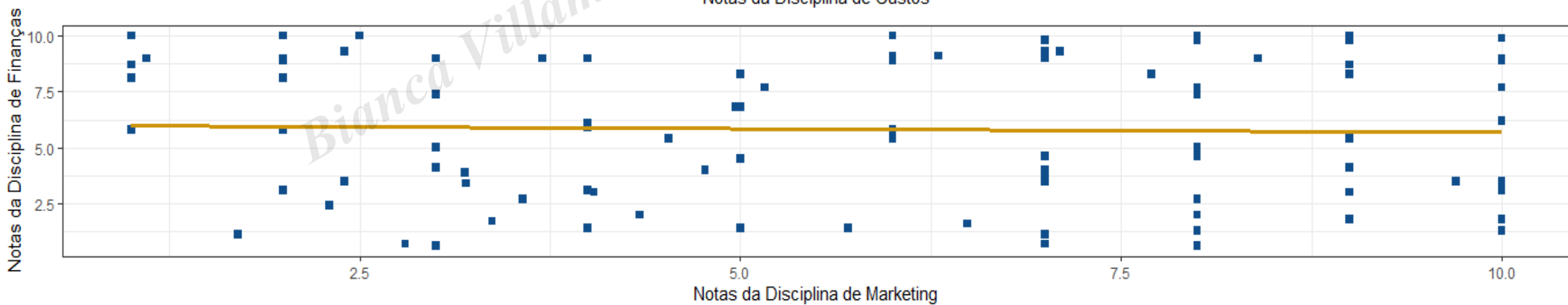
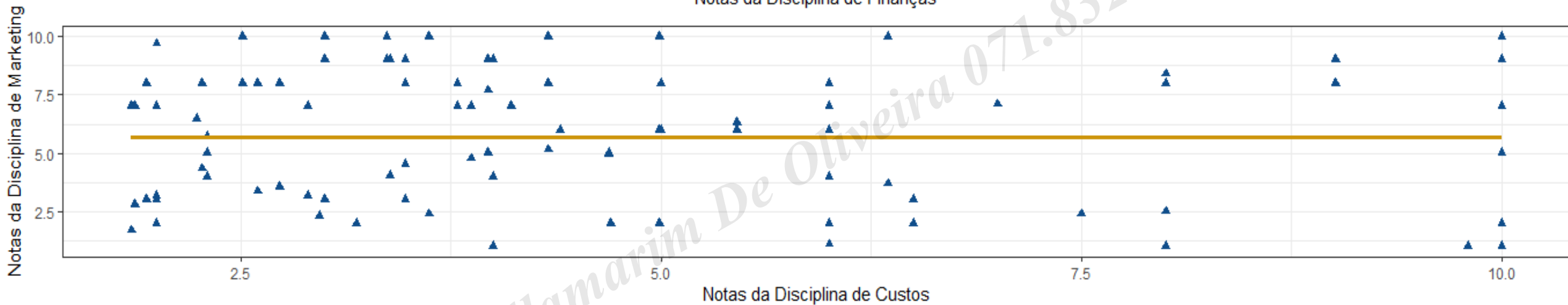
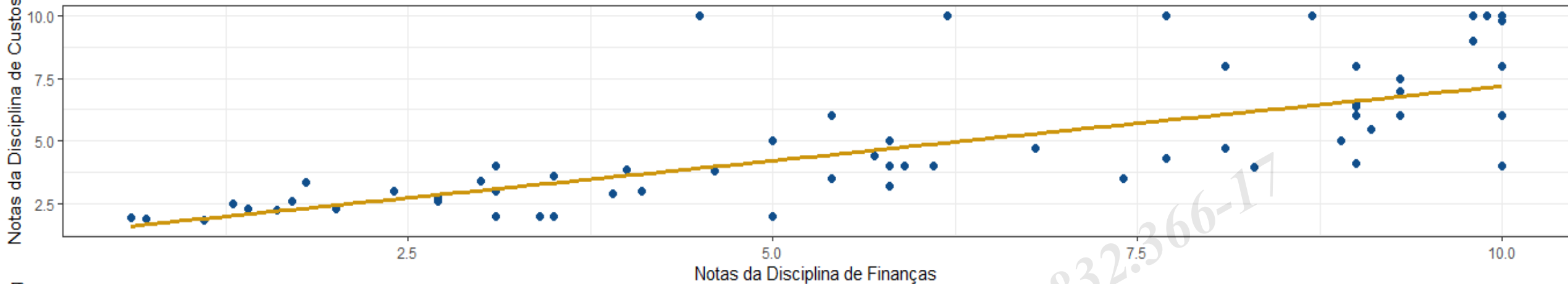


