Multivariate Statistical Analysis: Principal Component Analysis

rafael.poerschke@gmail.com

26 de maio de 2023

Resumo

Rudimentos de álgebra para Análise Multivariada, Análise dos Componentes Principais e Análise Fatorial Exploratória.

1 Introduction

Nas últimas três décadas, principalmente, as pesquisas acadêmicas e aplicada sofreram importantes avanços, muito em decorrência da evolução dos hardwares do computador pessoal (PC) e dos avanços e da facilidade de acesso aos softwares estatísticos livres, como: R-package, Gretl, Winbugs, J-Multi, bem como outros licenciados como: Stata, SAS, SPSS, MatLab, Eviews. As interfaces destes softwares estão cada vez mais amigáveis, o que facilita o acesso de usuários não muito familiarizados com a linguagem de programação.

A disponibilidade desses recursos computacionais permite que se possa trabalhar com banco de dados muito grandes, tanto em número de variáveis quanto em quantidade de observações, bem como utilizar qualquer modelo matemático ou estatístico/econométrico de interesse do pesquisador. Na área de economia, os modelos mais recorrentes são de regressão *cross-section* ou de série temporal, dados em painel, redes neurais artificiais (RNA), equilíbrio geral computável, *fuzzy* e ferramentas da estatística multivariada (componentes principais, análise fatorial exploratória, análise fatorial confirmatória, análise de discriminantes, análise de *clusters* e correlação canônica).

De acordo com Favero and Belfiore (2017), a humanidade, em tempos atuais, tem convivido com cinco dimensões dos dados: volume, velocidade, variedade, variabilidade e complexidade. O gigantesco volume de dados decorre, entre outros, do incremento da ca-

pacidade computacional, do aumento do monitoramento dos fenômenos estudados e das mídias sociais. A velocidade diz respeito à rapidez com que os dados são disponibilizados para tratamento e análise, já que muitos desses são coletados por meio de etiquetas eletrônicas e sistema com radiofrequência. Os autores ressaltam três variedades em que os dados podem ser acessados, como: texto, indicadores e discursos. Os dados podem sofrer de variabilidade cíclica ou sazonal, por vezes por alta frequência, diretamente observável ou não. Por fim, a complexidade, pois os dados em grandes volumes podem ser acessados com fontes, frequências ou critérios distintos, o que exige do pesquisador uma análise integrada para a consolidar criteriosamente a base dos dados.

É nesse contexto que a estatística multivariada desempenha papel fundamental, pois permite estudar fenômenos complexos ao realizar o tratamento de diversas variáveis simultaneamente, mesmo quando não se conhece um modelo teórico que relacione essas variáveis entre si ou às observações – casos. Talvez, essa seja também a maior crítica à técnica, uma vez que sua aplicação estaria descolada de um marco teórico[1] como cerne (JOHNSON and WICHERN, 1992). Muito embora a estatística multivariada seja utilizada em larga escala na biologia, na economia, nas engenharias e na psicologia, quando comparada à econometria, mas nas ciências sociais como um todo ela não é tão popular. Essa é uma afirmação que se confirma na medida em que não consta em programas de graduação, nas Áreas de Ciências Sociais e Ciências Sociais Aplicadas, uma disciplina formal afeita ao tema, assim como fica à margem da maioria das disciplinas oferecidas no nível da pós-graduação.

Para um biólogo[2], não é necessário ter uma teoria prévia, basta padrões observados por um conjunto de características entre indivíduos para inferir, por exemplo, se uma rã encontrada em uma úmida floresta distante pertence a uma família já determinada. Dependendo das características desse indivíduo, existe a possibilidade de se criar um novo gênero para caracterizar sua espécie. O mesmo vale na psicologia[3] e psiquiatria no âmbito de seus estudos comportamentais. Sem a presença de sintomas verificáveis por exames convencionais, como uma doença de pele, uma artéria obstruída, os transtornos comportamentais

[1] Essa é uma afirmação que não é validada para a Análise Fatorial.

[2] Reyment and Jvreskog (1993) é apenas um exemplo de manual de análise multivariada voltada às Ciências Naturais. Ainda, cabe apresentar o pacote do R chamado *ADE*4 ou Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences.

[3] Gorsuch (1993), por sua vez, é um manual para a aplicação da análise fatorial em psicologia comportamental. Cabe destacar que o trabalho seminal no campo da psicologia com análise fatorial foi desenvolvido por Spearman (1904). No R, existe o pacote *PSYCH*: Procedures for Psychological, Psychometric, and Personality Research para auxiliar os pesquisadores da área com esse tipo de análise.

são verificados quando se tenta enquadrar indivíduos dentro de um padrão de comportamento classificado como normal. Os indivíduos se distribuem ao redor desse normal, que pode ser entendido como a média, e, quanto mais distante do comportamento esperado ele estiver, mais desajustado serão os traços emocionais do indivíduo em relação ao grupo dos considerados normais.

Na economia, por sua vez, grande parte das pesquisas se deflagra com a definição clara de hipóteses sobre fenômenos observados antecipadamente, para então, investigar com profundidade sua ocorrência caso a caso. Caso exista um grande número de publicações nacionais e internacionais na área de economia, quer seja em livros e periódicos, que usaram as ferramentas da estatística multivariada, a opção foi pela restrição de algumas dessas publicações, destacadas a seguir.

Mingoti and da Silva (1997) utilizaram análise de agrupamentos para definir os itens que deveriam constar nos grupos de gastos das famílias em índices de preços ao consumidor. Muito embora haja parâmetros baseados em índices previamente construídos, não é necessário se utilizar desses, uma vez que a análise de agrupamentos não carece dessa informação *a priori* para responder ao problema de pesquisa proposto.

Kageyama and Leone (1999) buscaram construir uma tipologia de economias regionais em um conjunto de municípios a partir de suas principais características sociais e econômicas. Para as autoras, a maioria dos estudos à época consistiam de análises em profundidade de localidades específicas, então elas exploraram a possibilidade de gerar tipologias territoriais[4] passíveis de se obter uma compreensão mais abrangente da problemática por meio da geração de unidades territoriais maiores e mais homogêneas para o estudo das famílias agrícolas.

Por sua vez, e com um olhar centrado no Rio Grande do Sul, Schneider and Waquil (2001)[5], de Freitas et al. (2007) e Poerschke (2007)[6], procuraram agrupar e caracterizar municípios ou regiões do estado do Rio Grande do Sul conforme suas semelhanças, observando

[4]Os resultados das autoras mostraram cinco regiões relativamente homogêneas no estado de São Paulo: rural muito pobre, rural pobre, intermediária, urbano em expansão e urbano denso. Essas tipificações foram descritas em termos de renda, população e produção agrícola locais.

[5]Os autores classificaram o Rio Grande do Sul em cinco grupos homogêneos de municípios, sendo que dois deles tinham em comum a pobreza rual e a degradação dos recursos naturais. Ao fim e ao cabo, os autores acabaram por descartar a ideia de que o estado podia ser dividido em duas partes, isto é, entre uma metade sul mais atrasada, e o norte desenvolvido.

[6]Enquanto de Freitas et al. (2007) abordaram o estado do Rio Grande do Sul em sua totalidade, Poerschke (2007) explorou os Conselhos Regionais do Rio Grande do Sul (COREDES) predominantemente agropecuários. A ideia principal de ambos era identificar padrões dado o grau de utilização de insumos agrícolas modernos nos grupos formados e sua relação com a renda agrícola e a atividade desenvolvida no estabelecimento rural.

a matriz produtiva e variáveis socioeconômicas.

Firme and Vasconcelos (2015) buscaram identificar nichos de mercado relacionando os países importadores às exportações brasileiras. Nesse exercício, os autores identificaram que as exportações de produtos da indústria pesada (metalomecânica) e tecnológicos (máquinas, reatores e eletrônicos) estavam entre o principal fator de diferenciação dos países de destino.

Por que a estatística multivariada se apresenta útil em estudos aplicados? A estatística multivariada procura reduzir uma matriz de dados inicial em um novo conjunto, menor e mais homogêneo, ou seja, quando se usa a análise de componentes principais é possível reduzir as informações contidas em p -variáveis originais pelas informações em k componentes principais (CP), de maneira que o número de componentes principais seja menor que o número de variáveis ($k < p$) para que o pesquisador ou agente tomador de decisão possa deflagrar uma análise.

Assim, a análise multivariada é o ramo da análise estatística preocupado em verificar a relação entre grupos de variáveis correlacionadas entre si (Morrison et al., 1976). O objetivo de otimizar a interpretação de grandes conjuntos de dados, em um número bem menor de variáveis latentes é de grande utilidade na análise econômica, uma vez que é notório o acesso a dados de corte transversal, como é o caso dos dados dos Censos Agropecuários. Esses dados serão o objeto ilustrativo dessa técnica, a fim de mostrar a utilidade e justificar a necessidade que um pesquisador tem de interpretar e operacionalizar um conjunto de dados dessa magnitude.

Conforme já ressaltado, há uma série de modelos, ou ferramentas da estatística multivariada, que podem ser aplicadas quando o propósito é tornar essa quantidade de informação, contidas em uma base de dados, num número menor de variáveis latentes e ortogonais[7] (Mingoti and da Silva, 1997). Entretanto, o foco do presente estudo inicia com a análise dos componentes principais (ACP), com a utilização da linguagem estatística R e a sugestão de um algoritmo[8] o qual segue em detalhes na próxima seção.

==INTRO APRESENTANDO A ESTRUTURA DO TEXTO==

[7] A ortogonalidade implica correlação nula entre as variáveis.

