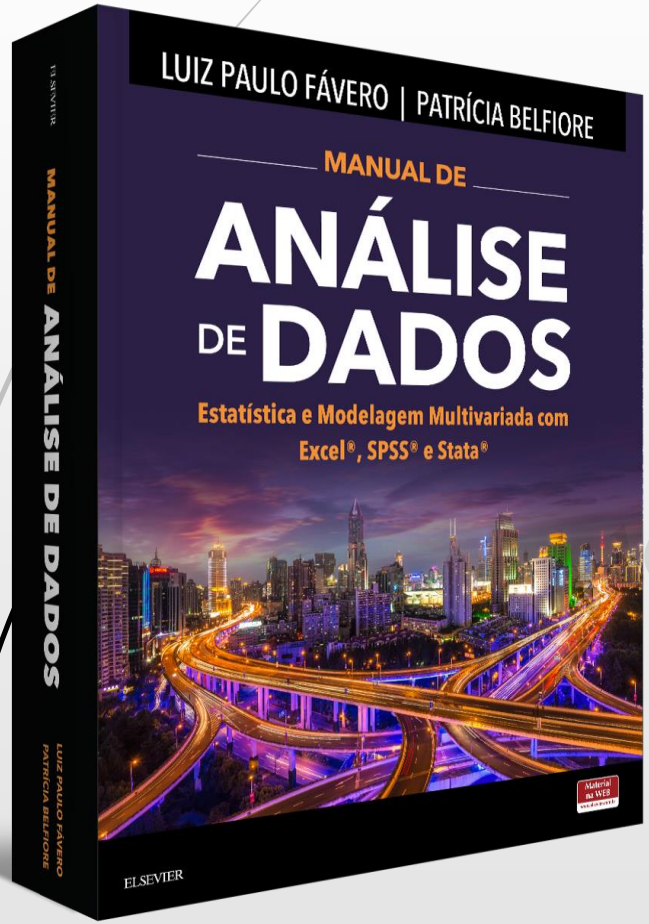
MBA USP ESALQ

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

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.

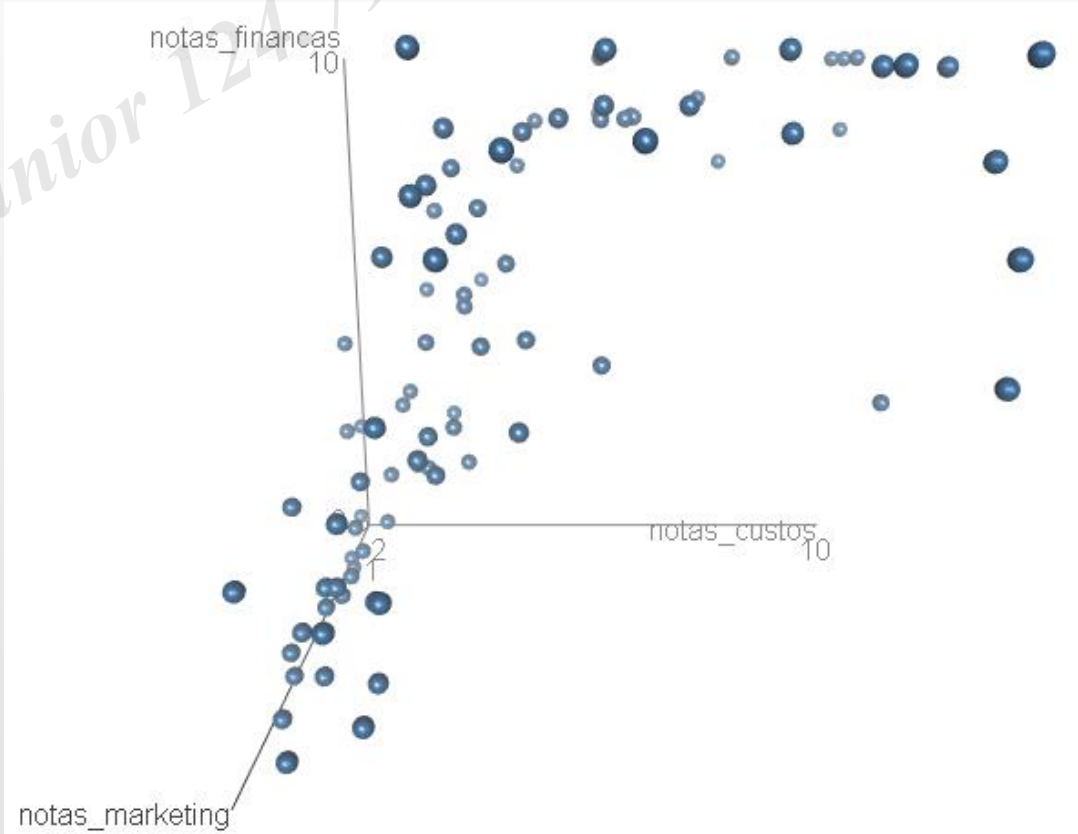
Análise Fatorial por Componentes Principais x Análise de Clusters

	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.