Redução Dimensional dos Dados

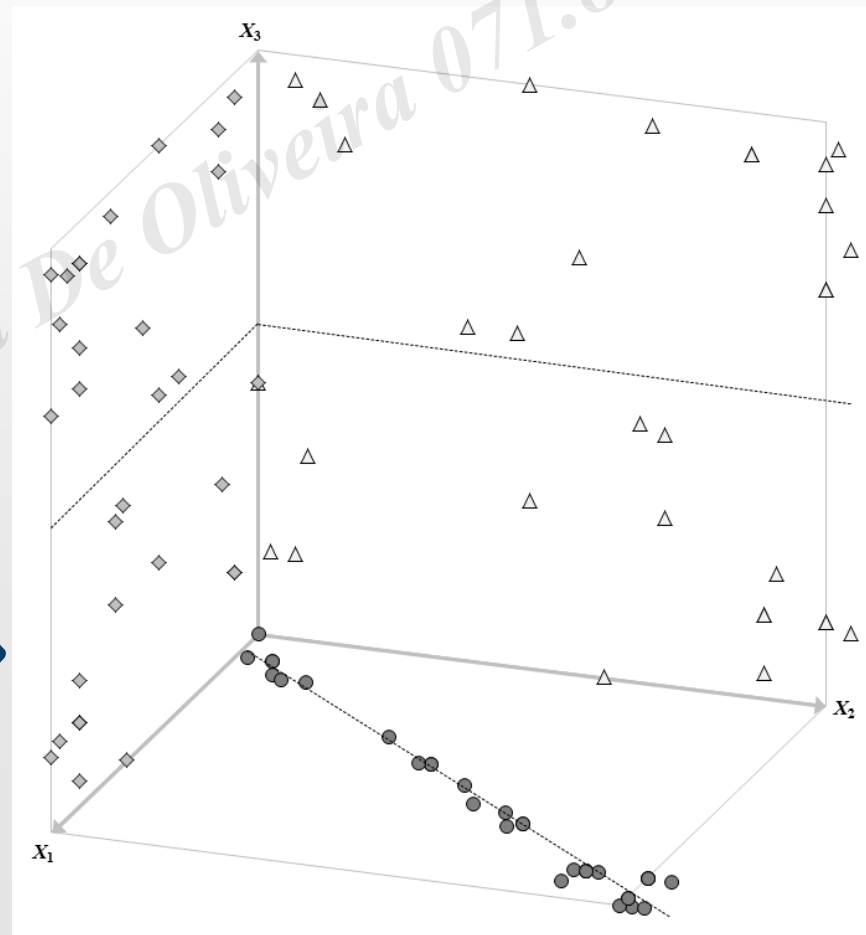


Redução Dimensional dos Dados



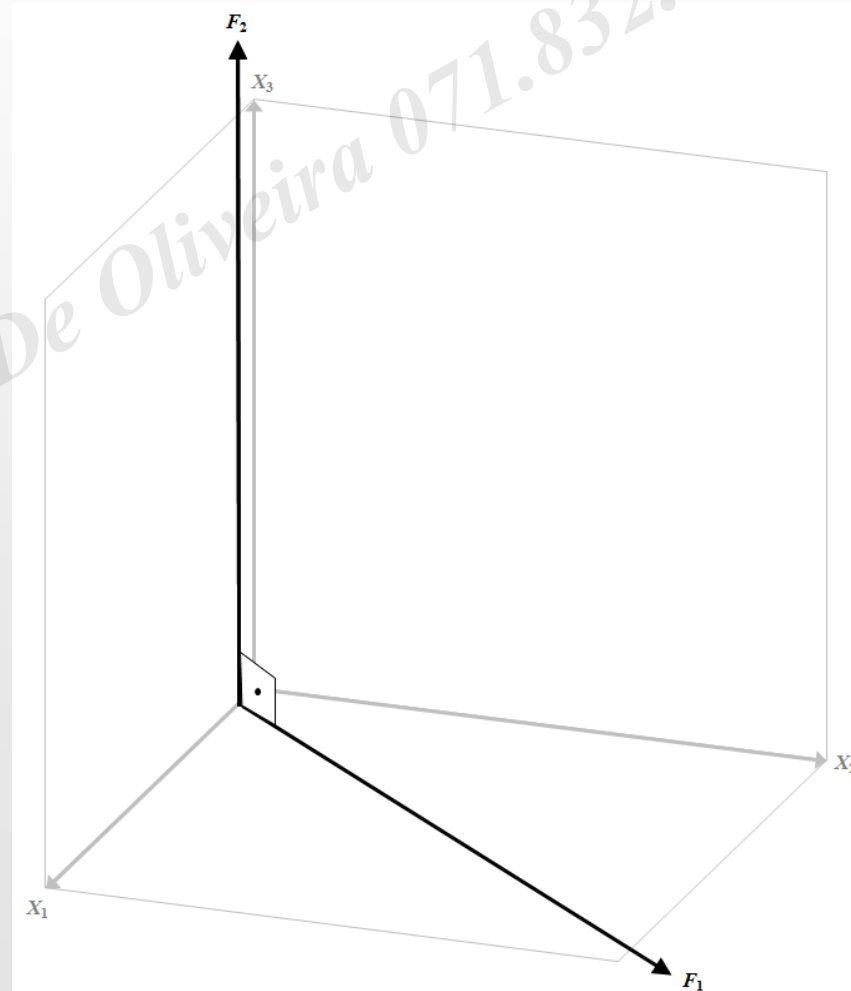


Redução Dimensional dos Dados



Fávero & Belfiore (2017)

Redução Dimensional dos Dados



Fávero & Belfiore (2017)

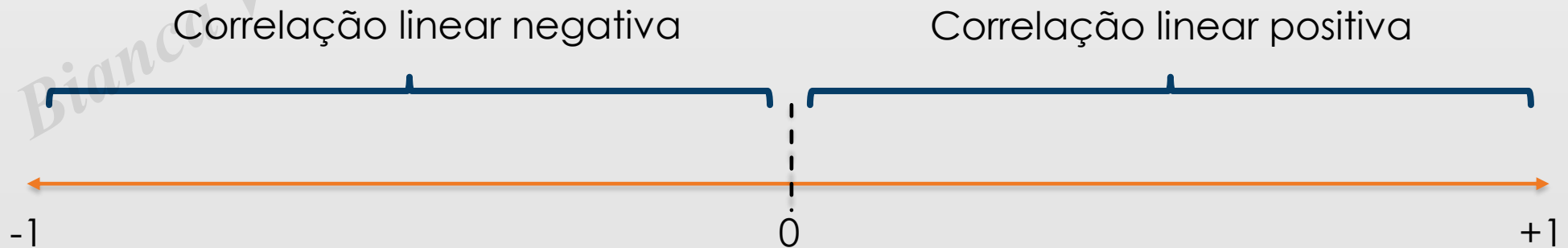


Primeiros Passos

Passos Iniciais: O Coeficiente de Correlação Linear de Pearson (ρ)

$$\rho = \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1) \cdot (X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2} \cdot \sqrt{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2}}$$

em que X_k , sendo $k = 1, 2, 3, \dots$, diz respeito às variáveis métricas da base de dados; \bar{X}_k refere-se às médias de cada variável X ; i representa cada uma das observações; e n aponta para o total de observações.



A Construção da Matriz de Correlações (ρ)

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix}$$

A matriz ρ deve ser simétrica em relação à diagonal principal; por óbvio, a diagonal principal apresentará valores iguais a 1.

A black and white photograph of a person from behind, standing in front of a chalkboard. The person has their hands clasped behind their back and is looking at two large white arrows drawn on the chalkboard. The arrows originate from a central point at the top and point outwards and downwards in opposite directions. A semi-transparent text box is overlaid on the bottom of the image.