[8] Um algoritmo nada mais é, do que uma receita ou um passo a passo de procedimentos encadeados que conduzem o usuário na solução de uma tarefa. Em termos técnicos, ele é uma sequência lógica, finita e definida de instruções que devem ser seguidas para resolver um problema ou executar uma tarefa (Carboni, 2003).

1.1 Análise dos Componentes Principais - ACP

Entre os principais objetivos da ACP, pode-se destacar a redução da dimensionalidade dos dados e a obtenção de combinações interpretáveis do conjunto das variáveis originais, o que, por sua vez, possibilita descrever e compreender a estrutura de correlação dessas variáveis.

Formalmente, podemos afirmar que a análise de componentes principais transforma linearmente um conjunto de p variáveis correlacionadas em um conjunto de k variáveis latentes ortogonais (com $k < p$), que explicam uma parcela substancial das informações do conjunto original. Essa abordagem possibilita a geração, seleção e interpretação dos componentes investigados, e, também, auxilia na identificação das variáveis de maior influência na formação de cada componente[9].

Portanto, a ACP procura explicar a estrutura da variância e covariância de um conjunto de variáveis com o mínimo de perda de informação. Antes de se esquadrinhar o método de ACP, apresenta-se uma breve revisão de alguns conceitos em estatística e álgebra matricial.

1.1.1 Álgebra linear e estatística aplicada à análise multivariada

Primeiramente, cabe salientar o procedimento *pari passu* a fim de preparar o leitor para a sequência do processo de estimação dos componentes principais. Sendo assim, o algoritmo se baseia na matriz de variâncias-covariâncias, ou, na matriz de correlações. Se a análise utilizar a matriz de variância e covariância é preciso que as variáveis sejam padronizadas, mas se for utilizada a matriz de correlação não é necessário recorrer a padronização para se alcançar as mesmas estimativas[10].

Dessas matrizes, se extraem os autovalores e, os respectivos, autovetores. Essa matriz de autovetores irá multiplicar a matriz original de dados. Em suma, o que o método

[9] Cabe destacar ainda uma propriedade da ACP, uma vez que o conjunto original possui distribuição normal multivariada, essa característica será preservada nos componentes estimados que, além de multivariados, serão independentes entre si. O vetor $\vec{X}_{p\times 1} = (X_1, \cdots, X_p)$ é um vetor de variáveis aleatórias com distribuição normal multivariada (p-variada) se sua função de densidade de probabilidade for dada por $f(\vec{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right\}$, com $-\infty < x_i < \infty$, $\forall\, i = 1,2,\cdots,p$ e Σ uma matriz positiva definida - $\Sigma > 0$. A notação convencional para uma distribuição normal multivariada é: $\vec{X}_p \sim N(\vec{\mu}, \Sigma)$.

[10] Se o caso for de uma padronização de média zero e variância constante (1), temos: $Z_{il} = \frac{X_{il}-\overline{X}_l}{s(X_l)}$, tal que i é a i-ésima linha da l-ésima coluna. A média amostral, por sua vez, pode ser encontrada por: $\overline{x}_l = \frac{1}{n}\sum_{i=1}^{n} x_{il}$, $\forall\, l \in \{1,2,\ldots,p\}$;

.

traduz, nada mais é que a criação de um conjunto de novas variáveis latentes ortogonais, ou seja, não correlacionadas entre si – linearmente independentes –, obtidas de combinações lineares das variáveis iniciais e apresentadas em ordem decrescente relativas ao seu poder de explicação.

A ACP conta com um conjunto de " n" indivíduos, constituindo o número e observações e " p" variáveis observadas. Assim, a matriz de dados tem a dimensão $n \times p$ condicionada a $n > p$, isto é, para aplicação da ACP é necessário que o número de variáveis seja menor que o de observações. Algebricamente pode-se representar a matriz $\mathbf{X}_{n\times p}$, com entradas x_{ij} para i-ésimo item da linha e j-ésima variável da coluna, como:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [x_{ij}],\ \forall i \in \{1, \ldots, n\},\ \forall j \in \{1, \ldots, p\} \tag{1.1}$$

As variáveis da matriz $\mathbf{X}$ podem apresentar escalas diferentes entre si, e essa é uma realidade dos dados utilizados na presente pesquisa[11], é necessário e suficiente a normalização do conjunto de dados $\mathbf{X}_{n\times p}$. Assim, quando a opção for pela utilização da matriz de variância e covariância para se calcular os autovalores e autovetores, será necessário a padronização das variáveis de $\mathbf{X}_{n\times p}$.

Dessa forma, começa-se com a estimação da matriz variâncias-covariâncias amostrais $\mathbf{S}$[12], tal que a covariância entre a k-ésima e a l-ésima variáveis é determinada pela expressão $s_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)(x_{il} - \overline{x}_l)$, $k \neq l$ e, por sua vez, a variância amostral da k-ésima variável é estimada por $s_{kk}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ik} - \overline{x}_k\right)^2$,. Portanto, observe que a matriz $\mathbf{S}$ é uma matriz simétrica, formada por entradas dadas pela primeira fórmula, exceto ao longo das entradas da diagonal principal, que é definida pela segunda. Formalmente, pode-se es-

[11] Os dados de Censo têm variáveis como o número de tratores, o efetivo bovino, o valor das receitas com agricultura, ainda, utilizamos aqui a distância, etc. Essas diferentes escalas de medidas, se não tratadas, podem prejudicar o resultado das medidas de correlação, isto é, uma determinada característica poderá assumir maior importância no componente principal que outras apenas pelo fato de ter uma maior escala de medida.

[12] Do ponto de vista populacional, podemos definir $\mathbf{\Sigma}$ como sendo a matriz populacional

crever:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \Longrightarrow \mathbf{\Sigma} = Cov(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \tag{1.2}$$

No entanto, se a opção for calcular os componentes principais usando a matriz de correlação amostral, as variáveis não precisam ser padronizadas, conforme já afirmado. A matriz de correlação pode ser representada por:

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{ip} \\ r_{12} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ r_{ip} & \cdots & \cdots & 1 \end{bmatrix}_{p \times p} . \tag{1.3}$$

Tal que a correlação amostral é dada por: $r_{il} = \frac{s_{il}}{s_i s_l}$, $\forall\, i,l \in \{1,2,\ldots,p\}$, tal que o coeficiente de correlação:

- varia no intervalo $-1 \leq r_{il} \leq 1$;
- descreve uma relação linear;
- não sofre alteração com a transformação linear das variáveis – propriedades do conjunto de matrizes;
- pode ser do tipo de Pearson, Spearman's (não linear) ou Kendall's (relação ordinal).

Como já salientado, a transformação linear parte da decomposição dos autovalores associados, isto é, dos valores característicos das matrizes (**R**) ou (**S**). Da matriz de variâncias-covariâncias amostrais[13] $\mathbf{S}_{(p\times p)} = \frac{1}{n-1}(\mathbf{X}^T\mathbf{X} - \vec{\bar{x}}\vec{\bar{x}}^T)$ são extraídos os λ_p autovalores (*eigenvalues*) para se estimar os $\vec{e}_p$ autovetores que dão origem a uma matriz ortogonal (**P**), na qual cada coluna é um autovetor, isto é, **P** será uma matriz quadrada com p colunas responsáveis pela rotação do sistema original das variáveis.

Um autovalor é uma solução algébrica para proporção da variância explicada dada por cada uma das colunas de **P**. A soma de todos os autovalores representa o total da variância do conjunto. Portanto, eles serão a medida da importância relativa de cada componente selecionado, logo eles são o cerne da análise, e precisa de uma maior formalidade em sua apresentação.

[13] A matriz de variâncias-covariâncias amostrais pode também ser escrita como: $\mathbf{S}_{(p\times p)} = \frac{1}{n-1}(\mathbf{X}^T\mathbf{X} - \frac{1}{n-1}\mathbf{X}^T\vec{1}\vec{1}^T\mathbf{X}$, tal que a $\vec{1}$ é um vetor coluna de uns.

A matriz $\mathbf{A} \in \mathbf{M}_{n\times n}(\mathbb{R})$ possui um autovalor se, e somente se, existe $\underbrace{\vec{e}}_{(n\times 1)}$ não nulo tal que $\mathbf{A}\vec{e} = \lambda\vec{e}$ onde $\lambda \in \mathbb{C}$. Nesse caso, dizemos que λ é um autovalor associado ao autovetor $\vec{e}$[14]. Pode-se definir esse conceito de maneira mais generalizada, mas se restringe ao caso das matrizes reais, que é de particular interesse do presente trabalho. Em alemão, o prefixo "eigen"[15] pode ser traduzido como "específico" ou "próprio", isto é, "eigen" é conhecido por "característico", então um autovetor descreve uma propriedade característica de **A**.