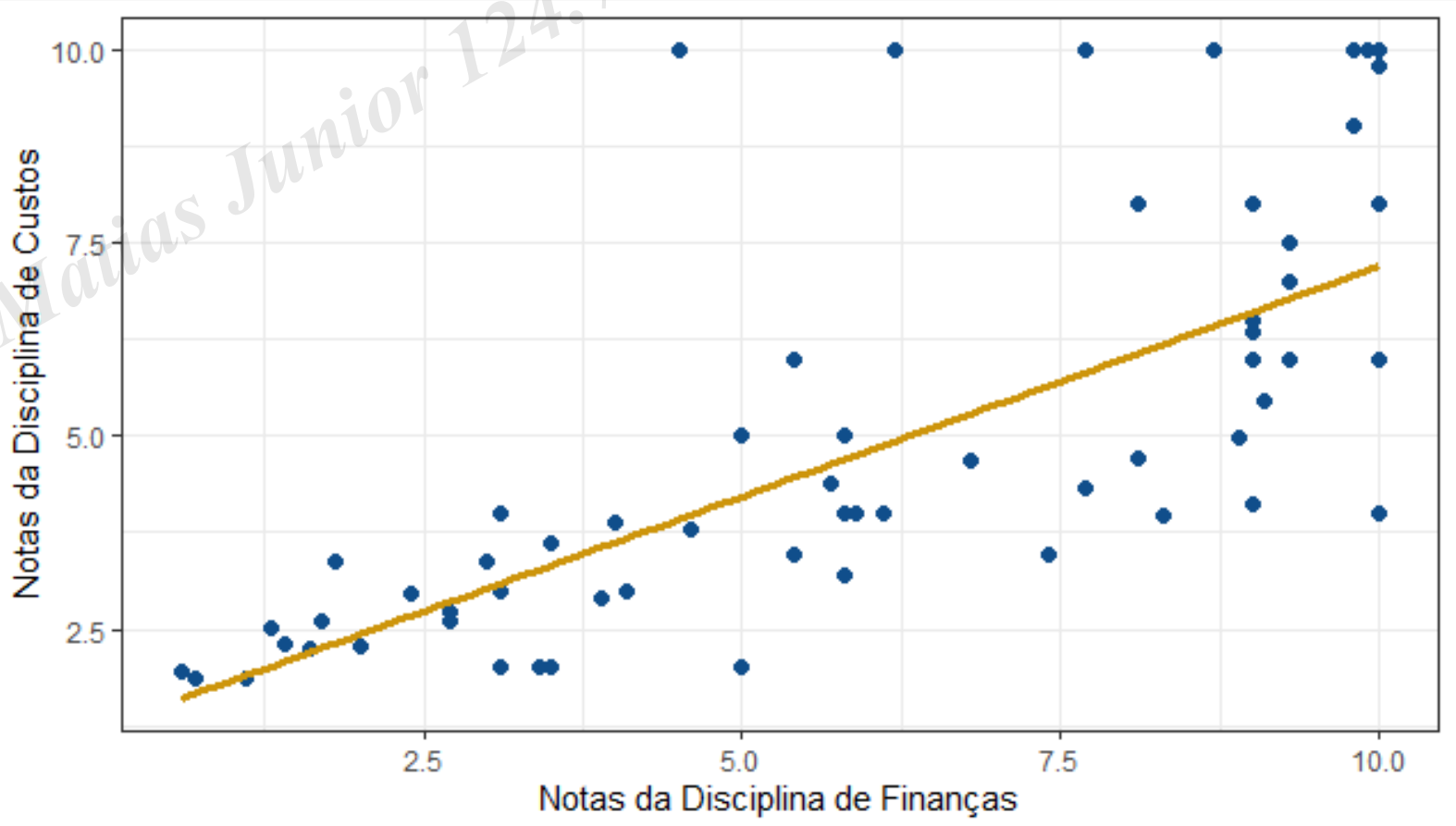
Outros Objetivos Relevantes da PCA

- 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.