Passo 2: Verificando a adequabilidade dos dados

A Adequabilidade dos Dados para a PCA

- A estatística Kaiser-Meyer-Olkin (KMO)
- O teste de esfericidade de Bartlett

$$\chi^2_{Bartlett} = - \left[(n-1) - \left(\frac{2k+5}{6} \right) \right] \cdot \ln|D|$$

com $\frac{k \cdot (k-1)}{2}$ graus de liberdade, em que D representa o determinante da matriz $\mathbf{\rho}$.

$$H_0: \mathbf{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} = \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$H_1: \mathbf{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \neq \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$



Relembrando o objetivo:

- Objetivo: determinar os fatores.
- Método: componentes principais.
- Combinações lineares das variáveis observadas;
- 1º componente principal: combinação que explica a maior quantidade de variância na amostra;
- 2º componente: 2ª maior quantidade de variância e não é correlacionada com a primeira; etc.
- Um conjunto de variáveis correlacionadas é transformado em um conjunto de variáveis não correlacionadas.



Passo 3: Definição dos fatores por componentes principais: determinação dos autovalores e autovetores da matriz de correlações ρ e cálculo dos scores fatoriais

Ainda Discutindo sobre os Fatores

- Um fator representa a combinação linear de k variáveis originais. Dessa forma, o número máximo possível de fatores (F) para determinada base de dados, será igual a k . Dessa forma, para o cálculo de F_k , teremos:

$$F_{1i} = s_{11} \cdot X_{1i} + s_{21} \cdot X_{2i} + \dots + s_{k1} \cdot X_{ki}$$

$$F_{2i} = s_{12} \cdot X_{1i} + s_{22} \cdot X_{2i} + \dots + s_{k2} \cdot X_{ki}$$

$$\vdots$$

$$F_{ki} = s_{1k} \cdot X_{1i} + s_{2k} \cdot X_{2i} + \dots + s_{kk} \cdot X_{ki}$$

em que s representa os scores fatoriais.

Scores Fatoriais (s)

- Os scores fatoriais dizem respeito a coeficientes que relacionam determinado fator com as variáveis originais. Seu cálculo perpassa pela determinação dos autovalores e autovetores da matriz de correlações ρ .

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix}$$

ρ terá sempre dimensões $k \times k$, com k autovalores $\lambda^2 (\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2)$, que podem ser obtidos com a solução da seguinte equação: $\det(\rho - \lambda^2 \cdot \mathbf{I}) = 0$.

- Como determinado fator representa o resultado do agrupamento de variáveis, é importante ressaltar que: $\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_k^2 = k$.

Autovalores (λ^2)

- Vetores independentes, advindos de uma matriz quadrada, que capturam o máximo da variabilidade dos dados. Pode-se entender, portanto, os autovalores como a parcelas da variância total capturada das variáveis estudadas.
- Então, matricialmente, $\det(\mathbf{p} - \lambda^2 \cdot \mathbf{I}) = 0$ será:

$$\det(\mathbf{p} - \lambda^2 \cdot \mathbf{I}) = \det \begin{pmatrix} 1 - \lambda^2 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 - \lambda^2 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 - \lambda^2 \end{pmatrix} = 0$$

Para fins didáticos, o polinômio característico do exemplo dado em aula é:

$$\lambda^4 - 4\lambda^3 + 4,2657473\lambda^2 - 1,4025534\lambda + 0,1371619$$

Autovalores (λ^2)