Para se extrair um autovalor assume-se a existência de um autovetor e procuramos pelas raízes características λ. Isso é feito, com o cálculo do determinante da matriz $(T-\lambda I)$. Então, se esse autovetor existir, haverá um vetor não nulo tal que $(T-\lambda I)\vec{e} = 0$, i.e. $(T-\lambda I)$ é singular e, consequentemente, tem determinante nulo. Assim, se reduz, portanto, o problema de encontrar autovalores de uma matriz ao de encontrar raízes de um certo polinômio em λ, denominado polinômio característico. Mais explicitamente,

$$\begin{aligned} \mathbf{A}\vec{e} = \lambda\vec{e} &\iff \mathbf{A}\vec{e} - \lambda\vec{e} = 0 \\ &\iff (\mathbf{A} - \lambda\mathbf{I})\vec{e} = 0 \\ &\iff \det(\mathbf{A} - \lambda\mathbf{I}) = 0 \end{aligned} \tag{1.4}$$

tal que: $\mathbf{A} = \lambda_1\vec{e}_1\vec{e}_1^T + \lambda_2\vec{e}_2\vec{e}_2^T + \cdots + \lambda_k\vec{e}_k\vec{e}_k^T$[16]. Geometricamente, a matriz **A** deforma o espaço e os autovetores são os únicos vetores que mudam somente por um fator λ. Eles podem ser alongados, contraídos, apontar para a direção contrária ou mesmo rotacionar, caso um dos autovalores for complexo, mas não mais do que isso. Se um autovalor for $\lambda = 1$, por exemplo, a matriz **A** devolveria o próprio autovetor associado a 1, i.e. os outros vetores (possivelmente com exceção dos autovetores associados a outros autovalores) ao seu redor seriam distorcidos e $\vec{e}_1$ ficaria inalterado. Deixa-se dito, por fim, que os autovetores são linearmente independentes, ou seja, perpendiculares entre si. Isso nos fornecerá uma importante informação sobre o padrão dos dados mais a frente.

[14] Normalmente, a notação e é utilizada para indicar a base canônica, contudo essa notação aqui, e doravante, irá denotar os autovetores associados aos respectivos autovalores

[15] A soma dos autovalores de uma matriz $\mathbf{A}$ é igual ao traço da matriz e o produto dessas raízes características é igual ao determinante da matriz $\mathbf{A}$.

[16] Conhecida como Decomposição Espectral, que pode ser escrita também como: $\sum_{i=1}^{k}\lambda_i\vec{e}_i\vec{e}_i^T = \mathbf{P\Lambda P^{-1}}$, tal que $\mathbf{\Lambda}$ é a matriz diagonal dos autovalores. Lembrando que uma matriz diagonal é uma matriz quadrada que possui todos os elementos acima e abaixo da diagonal principal nulos. Ainda, vale ressaltar para o leitor que existe uma segunda maneira para estimar os componentes, conhecida como Decomposição Singular, tal que a matriz $\mathbf{A} = \mathbf{U\Sigma V}^T$, em que $\mathbf{U} = \{u_1, u_2, \ldots, u_{k-1}\} = \left[\frac{1}{\sqrt{\lambda_i}}\right]\mathbf{A}\vec{e}_i$; $\mathbf{\Sigma} = \left\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_k}\right\}\mathbf{I}_{n\times k}$; $\mathbf{V} = \left\{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k\right\} = [\Lambda_1, \Lambda_2, \ldots, \Lambda_k]$.

Exemplo 1: Autovetores e Autovalores

1. Encontre os autovalores e autovetores associados a $\mathbf{A} = \begin{bmatrix} 0,8 & 0,3 \\ 0,2 & 0,7 \end{bmatrix}$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \iff \det\left(\begin{vmatrix} 0,8-\lambda & 0,3 \\ 0,2 & 0,7-\lambda \end{vmatrix}\right) = 0,56 - 1,5\lambda + \lambda^2 - 0,06 = 0$$

$$\iff \lambda^2 - 1,5\lambda + 0,5 = 0$$

As raízes características são $\lambda_1 = 1$ e $\lambda_2 = \frac{1}{2}$. Que-se, agora, encontrar o vetor que torna a igualdade $\mathbf{A}\vec{e}_i = \lambda_i\vec{e}_i$ verdadeira. Para o primeiro autovalor faz-se:

$$\begin{bmatrix} 0,8 & 0,3 \\ 0,2 & 0,7 \end{bmatrix}\vec{e}_1 = \lambda_1\vec{e}_1 \iff \begin{bmatrix} 0,8-1 & 0,3 \\ 0,2 & 0,7-1 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Isso nos fornece o seguinte sistema:

$$\begin{cases} -0,2a + 0,3b = 0 \\ 0,2a - 0,3b = 0 \end{cases}$$

Resolvendo para a, encontra-se $a = 1,5b$. Isso significa que existem infinitas soluções para o sistema. Pode-se escolher somente uma. Então, para $b = 2$, $a = 3$. Logo, o autovetor associado ao autovalor λ_1 é $\vec{e}_1^T = [3,\ 2]$. Para o segundo autovalor:

$$\begin{bmatrix} 0,8 & 0,3 \\ 0,2 & 0,7 \end{bmatrix}\vec{e}_2 = \lambda_2\vec{e}_2 \iff \begin{bmatrix} 0,8-0,5 & 0,3 \\ 0,2 & 0,7-0,5 \end{bmatrix}\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Tem-se o sistema:

$$\begin{cases} 0,3c + 0,3d = 0 \\ 0,2c + 0,2d = 0 \end{cases}$$

Resolvendo para c, encontramos $c = -d$. Ao escolher $d = -1$ e se obtém $c = 1$.

Portanto,

$$\vec{e}_2^T = [1, -1]$$

Enfim, busca-se os autovetores que tornassem a igualdade $\mathbf{A}\vec{e}_i = \lambda_i \vec{e}_i$ verdadeira. Procedendo um teste simples

Para $\vec{e}_1$:

$$\begin{bmatrix} 0,8 & 0,3 \\ 0,2 & 0,7 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} - 1 \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} + \begin{bmatrix} -3 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \checkmark$$

Para $\vec{e}_2$:

$$\begin{bmatrix} 0,8 & 0,3 \\ 0,2 & 0,7 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0,5 \\ -0,5 \end{bmatrix} + \begin{bmatrix} -0,5 \\ 0,5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \checkmark$$

Finalmente, a Decomposição Espectral $\mathbf{A} = \mathbf{PDP}^{-1}$ pode ser encontrada, tal que $\mathbf{A}$ é diagonalizável, isto é

$$\begin{bmatrix} 0,8 & 0,3 \\ 0,2 & 0,7 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0,5 \end{bmatrix} \begin{bmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{-3}{5} \end{bmatrix}$$

Exemplo 2: Autovetores e Autovalores

1. Encontre os autovalores e autovetores associados a $\mathbf{A}$ de correlações, tal que $\mathbf{\Sigma_A} = \begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix}$.

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \iff \det\left(\begin{vmatrix} 1-\lambda & -5 \\ -5 & 1-\lambda \end{vmatrix}\right) = (1-\lambda)(1-\lambda) - 25 = 0$$

$$\iff \lambda^2 - 2\lambda - 24 = 0$$

As raízes características são $\lambda_1 = 6$ e $\lambda_2 = -4$. Encontrando, primeiramente, o autovetor associado ao autovalor λ_1. Tem-se

$$\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix} \vec{e}_1 = \lambda_1 \vec{e}_1 \iff \begin{bmatrix} 1-6 & -5 \\ -5 & 1-6 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Isso permite estabelecer o seguinte sistema:

$$\begin{cases} -5e_{11} - 5e_{21} = 0 \\ -5e_{11} - 5e_{21} = 0 \end{cases}$$

Resolvendo para e_{11}, encontra-se $e_{11} = -e_{21}$. Escolhendo arbitrariamente $e_{21} = -1$, obtém-se $e_{11} = 1$. Logo $\vec{e}_1^T = [1, -1]$.

Para o segundo autovetor:

$$\begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix} \vec{e}_2 = \lambda_2 \vec{e}_2 \Longleftrightarrow \begin{bmatrix} 1-(-4) & -5 \\ -5 & 1-(-4) \end{bmatrix} \begin{bmatrix} e_{12} \\ e_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Isso dá origem ao sistema:

$$\begin{cases} 5e_{12} - 5e_{22} = 0 \\ -5e_{12} + 5e_{22} = 0 \end{cases}$$

Resolvendo para e_{12}, tem-se $e_{12} = e_{22}$. Pode-se, então, simplesmente escolher $e_{12} = e_{22} = 1$, i.e. $\vec{e}_2^T = [1, 1]$. Os autovetores, normalizados[a], são:

$$e_1^T = \left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right] \quad \text{e} \quad e_2^T = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$$

Busca-se autovetores que tornassem a igualdade $\mathbf{A} = \lambda_1 \vec{e}_1 \vec{e}_1^T + \lambda_2 \vec{e}_2 \vec{e}_2^T$ verdadeira. Verifica-se que este é o caso.

$$\begin{aligned} 6 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} + (-4) \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} &= 6 \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} + (-4) \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} 3 & -3 \\ -3 & 3 \end{bmatrix} + \begin{bmatrix} -2 & -2 \\ -2 & -2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -5 \\ -5 & 1 \end{bmatrix} \end{aligned}$$

[a] Considerando que o vetor $\vec{e}$ pode ser esticado/aumentado com a multiplicação de uma constante qualquer, é conveniente aplicar a restrição de que esses vetores sejam unitários (JOHNSON and WICHERN, 1992). Nesse sentido, cada elemento do vetor deve ser dividido pela norma desse.

A norma, ou comprimento de um vetor com m elementos é definida pela expressão Pitagórica: $|\vec{x}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_m^2}$.

Na figura 1, em azul e vermelho e traçados com linha sólida estão os vetores da coluna de nossa matriz original do exemplo 2[17], ao passo que o os vetores traçados em pontilhado traduzem a nova matriz, rotacionada pela matriz de autovetores. Como é possível notar, além da nova direção dos eixos, existe uma perpendicularidade dos vetores tal como mencionamos. Esse novo conjunto de vetores representam o vetor de *scores* de cada observação, isto é, são as coordenadas das observações sobre os componentes principais.