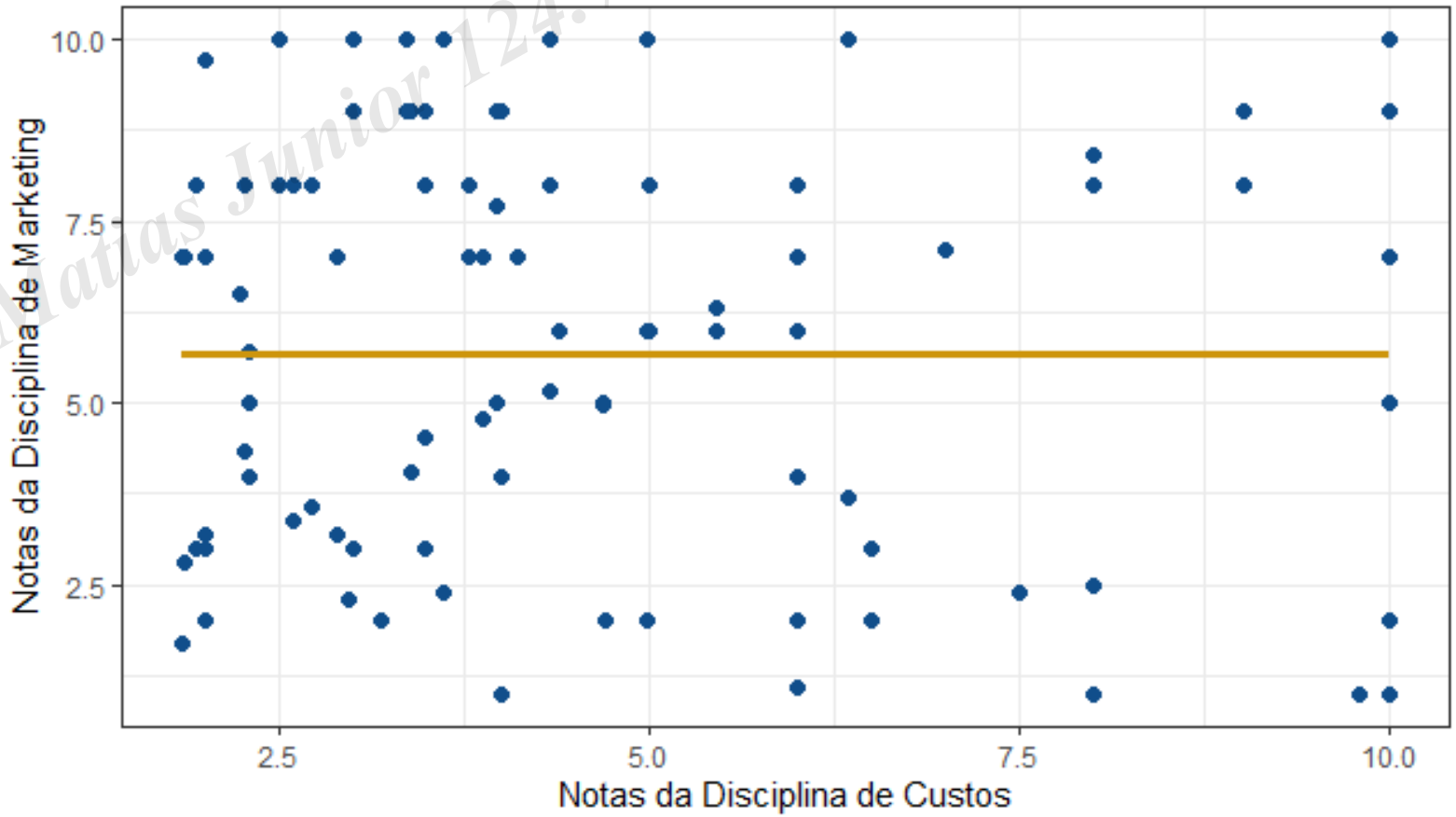
Redução Dimensional dos Dados



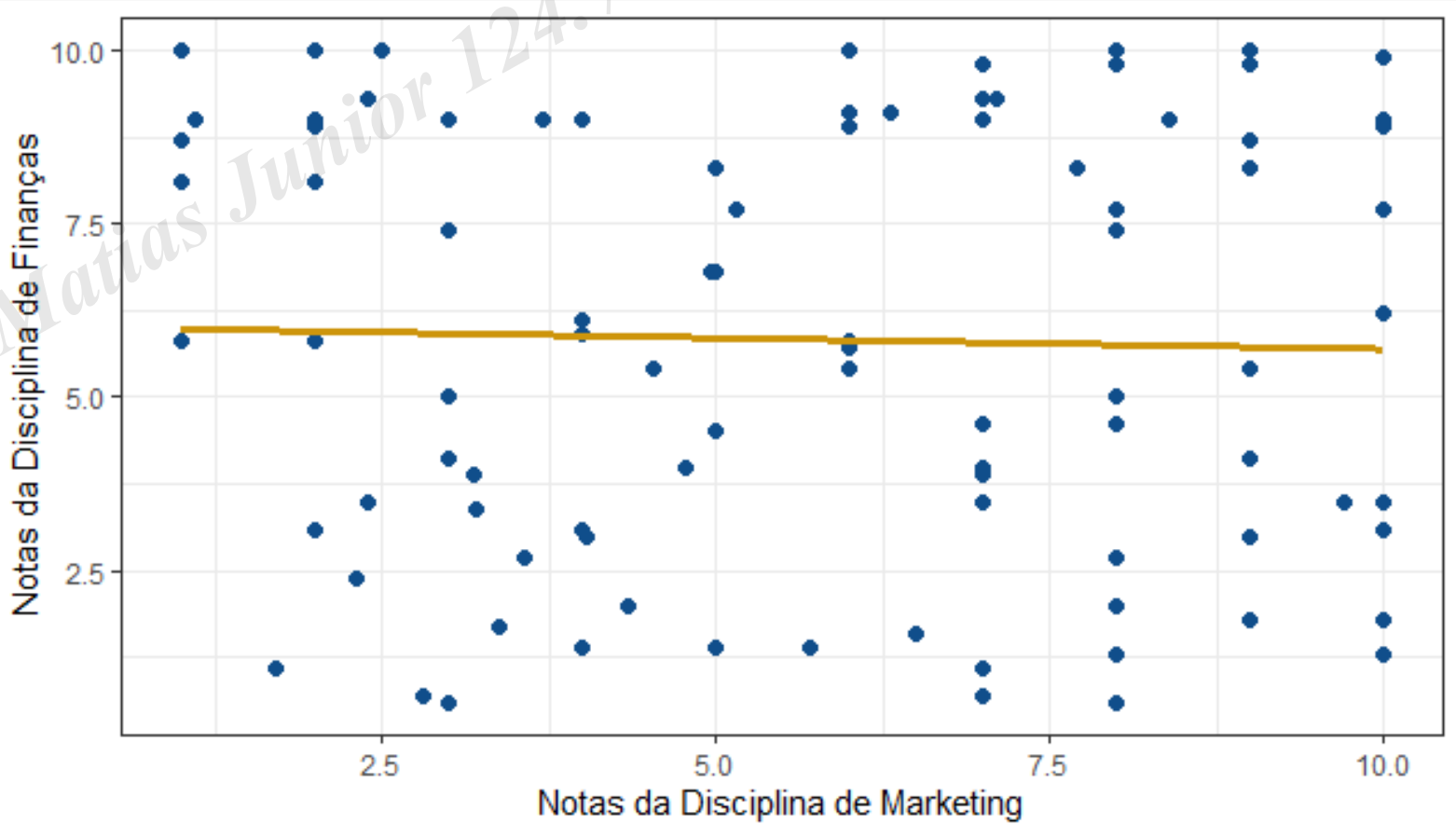
Redução Dimensional dos Dados