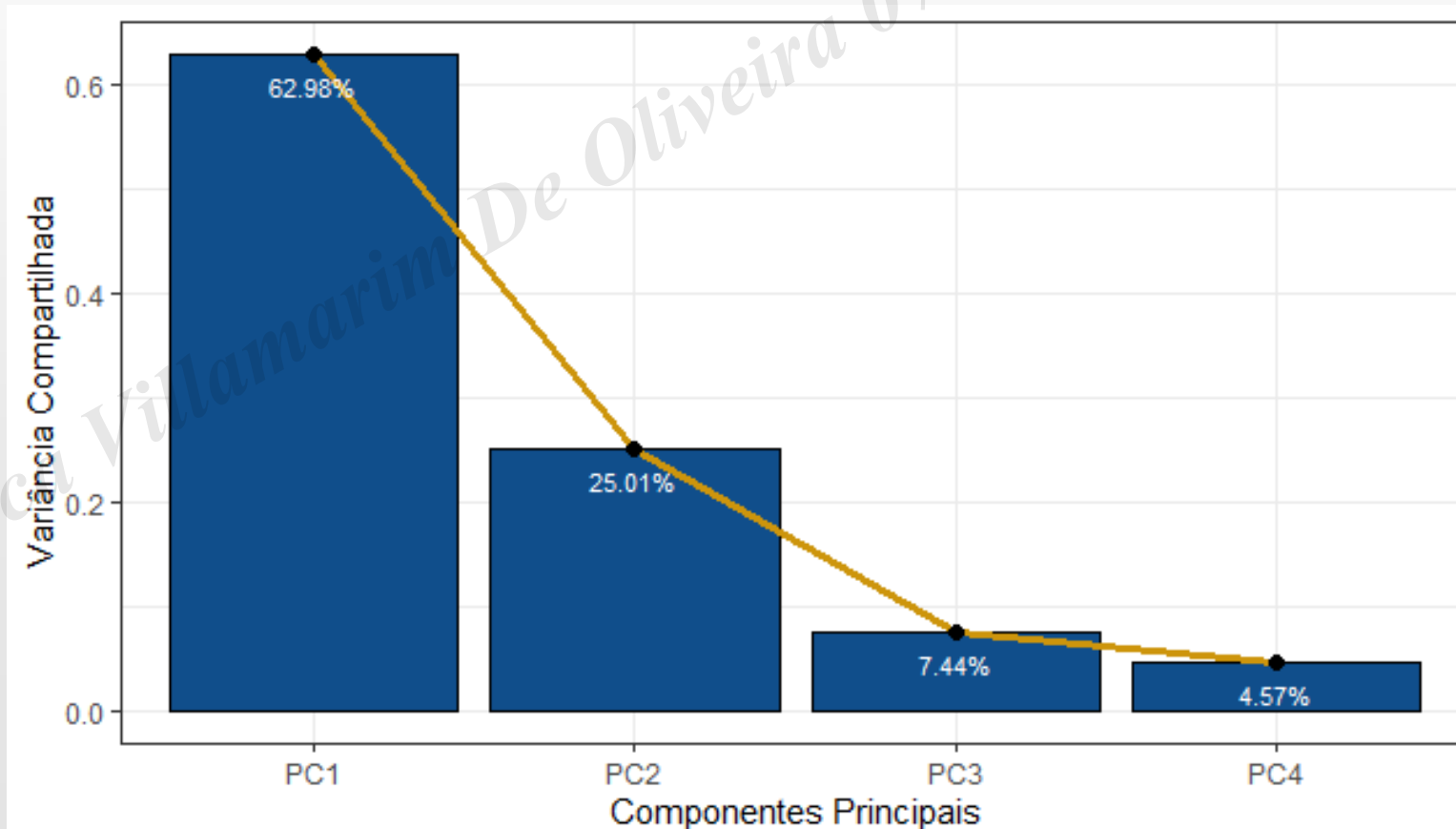
- As raízes do polinômio característico resultante de $\det(\mathbf{p} - \lambda^2 \cdot \mathbf{I}) = 0$ serão os eigenvalues λ_k^2 da matriz \mathbf{p} . A partir desses autovalores, se pode definir a matriz $\mathbf{\Lambda}^2$:

$$\mathbf{\Lambda}^2 = \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k^2 \end{pmatrix}$$

Posto isso, para que sejam definidos os autovetores (v) da matriz \mathbf{p} com base nos seus autovalores, devemos resolver alguns sistemas de equações para cada autovalor λ^2 ($\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2$).

Para as aulas, assumiremos o critério de Kaiser (critério da raiz latente), isto é, consideraremos os componentes principais com eigenvalues > 1 .