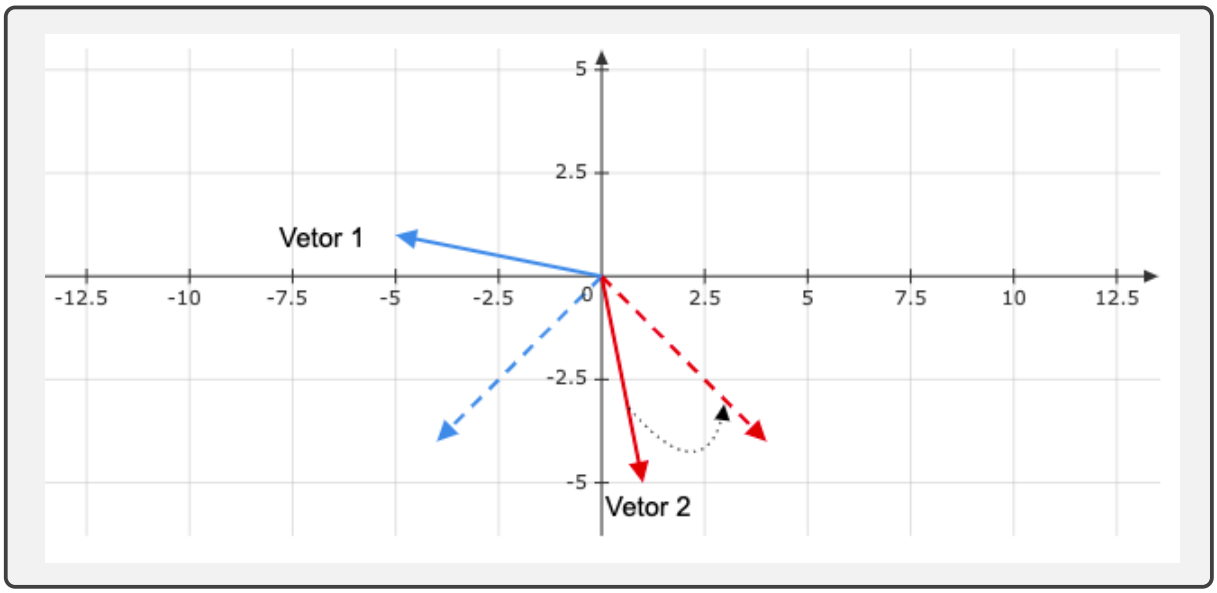


Figura 1: Ilustração dos Autovetores do Exemplo 2

Finalmente, para algumas técnicas de análise multivariada, especialmente, em análise de *cluster*, o conceito de distância entre dois pontos é de considerável importância. Assim, se temos valores observados sobre dois elementos distintos em uma amostra, como poderíamos mensurar o grau de semelhança entre eles?

Entre as diversas medidas existentes, a mais comum é a distância Euclideana, que é definida por

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(x_{ik} - x_{jk}\right)^2}, \tag{1.5}$$

[17] A matriz $\mathbf{A}$ do Exemplo 2 é simétrica e, por isso, os autovalores são sempre reais e distintos. Existe uma fórmula para matrizes 2×2 para encontrar os autovalores e, com ela podemos provar esse resultado no caso 2×2. A fórmula é $\lambda_{1,2} = \frac{\text{Tr}(\mathbf{A}) \pm \sqrt{\text{Tr}^2(\mathbf{A}) - 4\det(\mathbf{A})}}{2}$, em que $\text{Tr}(\mathbf{A})$ é o traço da matriz $\mathbf{A}$. É só perceber que sempre teremos $\text{Tr}^2(\mathbf{A}) - 4\det(\mathbf{A}) > 0$ e isso implica em autovalores reais e distintos. Seja $\mathbf{A}$ uma matriz $n \times n$. O traço de $\mathbf{A}$ é soma dos elementos da diagonal principal, $\text{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

em que x_{ik} e x_{jk}, $k = 1, \ldots, q$ são o s valores para as observações i e j, respectivamente[18]. Vale ressaltar que quando as variáveis são medidas com grandezas diferentes entre si, o ideal para operacionalizar esse cálculo será a padronização do conjunto de variáveis.

1.1.2 Análise dos Componentes Principais

Sabendo sobre a extração dos autovalores e seus autovetores associados, passa-se agora para o entendimento do que existe por trás da ACP. Sendo assim, iremos toma-se esses autovetores como um novo conjunto, que será arranjado em ordem decrescente em relação aos autovalores[19], irão compor as colunas da matriz $\mathbf{W}$. Esse conjunto ortogonal irá rotacionar a matriz original de variáveis tal que o novo conjunto será agrupado em $\mathbf{Y}_{p \times p}$, matriz que deverá conter a relação das variáveis com os p componentes principais estimados. Na prática, o primeiro componente será igual ao autovetor estimado pelo maior dos autovalores da matriz **S**. Na forma matricial, os autovetores são agrupados em uma matriz **W**, ordenada pelas raízes características λ e organizada em ordem decrescente para todo $\lambda_1, \lambda_2, \ldots, \lambda_p \geq 0$, e $(\lambda_1 \vec{e}_1), (\lambda_2 \vec{e}_2), \ldots, (\lambda_p \vec{e}_p)$ pares ordenados de autovalores-autovetores de $\mathbf{\Sigma}$.

Assim, a ACP consiste em encontrar os $\vec{e}_p$ vetores que maximizam o lado direito da expressão (1.7), sujeito às restrições de **W** e **X**. Portanto, o processo começa com o primeiro componente (aquele com maior variância) e segue até o p-ésimo e último componente (com a menor proporção de variância do conjunto original). A variabilidade total dos "p" componentes extraídos será igual à variabilidade total das "p" variáveis originais. Sendo assim, o processo consiste em passar de um conjunto de "p" variáveis para um conjunto alternativo. Esse novo conjunto de coordenadas para os indivíduos (observações) reunido nas linhas de **Y** será construído com base nos autovetores estimados.

A transformação linear do vetor $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ de variáveis correlacionadas, que possui matriz de variâncias-covariâncias, será transformado em novas variáveis não-correlacionadas $Y_1, Y_2, \ldots, Y_p$. As coordenadas dessas novas variáveis são descritas pelos

[18] No *R*, a distância Euclideana pode ser calculada com o comando *dist()*.

[19] Por definição, os autovalores podem ser complexos caso a matriz possua entradas reais, mas não se pode ordenar números complexos, pois não existe ordem nos números complexos. No entanto, se a matriz $\sum$ for auto-adjunta, os autovalores seriam sempre reais e faria sentido ordená-los.

vetores característicos $\vec{e}_j$ de $\mathbf{W}_{p\times p}$ usados na seguinte transformação:

$$\underbrace{\mathbf{Y}}_{(p\times p)} = \underbrace{\mathbf{W}^T}_{(p\times p)}\underbrace{\mathbf{X}}_{(p\times p)}, \tag{1.6}$$

desde que as condições do sistema linear (1.7) sejam satisfeitas:

$$\begin{aligned}
\vec{Y}_1 &= \vec{e}_1^T\vec{\mathbf{X}} = e_{11}X_1 + e_{21}X_2 + \cdots + e_{p1}X_p = \sum_{j=1}^{p} e_{j1}x_j \\
\vec{Y}_2 &= \vec{e}_2^T\vec{\mathbf{X}} = e_{12}X_1 + e_{22}X_2 + \cdots + e_{p2}X_p = \quad \vdots \\
&\vdots \\
\vec{Y}_p &= \vec{e}_p^T\vec{\mathbf{X}} = e_{1p}X_1 + e_{2p}X_2 + \cdots + e_{pp}X_p = \sum_{j=1}^{p} e_{jj}x_j
\end{aligned} \tag{1.7}$$

Ao considerar que $\mathbf{W}$ é uma matriz ortogonal $\mathbf{W}_{p\times p} = (\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_p)$, então pode-se afirmar que $\vec{Y}_1 = \vec{e}_1^T\vec{\mathbf{X}}$, para $\vec{\mathbf{X}}^T = [X_1, X_2; \ldots, X_p]$, é o primeiro vetor de componentes principais extraído de $\mathbf{\Sigma}$. As colunas de $\mathbf{Y}$ contêm a relação dos autovetores associados aos componentes, cada elemento e representa o peso, isto é, quanto cada variável contribui para o componente correspondente. E, do ponto de vista de variáveis normalizadas, podemos escrever essa nova variável Z como o componente principal de $\mathbf{X}$. Então o j-ésimo componente principal será

$$\vec{Z}_j = \vec{e}_j^T\left[\vec{X} - \vec{\overline{X}}\right], \tag{1.8}$$

tal que $\vec{X}_j$ e $\vec{\overline{X}}$ são vetores com dimensão $p\times 1$ que contêm as observações medidas nas variáveis originais e suas médias, respectivamente. Essa etapa do processo de estimação pode ser ilustrada com os resultados do Exemplo 2, discutido anteriormente. De posse dos autovetores do Exemplo 2, se escreve cada componente de (1.6), no caso do exemplo, apenas dois componentes são possíveis, assim:

$$\begin{aligned}
Y_1 &= \vec{e}_1^T\mathbf{X} = \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2 \\
Y_2 &= \vec{e}_2^T\mathbf{X} = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2
\end{aligned}$$

Para derivarmos essa forma apresentada dos Componentes Principais, na equação (1.7), considere primeiro que $\vec{e}_1^T\vec{\mathbf{X}}$; tal que $\vec{e}_1$ será o vetor responsável por maximizar $var\left[\vec{e}_1^T\vec{\mathbf{X}}\right] = \vec{e}_1^T\mathbf{\Sigma}\vec{e}_1$. Esse modelo assume que os autovetores são normais, tal que $\vec{e}_1^T\vec{e}_1 = 1$, isto é, a soma

do quadro dos elementos de $\vec{e}_1$ será igual a 1.