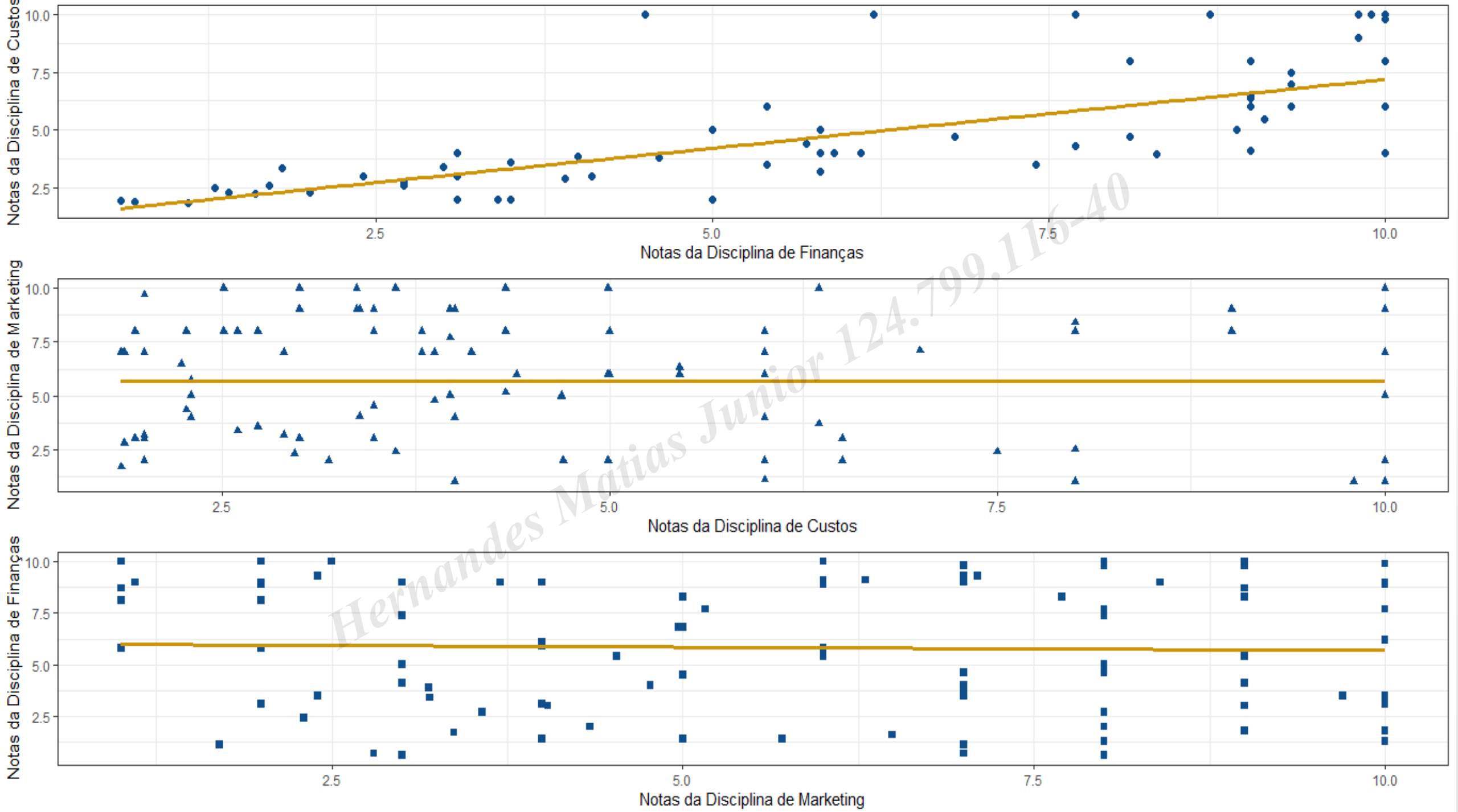


Redução Dimensional dos Dados

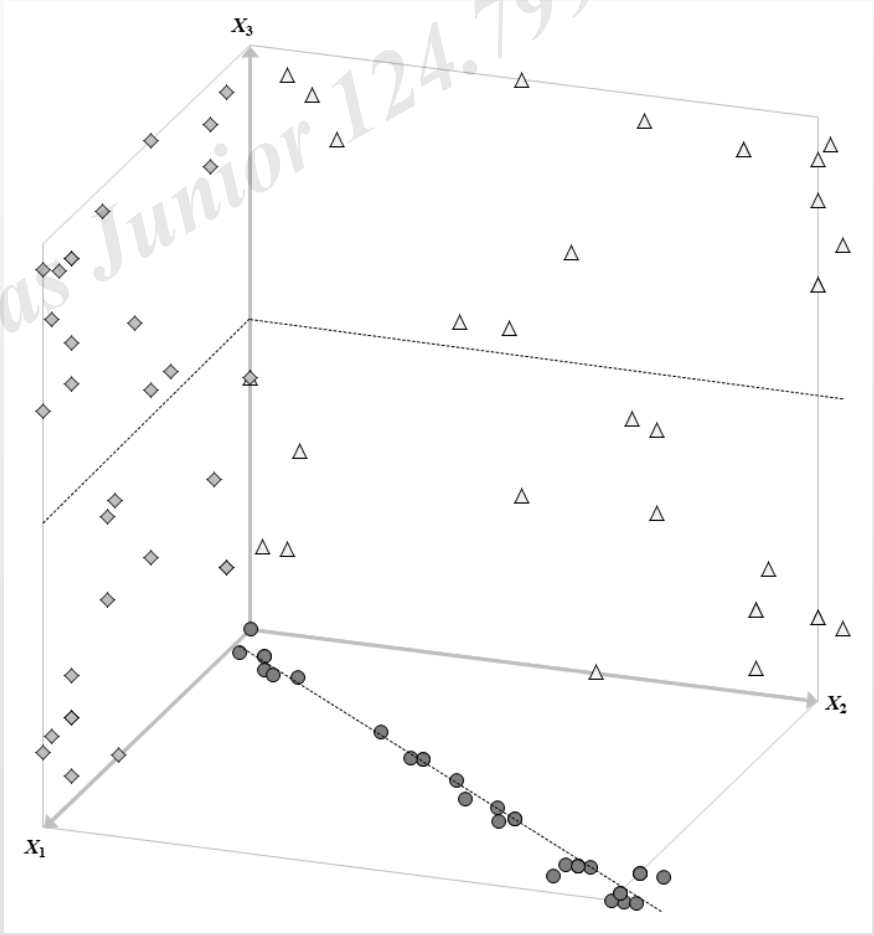