Demonstração Exemplificativa da Captura da Variância dos Dados



Autovetores (v)

- Os eigenvectors são projeções da variância capturada, utilizados para o cálculo dos scores fatoriais e das cargas fatoriais.
- A determinação dos autovetores $v_{11}, v_{21}, \dots, v_{k1}$, a partir do primeiro autovalor (λ_1^2), pode ser dada por:

$$\begin{pmatrix} \lambda_1^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_1^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_1^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{k1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

de onde vem que:

$$\begin{cases} (\lambda_1^2 - 1) \cdot v_{11} - \rho_{12} \cdot v_{21} - \cdots - \rho_{1k} \cdot v_{k1} = 0 \\ -\rho_{21} \cdot v_{11} + (\lambda_1^2 - 1) \cdot v_{21} - \cdots - \rho_{2k} \cdot v_{k1} = 0 \\ \vdots \\ -\rho_{k1} \cdot v_{11} - \rho_{k2} \cdot v_{21} - \cdots + (\lambda_1^2 - 1) \cdot v_{k1} = 0 \end{cases}$$

Autovetores (v)

- A determinação dos autovetores $v_{12}, v_{22}, \dots, v_{k2}$, a partir do segundo autovalor (λ_2^2), pode ser dada por:

$$\begin{pmatrix} \lambda_2^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_2^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_2^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{k2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

de onde vem que:

$$\begin{cases} (\lambda_2^2 - 1) \cdot v_{12} - \rho_{12} \cdot v_{22} - \cdots - \rho_{1k} \cdot v_{k2} = 0 \\ -\rho_{21} \cdot v_{12} + (\lambda_2^2 - 1) \cdot v_{22} - \cdots - \rho_{2k} \cdot v_{k2} = 0 \\ \vdots \\ -\rho_{k1} \cdot v_{12} - \rho_{k2} \cdot v_{22} - \cdots + (\lambda_2^2 - 1) \cdot v_{k2} = 0 \end{cases}$$

Autovetores (v)

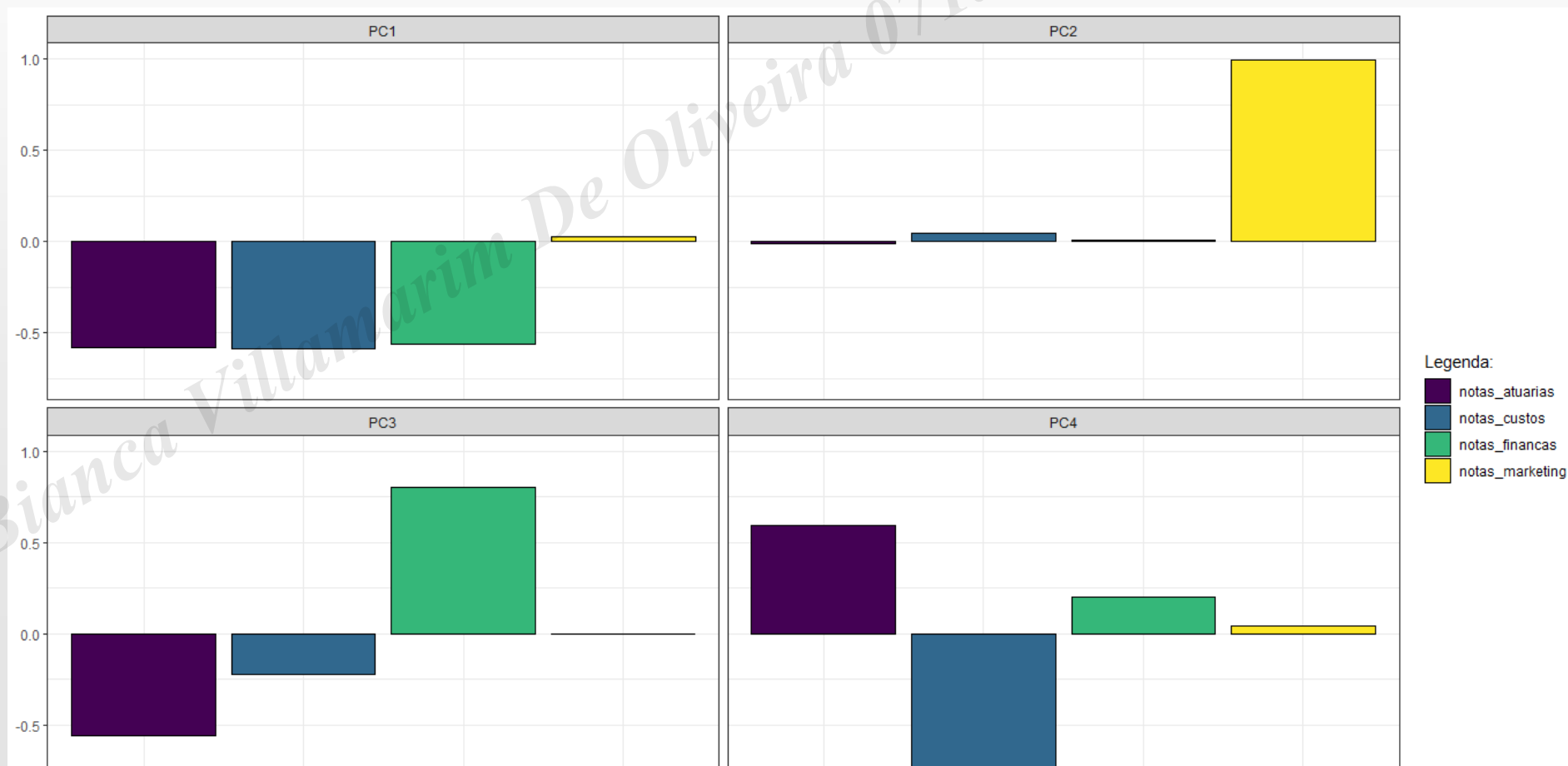
- A determinação dos autovetores $v_{1k}, v_{2k}, \dots, v_{kk}$, a partir do k -ésimo autovalor (λ_k^2), pode ser dada por:

$$\begin{pmatrix} \lambda_k^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_k^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_k^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{1k} \\ v_{2k} \\ \vdots \\ v_{kk} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

de onde vem que:

$$\begin{cases} (\lambda_k^2 - 1) \cdot v_{1k} - \rho_{12} \cdot v_{2k} - \cdots - \rho_{1k} \cdot v_{kk} = 0 \\ -\rho_{21} \cdot v_{1k} + (\lambda_k^2 - 1) \cdot v_{2k} - \cdots - \rho_{2k} \cdot v_{kk} = 0 \\ \vdots \\ -\rho_{k1} \cdot v_{1k} - \rho_{k2} \cdot v_{2k} - \cdots + (\lambda_k^2 - 1) \cdot v_{kk} = 0 \end{cases}$$

Demonstração Exemplificativa das Projeções da Variância Capturada



○ Cálculo dos Scores Fatoriais (s)

Scores fatoriais do primeiro fator:

$$s_1 = \begin{pmatrix} s_{11} \\ s_{21} \\ \vdots \\ s_{k1} \end{pmatrix} = \begin{pmatrix} \frac{v_{11}}{\sqrt{\lambda_1^2}} \\ \frac{v_{21}}{\sqrt{\lambda_1^2}} \\ \vdots \\ \frac{v_{k1}}{\sqrt{\lambda_1^2}} \end{pmatrix}$$

Scores fatoriais do

fatoriais do segundo

$$s_2 = \begin{pmatrix} s_{12} \\ s_{22} \\ \vdots \\ s_{k2} \end{pmatrix} = \begin{pmatrix} \frac{v_{12}}{\sqrt{\lambda_2^2}} \\ \frac{v_{22}}{\sqrt{\lambda_2^2}} \\ \vdots \\ \frac{v_{k2}}{\sqrt{\lambda_2^2}} \end{pmatrix}$$

fatoriais do

Scores fatoriais do k -ésimo fator:

$$s_k = \begin{pmatrix} s_{1k} \\ s_{2k} \\ \vdots \\ s_{kk} \end{pmatrix} = \begin{pmatrix} \frac{v_{1k}}{\sqrt{\lambda_k^2}} \\ \frac{v_{2k}}{\sqrt{\lambda_k^2}} \\ \vdots \\ \frac{v_{kk}}{\sqrt{\lambda_k^2}} \end{pmatrix}$$

Cargas Fatoriais e Comunalidades

- As cargas fatoriais correspondem às correlações de Pearson entre os fatores e as variáveis originais utilizadas para a elaboração da PCA.



- As comunalidades representam a variância total compartilhada de cada uma das variáveis originais com todos os fatores extraídos.