Para maximizar $\vec{e}_1^T \mathbf{\Sigma} \vec{e}_1$, sujeito a $\vec{e}_1^T \vec{e}_1 = 1$, recorremos a abordagem dos multiplicadores de Lagrange. Assim, maximiza-se

$$\vec{e}_1^T \mathbf{\Sigma} \vec{e}_1 - \lambda \left(\vec{e}_1^T \vec{e}_1 - 1 \right),$$

em que λ é o multiplicador de Lagrange. Nesse sentido, diferenciamos em relação a $\vec{e}_1$ para se obter

$$\mathbf{\Sigma} \vec{e}_1 - \lambda \vec{e}_1 = 0,$$

$$\left(\mathbf{\Sigma} - \lambda \mathbf{I} \right) \vec{e}_1 = 0,$$

no qual $\mathbf{I}_p$ é a matriz identidade $\left(p \times p \right)$. Então, λ é um autovalor de $\mathbf{\Sigma}$ e $\vec{e}_1$ corresponde ao autovetor. Para decidir qual dos p autovetores leva a $\vec{e}_1^T$ com maior parcela de variância, observe que a quantidade a ser maximizada é

$$\vec{e}_1^T \mathbf{\Sigma} \vec{e}_1 \; = \; \vec{e}_1^T \lambda \vec{e}_1 \; = \; \lambda \vec{e}_1^T \vec{e}_1 \; = \; \lambda$$

logo, λ deve ser o maior possível. Assim, $\vec{e}_1$ é o autovetor correspondente ao maior autovalor de $\mathbf{\Sigma}$, e $var\left(\vec{e}_1^T \right) = \vec{e}_1^T \mathbf{\Sigma} \vec{e}_1 = \lambda_1$ ao maior autovalor.

Mas o leitor deve lembrar que o objetivo principal da técnica é reduzir a dimensão dos dados para um número menor de componentes principais, ou seja, condensa-se as p variáveis em k componentes, para $k \leq p$. Assim, o k-ésimo Componente Principal é $\vec{e}_k^T \mathbf{X}$ e $var\left(\vec{e}_k^T \right) = \lambda_k$, no qual λ_k é o k-ésimo maior autovalor de $\mathbf{\Sigma}$, e $\vec{e}_k$ é o autovetor correspondente. De maneira similar, pode-se mostrar para $k = 2$, mas para $k \geq 3$ seria mais rebuscado, uma vez que essa demonstração ultrapassa o escopo desse documento[20].

[20] Para mais detalhes ver (JOLLIFFE, 2002).

Exemplo 3: Componentes Principais de dados padronizados

1. Exemplo 8.5 - (JOHNSON and WICHERN, 1992) - Estes são os dados de 100 semanas sucessivas da taxa de retorno das ações de cinco empresas (Allied Chemical, du Pont, Union Carbide, Exxon e Texaco), denotadas por x_1, x_2, ..., x_5. Então

$$\bar{\mathbf{x}}^T = [,0054\,,0048\,,0057\,,0063\,,0037]$$

e

$$\mathbf{R} = \begin{bmatrix} 1,000 & ,577 & ,509 & ,387 & ,462 \\ ,577 & 1,000 & ,599 & ,389 & ,322 \\ ,509 & ,599 & 1,000 & ,436 & ,426 \\ ,387 & ,389 & ,436 & 1,000 & ,523 \\ ,462 & ,322 & ,426 & ,523 & 1,000 \end{bmatrix}$$

Observe que **R** é a matriz de variâncias-covariâncias dos dados padronizados

$$z_{j1} = \frac{x_{j1} - \bar{x}_1}{\sqrt{s_{11}}},\ z_{j2} = \frac{x_{j2} - \bar{x}_2}{\sqrt{s_{22}}},\ \ldots,\ z_{j5} = \frac{x_{j5} - \bar{x}_5}{\sqrt{s_{55}}},\ \forall\ j = 1, \ldots, n.$$

Os autovalores e autovetores normalizados são

$$\hat{\lambda}_1 = 2,587,\quad \hat{\vec{e}}_1^T = [,464\ ,457\ \ ,470\ \ ,421\ \ ,421]$$

$$\hat{\lambda}_2 = ,809,\quad \hat{\vec{e}}_2^T = [,240\ \ ,509\ \ ,260\ \ -,526\ \ -,582]$$

$$\hat{\lambda}_3 = ,540,\quad \hat{\vec{e}}_3^T = [-,612\ \ ,178\ \ ,335\ \ ,541\ \ -,435]$$

$$\hat{\lambda}_4 = ,452,\quad \hat{\vec{e}}_4^T = [,387\ \ ,206\ \ -,662\ \ ,472\ \ -,382]$$

$$\hat{\lambda}_5 = ,343,\ \hat{\vec{e}}_5^T = [-,451\ \ ,676\ \ -,400\ \ -,176\ \ ,385].$$

Usando as variáveis padronizadas, opta-se por dois componentes da amostra:

$$\hat{\vec{y}}_1 = \hat{\vec{e}}_1^T\mathbf{Z} = ,464z_{11} + ,457z_{21} + ,470z_{31} + ,421z_{41} + ,421z_{51}$$

$$\hat{\vec{y}}_2 = \hat{\vec{e}}_2^T\mathbf{Z} = ,240z_{12} + ,509z_{22} + ,260z_{32} - ,526z_{42} - ,582z_{52}$$

Esses dois componentes são responsáveis por

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p}\right)100\% = \left(\frac{2{,}857 + {,}809}{5}\right)100\% = 73\%$$

da variância total dos dados. Os dois vetores $\vec{y}$ apresentam a relação das empresas com os componentes, e cada elemento do vetor é um *índice*. Como se vê no primeiro componente, as empresas, em sua totalidade, se relacionam positivamente com o CP1, logo esse componente pode ser "batizado"de componente geral do mercado de ações, ou mesmo, *componente do mercado*. Já no segundo componente, tem-se um contraste entre as empresas químicas (positiva) e de petróleo (negativamente correlacionadas). Portanto, podemos chamar o CP2 de *componente industrial*. Como o primeiro componente tem uma carga maior, pode-se dizer que a maior parte do movimento no retorno das ações é medido pelo componente do mercado (CP1).

Finalmente, pode-se obter a relação individual de cada observação com os autovetores. Mas muita atenção, pois o resultado final da rotação traz as observações exibidas nas colunas ao passo que cada linha será uma dimensão das observações. Portanto, após a transformação, a matriz **F** que contêm as novas projeções de cada entrada da matriz original, matricialmente, passa a ser representada por:

$$\begin{aligned} F_1 &= e_{11}x_{11} + e_{21}x_{12} + \cdots + e_{p1}x_{1n} \\ F_2 &= e_{12}x_{21} + e_{22}x_{22} + \cdots + e_{p2}x_{2n} \\ &\vdots \\ F_p &= e_{1p}x_{p1} + e_{2p}x_{p2} + \cdots + e_{pp}x_{pn} \end{aligned} \tag{1.9}$$

Mais ao final, se retornará ao assunto dos escores individuais uma vez que, dependendo de como são são apresentados, pode haver divergência na nomenclatura e forma de se encontrar tais escores. Essa matriz **F** de escores retornará ao centro das atenções, porém sujeita a algumas alterações na sua geração e, portanto, nomenclatura.

$$\underbrace{\mathbf{F}}_{(n\times p)} = \underbrace{\mathbf{X}}_{(n\times p)}\underbrace{\mathbf{W}}_{(p\times p)}. \tag{1.10}$$

Teorema 1 : Propriedades

$$Y_j = e_j^T \mathbf{X} = e_{1j}X_1 + e_{2j}X_2 + \cdots + e_{pj}X_p\,,\ \forall\ j = 1,2,\cdots,p.$$

Tal que

$$E(Y_j) = \vec{e}_j^T.$$

$$\text{Var}(Y_j) = \vec{e}_j^T \mathbf{\Sigma} \vec{e}_j = \lambda_j\,,\ \forall\ j = 1,2,\cdots,p.$$

$$\text{Cov}(Y_j, Y_k) = \vec{e}_j^T \mathbf{\Sigma} \vec{e}_k = 0\,,\ \forall\, j,k = 1,2,\cdots,p \text{ e } j \neq k.$$

A proporção da variância total de $\vec{x}$ que é explicada pela j-ésima componente principal é

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i},\ \forall\ j = 1,2,\cdots,p.$$

Como os autovalores estão ordenados em ordem decrescente, isso implica que o primeiro componente principal será a combinação linear de maior variância. Esse componente maximiza $Var(Y_1) = \vec{e}_1^T \mathbf{\Sigma} \vec{e}_1$, o que possibilita definir os componentes principais como

Y_1: combinação linear $\vec{e}_1^T \vec{\mathbf{X}}$ que maximiza $\begin{cases} \text{Máx.} & Var\left(\vec{e}_1^T \vec{\mathbf{X}}\right) \\ \text{s.a} & \vec{e}_1^T \vec{e}_1 = 1 \end{cases}$

Y_2: combinação linear $\vec{e}_2^T \vec{\mathbf{X}}$ que maximiza $\begin{cases} \text{Máx.} & Var\left(\vec{e}_2^T \vec{\mathbf{X}}\right) \\ \text{s.a} & \vec{e}_2^T \vec{e}_2 = 1\,,\ \vec{e}_1^T \vec{e}_2 = 0 \text{ e } Cov\left(\vec{e}_1^T \vec{\mathbf{X}}, \vec{e}_s^T \vec{\mathbf{X}}\right) \end{cases}$

p-ésimo CP: combinação linear $\vec{e}_p^T \vec{\mathbf{X}}$ que maximiza $\begin{cases} \text{Máx.} & Var\left(\vec{p}_2^T \vec{\mathbf{X}}\right) \\ \text{s.a} & \vec{e}_p^T \vec{e}_p = 1 \\ & \text{e } Cov\left(\vec{e}_i^T \vec{\mathbf{X}}, \vec{e}_p^T \vec{\mathbf{X}}\right) = 0 \text{ para } p < i. \end{cases}$

Esse é um problema de maximização semelhante a escolha da cesta que maximiza a utilidade de um consumidor em microeconomia e o método é o mesmo, isto é, a construção de uma nova função com a utilização do método de Lagrange. Considerando que para maximizar $Y_j = e_j^T \vec{x}$ temos $Var\left(Y_j\right) = \vec{e}_j^T\ \mathbf{\Sigma}\ \vec{e}_j$ sujeito à ${e_{1j}}^2 + {e_{2j}}^2 + \cdots + {e_{pj}}^2 = 1$ ou, para simplificar a notação matricial, podemos escrever $\vec{e}_j^T\ \vec{e}_j = 1$.