Redução Dimensional dos Dados



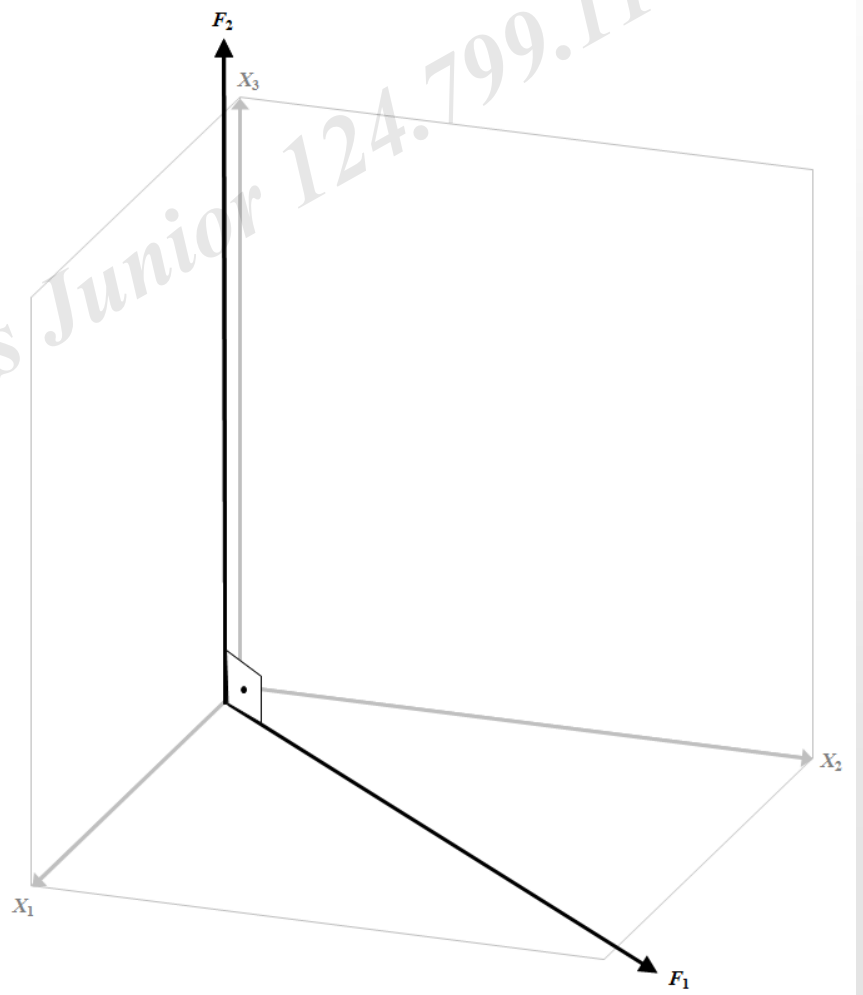


Redução Dimensional dos Dados



Fávero & Belfiore (2017)