Resultado 1 Solução do Problema de Maximização em matrizes e vetores

Uma nova equação (L) deve ser construída:

$$L(\vec{e}, \mathbf{\Sigma}, \boldsymbol{\lambda}) = \vec{e}_1^T \, \mathbf{\Sigma} \, \vec{e}_1 + \boldsymbol{\lambda} \left(1 - \vec{e}_1^T \vec{e}_1\right). \tag{1.a}$$

Para o primeiro componente, temos então:

$$\frac{\partial L}{\partial e_1} = 0 \;\Rightarrow\; \vec{\nabla}\left(\vec{e}_1^T\right) \mathbf{\Sigma} \, \vec{e}_1 + \boldsymbol{\lambda}\left(1 - \vec{e}_1^T \vec{e}_1\right) \;\Leftrightarrow\; \vec{\nabla}\left(\vec{e}_1^T \, \mathbf{\Sigma} \, \vec{e}_1\right) - \vec{\nabla}\left(\boldsymbol{\lambda} \vec{e}_1^T \, \vec{e}_1\right). \tag{1.b}$$

Mas não se engane, esse é um problema de derivada de matrizes, que pelo "cook book"de operações com matrizes, no caso de segunda ordem, encontra-se uma receita para essa solução no caso particular:

$$\frac{\partial \left(\mathbf{B}\vec{x} + \vec{b}\right)^T \mathbf{C}\left(\mathbf{D}\vec{x} + \vec{d}\right)}{\partial \vec{x}} = \mathbf{B}^T \mathbf{C}\left(\mathbf{D}\vec{x} + \vec{d}\right) + \mathbf{D}^T \mathbf{C}^T \left(\mathbf{B}\vec{x} + \vec{b}\right).$$

Pode-se olhar (1.b) pelo termo da esquerda da expressão que queremos encontrar o gradiente. Consideramos que $\vec{b}$ e $\vec{d}$ são vetores nulos, e tomando as matrizes $\mathbf{B}$ e $\mathbf{D}$ como matrizes identidade, ficamos apenas com $\mathbf{C}$, que representa a nossa matriz simétrica $\mathbf{\Sigma}$ dos dados. O processo será o mesmo no segundo termo de (1.2), uma vez que $\boldsymbol{\lambda}$ é a matriz diagonal de autovalores, tal que

$$\begin{aligned}
&\Leftrightarrow \mathbf{I}^T \mathbf{S}\left(\mathbf{I}\vec{e}_1 + \vec{0}\right) + \mathbf{I}^T \mathbf{\Sigma}^T \left(\mathbf{I}\vec{e}_1 + \vec{0}\right) - \vec{\nabla}\left(\boldsymbol{\lambda} \vec{e}_1^T \, \vec{e}_1\right) \\
&\Leftrightarrow \mathbf{\Sigma}\left(\vec{e}_1\right) + \mathbf{\Sigma}^T \left(\vec{e}_1\right) - \vec{\nabla}\left(\boldsymbol{\lambda} \vec{e}_1^T \, \mathbf{I} \, \vec{e}_1\right) \\
&\Leftrightarrow 2\,\mathbf{\Sigma}\,\vec{e}_1 - 2\,\boldsymbol{\lambda}\,\vec{e}_1 = 0 \\
&\Leftrightarrow \mathbf{\Sigma}\,\vec{e}_1 - \boldsymbol{\lambda}\,\vec{e}_1 = 0 \\
&\Leftrightarrow \boxed{\mathbf{\Sigma}\,\vec{e}_1 = \boldsymbol{\lambda}\,\vec{e}_1} \qquad (1.c) \\
&\Leftrightarrow (\mathbf{\Sigma} - \boldsymbol{\lambda})\, e_1 = 0 \Leftrightarrow \boxed{(\mathbf{\Sigma} - \boldsymbol{\lambda}\,\mathbf{I})\, e_1 = 0} \qquad (1.d)
\end{aligned}$$

Logo, a solução de $(\mathbf{\Sigma} - \lambda\, I)\, e_1 = 0$ tem as p raízes características que satisfazem nosso problema de autovalores já discutido anteriormente. Contudo, tem-se $\lambda_1, \lambda_2, \cdots, \lambda_p$, e ao tomar o maior λ, isto é, apenas aquele que maximiza a variância do componente $\vec{Y}_1$. Assim, ao $max\,[\mathrm{Var}\,(Y_1)]$, sabendo que $\mathrm{Var}\,(Y_1) = \vec{e}_1^T \mathbf{\Sigma} \vec{e}_1$ e o resul-

tado de (1.c), pode-se afirmar que na situação em que a variância será máxima pode ser definida por $\text{Var}(Y_1) = \vec{e}_1^T \boldsymbol{\lambda} \vec{e}_1$, ou seja, a variação máxima será dada pelo maior autovalor $\text{Var}(Y_1) = \lambda$. Para segundo componente, o processo seria o mesmo, exceto pelo fato de se ter que adicionar a restrição da ortogonalidade (produto escalar 0) entre o PC1 e o PC2 ($\vec{e}_1^T \vec{e}_2 = 0$).

Assim, usando Lagrange encontra-se um argumento capaz de solucionar os componentes principais. A matriz populacional $\boldsymbol{\Sigma}$ é uma matriz simétrica de dimensão $p \times p$. Isto é, se está diante de um problema mais amplo em termos de dimensões, busca-se os gradientes $\vec{\nabla}$ do conjunto.

Finalmente, pode-se concluir que a ACP transformou os dados de entrada em termos dos padrões apresentados entre eles. Ainda, permitiu calcular os escores fatoriais, ou contribuição de cada observação para cada fator. Em termos de autovetores, o que foi realizado representa uma projeção dos dados pelos autovetores diferente dos eixos originais.

1.1.3 Exemplo aplicado no R: dados do Censo Agropecuário 1994/95

Da teoria à prática! Tem-se nossos dados em escala original, listados no arquivo COREDES.xls, com variáveis nas colunas e observações (municípios) dispostas nas linhas da planilha. O primeiro passo é a importação da planilha Excel para o R e, posteriormente, podemos verificar na parte superior do *tibble* criado com o nome de "df". Para a leitura de dados xls, csv ou txt, dois pacotes devem ser instalados e carregados, a saber: *readr* e *readxl*.

1: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```
df <- read_excel("COREDES.xls",
    #skip = 20,       # Pula as 20 primeiras linhas
    sheet = 1,       # Informar qual a posição na planilha, nesse caso, coincidiu ser
     ↪  a primeira
    col_names = TRUE  # TRUE informa que o dado tem cabeçalho
)#[, c(1:4, 8, 14)]   # Não foi possível filtrar dentro da função, então deve ser
 ↪  feito separado

head(df)   #Comando para visualizar o arquivo importado (parte superior)
```

```
    > head(df)
# A tibble: 6 x 21
  municipio  id   corede   gado   pib    pop   pea   asspec assagr trator adubos
   ↪  irriga   recveg      recani
  <chr>     <dbl> <chr>   <dbl>  <dbl>  <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
   ↪  <dbl>    <dbl>       <dbl>
1 Alegrete    1    FO     536536 1.20e8  1401  9067    854    537   1905   1736    477
   ↪  48204160    30871591
2 Itacurubi   2    FO      89072 1.39e7  2482  1774    165    103    274    220     74
   ↪  3012431     4865929
3 Itaqui      3    FO     295104 1.28e8  8175  4419    381    316   1635    511    291
   ↪  66259965    19505423
4 Manoel_V    4    FO      88597 1.68e7  1248  1492    381     54    296    253     45
   ↪  4121898     3882005
5 Quaraí      5    FO     197706 3.08e7  2371  3011    323    102    343    251     63
   ↪  8437765    10662991
6 Rosário_    6    FO     322948 5.88e7  5610  6521    432    300   1040    784    222
   ↪  21800270    17251636
# . . .  with 6 more variables: magcol <dbl>, valveg <dbl>, valani <dbl>, arexpl <dbl>,
   ↪  distrg <dbl>,
renda <dbl> financ<dbl>
```

Os próximos passos são as estimações das matrizes de correlação e variâncias-covariâncias amostrais. Contudo, um *data frame* foi criado com as colunas de interesse, isto é, se excluí as três primeiras, uma vez que elas contêm caracteres ou dados não relevantes. Ainda, transformando o *tibble* em um *data frame*, o qual tem a estrutura muito semelhante a uma planilha de Excel. Os dados para análise estão agrupados no *dataframe* de nome "data", bem como se constroi uma matriz chamada de "my.cov"e "R" que contém as medidas estatísticas de interesse.

2: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```
# Gerando a matriz de covariâncias dos dados em data
# Agora chamaremos de "data" o nosso dataframe que será criado pegando apenas as
   ↪  colunas de interesse (com valores numéricos)

data <- as.data.frame(df[4:21])   #no caso, são as colunas 4,5,...,21.

```

```r
my.cov = cov(data)                #salvando a matriz em "my.cov"
my.cov

# Vale lembrar que os resultados com matriz de covariâncias ou correlações irão produzir valores de Componentes Principais diferentes. Qual utilizar?
```

```
    > my.cov = cov(data)
> my.cov
                gado           pop           pea        asspec        assagr
                 trator        adubos
gado         7.652550e+09  7.782745e+07  8.903632e+07  5.839452e+06  1.913807e+06
 2.675184e+07  8.518162e+06
pop          7.782745e+07  7.197673e+06  4.801263e+06  2.849402e+05  3.554323e+05
 7.407777e+05  1.120672e+06
pea          8.903632e+07  4.801263e+06  4.151950e+06  2.877193e+05  2.901086e+05
 6.216654e+05  9.562224e+05
asspec       5.839452e+06  2.849402e+05  2.877193e+05  5.632198e+04  3.989577e+04
 6.460193e+04  8.226115e+04
assagr       1.913807e+06  3.554323e+05  2.901086e+05  3.989577e+04  6.876275e+04
 6.515630e+04  9.071310e+04
trator       2.675184e+07  7.407777e+05  6.216654e+05  6.460193e+04  6.515630e+04
 2.247412e+05  1.337579e+05
adubos       8.518162e+06  1.120672e+06  9.562224e+05  8.226115e+04  9.071310e+04
 1.337579e+05  2.730295e+05
irriga       7.572957e+06  1.694098e+05  1.298236e+05  6.721291e+03  8.112035e+03
 4.220262e+04  2.090284e+04
recveg        7.965774e+11  1.642525e+10  1.302451e+10  1.173535e+09  1.185261e+09
 5.108313e+09  1.911301e+09
recani        4.250704e+11  5.969004e+09  6.325921e+09  5.816922e+08  3.585348e+08
 1.788541e+09  9.688521e+08
financ       1.928350e+11  3.316855e+09  2.816363e+09  2.546102e+08  2.370325e+08
 1.118100e+09  3.639290e+08
magcol       7.263426e+06  2.149370e+05  1.775743e+05  2.037011e+04  2.076450e+04
 7.090042e+04  3.855689e+04
valveg        8.839912e+11  2.028095e+10  1.627282e+10  1.431345e+09  1.405470e+09
 5.803155e+09  2.639002e+09
```

```
valani        4.110167e+11  5.498475e+09  5.866751e+09  4.626666e+08  2.644452e+08
↪  1.711388e+09  8.048042e+08
arexpl       9.749279e+12  1.053966e+11  1.188824e+11  8.475518e+09  3.752698e+09
↪  3.752217e+10  1.239819e+10
distrg      -2.658996e+06 -2.481543e+04 -2.557957e+04 -3.447027e+03 -6.518251e+03
↪  -1.649298e+04 -4.587618e+03
renda        1.006578e+03 -2.075418e+00  1.463934e+00  6.806293e+00  3.053651e+00
↪  1.878057e+01  3.895806e+00
                  irriga     recveg        recani        financ        magcol
                  ↪  valveg       valani
gado         7.572957e+06  7.965774e+11  4.250704e+11  1.928350e+11  7.263426e+06
↪  8.839912e+11  4.110167e+11
pop          1.694098e+05  1.642525e+10  5.969004e+09  3.316855e+09  2.149370e+05
↪  2.028095e+10  5.498475e+09
pea          1.298236e+05  1.302451e+10  6.325921e+09  2.816363e+09  1.775743e+05
↪  1.627282e+10  5.866751e+09
asspec       6.721291e+03  1.173535e+09  5.816922e+08  2.546102e+08  2.037011e+04
↪  1.431345e+09  4.626666e+08
assagr       8.112035e+03  1.185261e+09  3.585348e+08  2.370325e+08  2.076450e+04
↪  1.405470e+09  2.644452e+08
trator       4.220262e+04  5.108313e+09  1.788541e+09  1.118100e+09  7.090042e+04
↪  5.803155e+09  1.711388e+09
adubos       2.090284e+04  1.911301e+09  9.688521e+08  3.639290e+08  3.855689e+04
↪  2.639002e+09  8.048042e+08
irriga       1.520284e+04  1.044449e+09  4.261482e+08  2.552529e+08  1.083543e+04
↪  1.214050e+09  4.269893e+08
recveg        1.044449e+09  1.584904e+14  4.948771e+13  3.386032e+13  1.673864e+09
↪  1.744945e+14  4.742269e+13
recani        4.261482e+08  4.948771e+13  2.741834e+13  1.158889e+13  4.973556e+08
↪  5.564970e+13  2.413845e+13
financ       2.552529e+08  3.386032e+13  1.158889e+13  8.482681e+12  3.499067e+08
↪  3.743290e+13  1.116643e+13
magcol       1.083543e+04  1.673864e+09  4.973556e+08  3.499067e+08  2.435116e+04
↪  1.877233e+09  4.851891e+08
valveg        1.214050e+09  1.744945e+14  5.564970e+13  3.743290e+13  1.877233e+09
↪  1.961896e+14  5.383890e+13
```

```
valani        4.269893e+08  4.742269e+13  2.413845e+13  1.116643e+13  4.851891e+08
↪  5.383890e+13  2.328957e+13
arexpl       9.777466e+09  1.104148e+15  5.511395e+14  2.605846e+14  1.072240e+10
↪  1.222714e+15  5.334506e+14
distrg      -5.327508e+03 -2.333935e+08 -1.258840e+08 -6.383165e+07 -4.335136e+03
↪  -2.789443e+08 -1.526888e+08
renda        1.622109e+00  3.285826e+05  8.967399e+04  6.534510e+04  6.761450e+00
↪  3.660559e+05  8.569806e+04
                    arexpl        distrg         renda
gado         9.749279e+12 -2.658996e+06  1.006578e+03
pop          1.053966e+11 -2.481543e+04 -2.075418e+00
pea          1.188824e+11 -2.557957e+04  1.463934e+00
asspec       8.475518e+09 -3.447027e+03  6.806293e+00
assagr       3.752698e+09 -6.518251e+03  3.053651e+00
trator       3.752217e+10 -1.649298e+04  1.878057e+01
adubos       1.239819e+10 -4.587618e+03  3.895806e+00
irriga       9.777466e+09 -5.327508e+03  1.622109e+00
recveg       1.104148e+15 -2.333935e+08  3.285826e+05
recani       5.511395e+14 -1.258840e+08  8.967399e+04
financ       2.605846e+14 -6.383165e+07  6.534510e+04
magcol       1.072240e+10 -4.335136e+03  6.761450e+00
valveg       1.222714e+15 -2.789443e+08  3.660559e+05
valani       5.334506e+14 -1.526888e+08  8.569806e+04
arexpl       1.270191e+16 -3.499169e+09  1.693937e+06
distrg      -3.499169e+09  1.019677e+04 -1.315396e+00
renda        1.693937e+06 -1.315396e+00  8.341034e-03
```

3: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```r
########################################################################
# "cor"comando para gerar a matriz de correlações entre as variáveis

R <- cor(data)  # Tanto faz, estimar pela standardized ou pela matriz original

print(R)
```

```
> R <- cor(data)
> print(R)
              gado    pop     pea     asspe   assagr  trator  adubo   irriga   recveg
gado         1.0000  0.3316  0.4995  0.2812  0.0834  0.6450  0.1863  0.7021   0.7233
pop          0.3316  1.0000  0.8782  0.4475  0.5052  0.5824  0.7994  0.5121   0.4863
pea          0.4995  0.8782  1.0000  0.5949  0.5429  0.6435  0.8981  0.5167   0.5077
asspec       0.2812  0.4475  0.5949  1.0000  0.6410  0.5742  0.6633  0.2296   0.3927
assagr       0.0834  0.5052  0.5429  0.6410  1.0000  0.5241  0.6620  0.2508   0.3590
trator       0.6450  0.5824  0.6435  0.5742  0.5241  1.0000  0.5399  0.7219   0.8559
adubos       0.1863  0.7994  0.8981  0.6633  0.6620  0.5399  1.0000  0.3244   0.2905
irriga       0.7021  0.5121  0.5167  0.2296  0.2508  0.7219  0.3244  1.0000   0.6728
recveg       0.7233  0.4863  0.5077  0.3927  0.3590  0.8559  0.2905  0.6728   1.0000
recani       0.9279  0.4248  0.5928  0.4680  0.2611  0.7205  0.3541  0.6600   0.7507
financ       0.7568  0.4244  0.4745  0.3683  0.3103  0.8097  0.2391  0.7107   0.9234
maqcol       0.5320  0.5133  0.5584  0.5500  0.5074  0.9584  0.4728  0.5631   0.8520
valveg       0.7214  0.5397  0.5701  0.4305  0.3826  0.8739  0.3605  0.7029   0.9895
valani       0.9735  0.4246  0.5966  0.4039  0.2089  0.7480  0.3191  0.7175   0.7805
arexpl       0.9888  0.3485  0.5176  0.3168  0.1269  0.7022  0.2105  0.7036   0.7782
distrg      -0.3010 -0.0915 -0.1243 -0.1438 -0.2461 -0.3445 -0.0869 -0.4278  -0.1835
renda        0.1259 -0.0084  0.0078  0.3140  0.1275  0.4337  0.0816  0.1440   0.2857
              recani  financ  magcol  valveg    valani  arexpl  distrg  renda
gado          0.9279  0.7568  0.5320   0.7214   0.9735  0.9888 -0.3010  0.1259
pop           0.4248  0.4244  0.5133   0.5397   0.4246  0.3485 -0.0915 -0.0084
pea           0.5928  0.4745  0.5584   0.5701   0.5966  0.5176 -0.1243  0.0078
asspec        0.4680  0.3683  0.5500   0.4305   0.4039  0.3168 -0.1438  0.3140
assagr        0.2611  0.3103  0.5074   0.3826   0.2089  0.1269 -0.2461  0.1275
trator        0.7205  0.8097  0.9584   0.8739   0.7480  0.7022 -0.3445  0.4337
adubos        0.3541  0.2391  0.4728   0.3605   0.3191  0.2105 -0.0869  0.0816
irriga        0.6600  0.7107  0.5631   0.7029   0.7175  0.7036 -0.4278  0.1440
recveg        0.7507  0.9234  0.8520   0.9895   0.7805  0.7782 -0.1835  0.2857
recani        1.0000  0.7598  0.6086   0.7587   0.9552  0.9339 -0.2380  0.1875
financ        0.7598  1.0000  0.7698   0.9175   0.7944  0.7938 -0.2170  0.2456
maqcol        0.6086  0.7698  1.0000   0.8588   0.6442  0.6096 -0.2751  0.4744
valveg        0.7587  0.9175  0.8588   1.0000   0.7964  0.7745 -0.1972  0.2861
valani        0.9552  0.7944  0.6442   0.7964   1.0000  0.9807 -0.3133  0.1944
arexpl        0.9339  0.7938  0.6096   0.7745   0.9807  1.0000 -0.3074  0.1645
distrg       -0.2380 -0.2170 -0.2751  -0.1972  -0.3133 -0.3074  1.0000 -0.1426
```