Redução Dimensional dos Dados



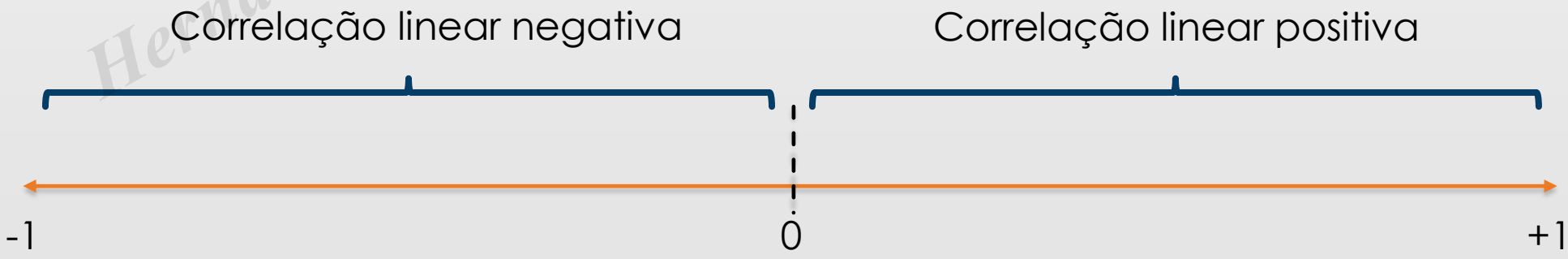
Fávero & Belfiore (2017)

Primeiros Passos

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

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

em que X_k, sendo $k = 1,2,3, \ldots$, 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 ($\boldsymbol{\rho}$)

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

A matriz $\boldsymbol{\rho}$ deve ser simétrica em relação à diagonal principal; por óbvio, a diagonal principal apresentará valores iguais a 1.

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

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

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

$$\mathrm{H}_0: \boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} = \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$\mathrm{H}_1: \boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \neq \mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Relembrando o objetivo:

- Objetivo: determinar os fatores.
- Método: componentes principais.
- Combinações lineares das variáveis observadas;
- 1° componente principal: combinação que explica a maior quantidade de variância na amostra;
- 2° 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:

$$F_{1i} = s_{11}.X_{1i} + s_{21}.X_{2i} + \cdots + s_{k1}.X_{ki}$$
$$F_{2i} = s_{12}.X_{1i} + s_{22}.X_{2i} + \cdots + s_{k2}.X_{ki}$$
$$\vdots$$
$$F_{ki} = s_{1k}.X_{1i} + s_{2k}.X_{2i} + \cdots + s_{kk}.X_{ki}$$

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 $\boldsymbol{\rho}$.

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

$\boldsymbol{\rho}$ 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: $\det(\boldsymbol{\rho} - \lambda^2.\mathbf{I}) = 0$.

- 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, $\det(\boldsymbol{\rho} - \lambda^2.\mathbf{I}) = 0$ será:

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

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

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

Autovalores (λ^2)

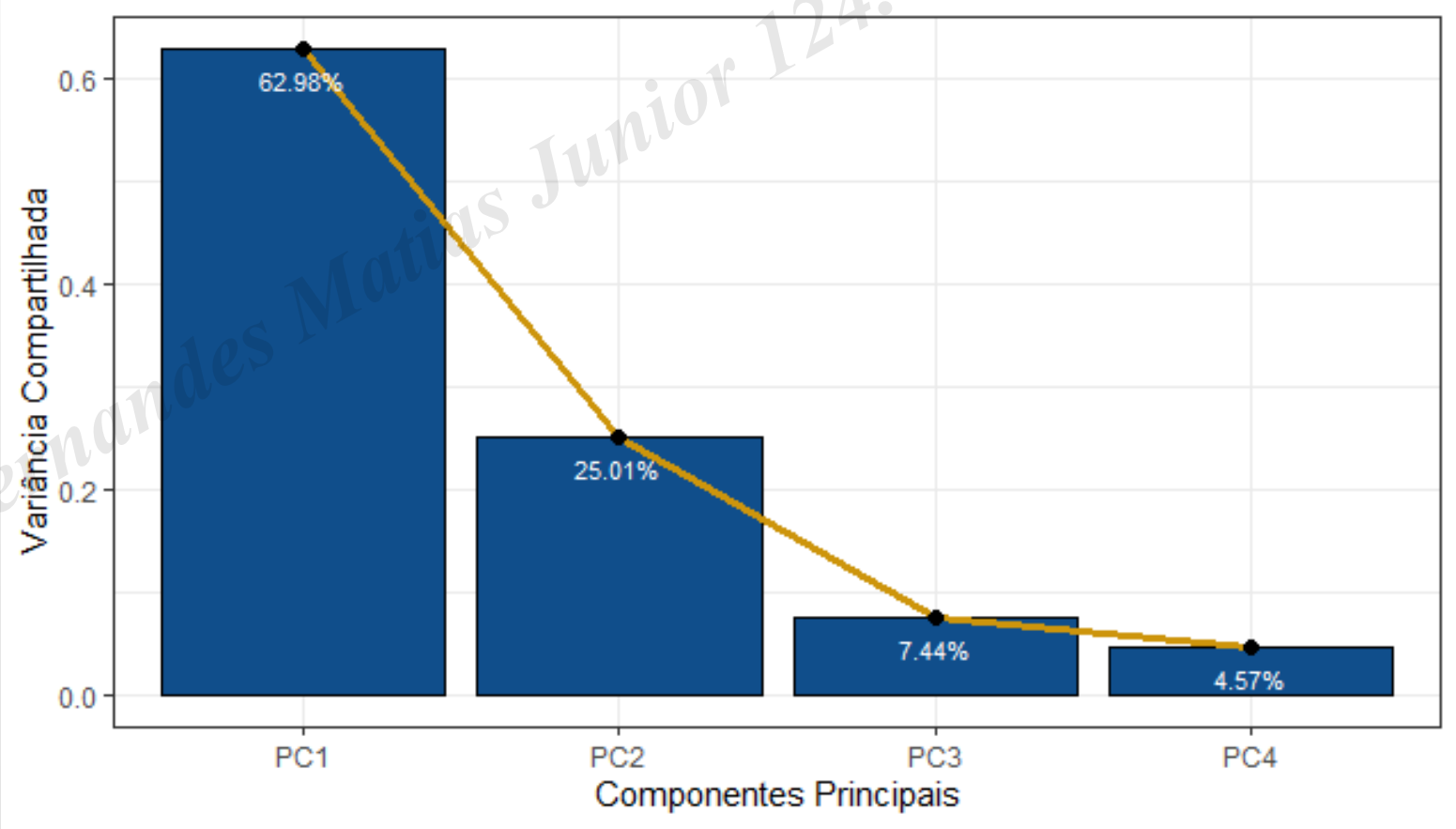
- As raízes do polinômio característico resultante de $\det(\boldsymbol{\rho} - \lambda^2.\mathbf{I}) = 0$ serão os eigenvalues λ_k^2 da matriz $\boldsymbol{\rho}$. A partir desses autovalores, se pode definir a matriz $\boldsymbol{\Lambda}^2$:

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

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

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 $v_{11}, v_{21}, \ldots, v_{k1}$, a partir do primeiro autovalor (λ_1^2), pode ser dada por:

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

de onde vem que:

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

Autovetores (v)

- A determinação dos autovetores $v_{12}, v_{22}, \ldots, v_{k2}$, 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:

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

Autovetores (v)

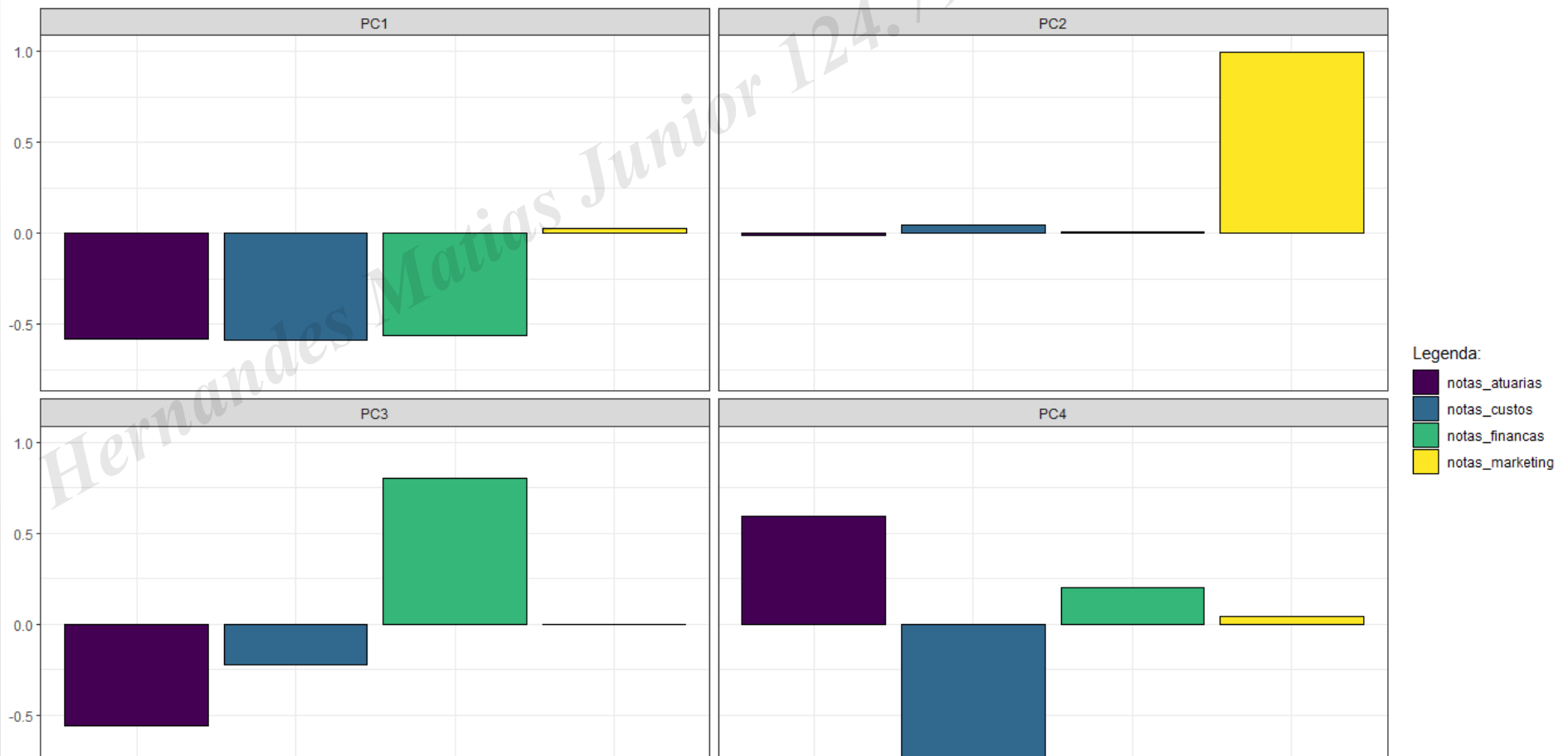
- A determinação dos autovetores $v_{1k}, v_{2k}, \ldots, v_{kk}$, a partir do k-ésimo autovalor (λ_k^2), pode ser dada por:

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

de onde vem que:

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

Demonstração Exemplificativa das Projeções da Variância Capturada

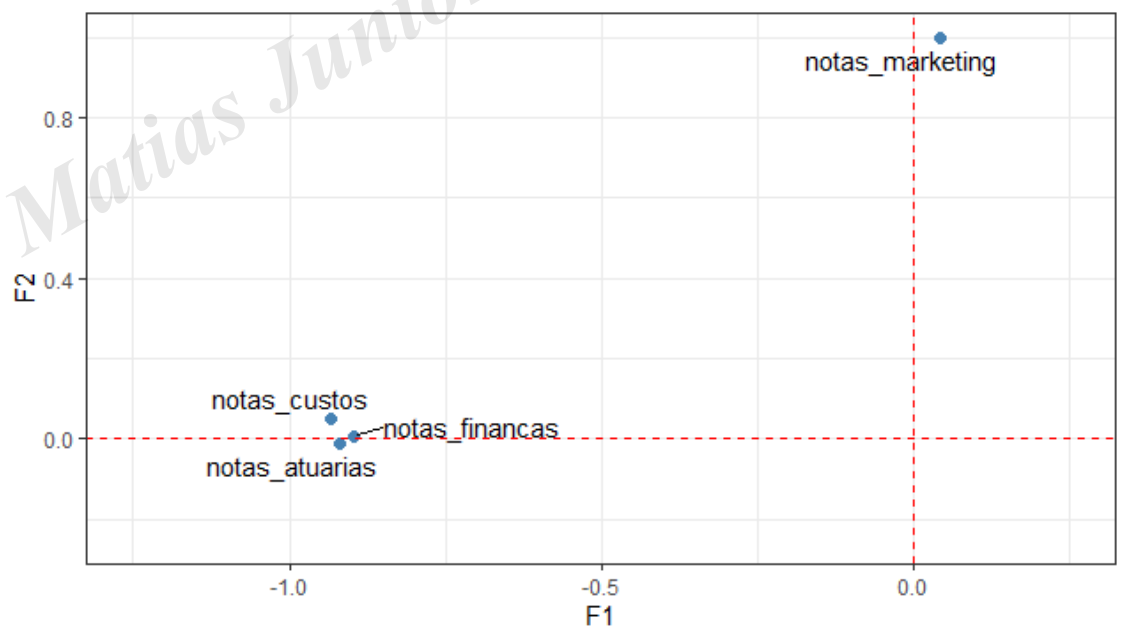


O Cálculo dos *Scores* Fatoriais (*s*)

Scores fatoriais do primeiro fator:	Scores fatoriais do segundo fator:	Scores fatoriais do k-ésimo fator:
$s_1 = \begin{pmatrix} s_{11} \\ s_{21} \\ \vdots \\ s_{k1} \end{pmatrix} = \begin{pmatrix} \frac{v_{11}}{\sqrt{\lambda_1^2}} \\ \frac{v_{21}}{\sqrt{\lambda_1^2}} \\ \vdots \\ \frac{v_{k1}}{\sqrt{\lambda_1^2}} \end{pmatrix}$	$s_2 = \begin{pmatrix} s_{12} \\ s_{22} \\ \vdots \\ s_{k2} \end{pmatrix} = \begin{pmatrix} \frac{v_{12}}{\sqrt{\lambda_2^2}} \\ \frac{v_{22}}{\sqrt{\lambda_2^2}} \\ \vdots \\ \frac{v_{k2}}{\sqrt{\lambda_2^2}} \end{pmatrix}$	$s_k = \begin{pmatrix} s_{1k} \\ s_{2k} \\ \vdots \\ s_{kk} \end{pmatrix} = \begin{pmatrix} \frac{v_{1k}}{\sqrt{\lambda_k^2}} \\ \frac{v_{21k}}{\sqrt{\lambda_k^2}} \\ \vdots \\ \frac{v_{kk}}{\sqrt{\lambda_k^2}} \end{pmatrix}$

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.