```
renda          0.1875  0.2456  0.4744   0.2861   0.1944  0.1645 -0.1426  1.0000
```

De fato, pode-se agora inferir sobre as relações existentes entre as variáveis e sua magnitude. Há uma ligação forte entre as receitas (recveg e recani) com o valor da produção animal (valveg) e o número de cabeças de gado (gado) e a área explorada (arexpl). Uma vez que se procura esse tipo de relação, a ACP poderá ser implementada. Esses resultados preliminares já indicam sinais que pode levar a formação de um componente que reunirá esse grupo de variáveis. Mas, para a análise, qual matriz deve ser utilizada? A análise de componentes principais pode ser aplicada em ambas, mas com diferentes resultados. Para definir, precisamos observar se as unidades de medida das variáveis são iguais. Em caso afirmativo, podemos escolher a matriz de variâncias-covariâncias. Agora, quando as variáveis medem grandezas diferentes, e como buscamos que ambas variáveis tenham o mesmo peso na análise, exceto por sua variabilidade, a matriz de correlação será a mais apropriada. Ainda, resta uma terceira via, caso as unidades de medida das variáveis difiram entre si, isto é, podemos fazer um reescala das mesmas em forma de índices ou $z - score$ (padronização).

To standardize or not to standardize? No nosso exemplo, faz-se necessário, primeiro, padronizar as variáveis, uma vez que as unidades de medida são diferentes. Pode-se fazer isso de várias maneiras no R, eis aqui uma sugestão: nesse caso, se cria uma função chamada de "*standardize*, "que irá tomar uma variável "x", calcular o desvio da média e dividí-lo pelo desvio-padrão dessa variável. Nesse sentido, se cria um novo *dataframe*, agora denominado *"df_st"*.

4: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```r
############################################################
# Standardize nos dados
############################################################
# Vamos criar uma função que pega a coluna e faz seu desvio da média e divide pelo
 ↪  desvio-padrão:

standardize <- function(x){
  return ((x - mean(x, na.rm = TRUE))/sd(x, na.rm = TRUE))
}

```

```r
df_st = as.data.frame(apply(df[4:21],2,standardize)) # Só nas colunas de interesse, 2 significa all lines

# "na.rm" significa literalmente NA remove, se TRUE,  a função irá desconsiderar qualquer valor NA (not available).
# Queremos somente as variáveis numéricas [colunas 4 até 21, totalizando 17 colunas]
```

```r
head(df_st)
> head(df_st)
   gado       pop       pea       asspec    assagr    trator    adubos     irriga     recveg    recani
1 5.671029 -1.101586  2.771553  2.3246450  0.709090  3.108934  1.7305931  3.4173011  3.1816790 5.16114144
2 0.555919 -0.698656 -0.807598 -0.5785769 -0.945966 -0.331494 -1.1707188  0.1488446 -0.4080158 0.19467718
3 2.911139  1.423342  0.490475  0.3315768 -0.133691  2.539397 -0.6138047  1.9087827  4.6158979 2.99047310
4 0.550489 -1.158615 -0.945994  0.3315768 -1.132828 -0.285087 -1.1075636 -0.0863544 -0.3198880 0.00677104
5 1.797750 -0.740030 -0.200521  0.0871836 -0.949780 -0.185945 -1.1113912  0.0596312  0.0229323 1.30177844
6 3.229433  0.467269  1.522064  0.5464742 -0.194707  1.284305 -0.0913389  1.3491711  1.0843503 2.56005327
      financ     maqcol    valveg    valani    arexpl     distrg       renda
1  3.8171367  2.3269363  3.07910498   5.3552328 5.5823447 -0.8536566  0.47555466
2 -0.3046018 -0.4606571 -0.42430068   0.3665802 0.3953186 -0.1901527 -1.00130796
3  4.6065028  2.4679180  4.53495519   3.1771902 2.7289456  1.4240434  1.21958706
4 -0.2359987 -0.2940423 -0.34186769   0.2267746 0.4903593 -0.4080196  0.23918387
5  0.2262276 -0.4414323 -0.07784675   1.3665896 1.8917074 -0.9526870  0.04722997
6  1.3147747  0.8914859  1.10882016   2.7408547 2.7462477 -1.8538639 -0.12345475
```

Como, geralmente, se desconhece a matriz $\boldsymbol{\Sigma}$, então se utiliza a matriz de variâncias e covariâncias amostrais $\mathbf{S}$ para estimar os componentes principais. Portanto, considere $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ os autovalores de $\mathbf{S}$, com autovetores correspondentes normalizados $\hat{\vec{e}}_1, \hat{\vec{e}}_2, \cdots, \hat{\vec{e}}_p$, em que a j-ésima componente principal amostral, conforme visto na equação (1.7), será dada por

$$\hat{\vec{Y}}_j = \hat{\vec{e}}_j^T \mathbf{X}.$$

Outrossim, considerando que o exemplo usado tem escalas diferentes, então aplica-se uma padronização (**Z**) dos dados, antes da análise de componentes principais para não afetar a análise . Essa análise é conhecida então como ACP pela matriz de correlações, pois

$$Cov(\mathbf{Z}) = \mathbf{R} = Cov(\mathbf{X}).$$

Somente assim, pode-se prosseguir na estimação dos autovalores e autovetores do nosso exemplo. Vale ressaltar que o quadro (5) traz a matriz de variâncias-covariâncias sobre as variáveis padronizadas, e, se o leitor verificar, irá perceber que ela é idêntica à matriz de correlações **R** que aparece no quadro (3).

5: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```
############################################################
# Variâncias-Covariâncias tomadas sobre a matriz Z
############################################################
my.cov = cov(df_st)
head(my.cov)
```

```
> head(my.cov)
        gado   pop    pea   asspec assagr trator adubos irriga recveg recani financ
gado   1.0000 0.3316 0.4995 0.2812 0.0834 0.6450 0.1863 0.7021 0.7233 0.9279 0.7568
pop    0.3316 1.0000 0.8782 0.4475 0.5052 0.5824 0.7994 0.5121 0.4863 0.4248 0.4244
pea    0.4995 0.8782 1.0000 0.5949 0.5429 0.6435 0.8981 0.5167 0.5077 0.5928 0.4745
asspec 0.2812 0.4475 0.5949 1.0000 0.6410 0.5742 0.6633 0.2296 0.3927 0.4680 0.3683
assagr 0.0834 0.5052 0.5429 0.6410 1.0000 0.5241 0.6620 0.2508 0.3590 0.2611 0.3103
trator 0.6450 0.5824 0.6435 0.5742 0.5241 1.0000 0.5399 0.7219 0.8559 0.7205 0.8097
       maqcol    valveg   valani   arexpl    distrg      renda
gado   0.532082 0.721451 0.973589 0.988860 -0.3010115  0.12598952
pop    0.513399 0.539702 0.424683 0.348574 -0.0915997 -0.00847031
pea    0.558462 0.570162 0.596610 0.517674 -0.1243185  0.00786657
asspec 0.550040 0.430593 0.403969 0.316878 -0.1438382  0.31402320
assagr 0.507440 0.382654 0.208967 0.126979 -0.2461632  0.12750664
trator 0.958401 0.873946 0.748042 0.702282 -0.3445295  0.43376803
```

Agora, já é possível conhecer os autovalores e autovetores da nossa matriz **S** de dados padronizados. Vale ressaltar que o objetivo aqui é o exercício para demonstrar cada

passo da obtenção dos componentes. Os pacotes do R, quando solicitados para extrair os componentes, farão diretamente sobre a matriz original de dados **X**, bem como automaticamente será considerado a matriz de variâncias-covariâncias dos dados normalizados (**Z**). De posse da matriz de interesse, pode-se extrair os autovalores, que no presente caso são 17. O comando será "eigen", exatamente como traduzido do alemão a palavra autovalor. Logo, armazena-se essa informação em *my.eigen*. Não é difícil verificar, no quadro (6), e no ambiente do R Studio, *my.eigen* que contém uma segunda informação, que, ao ser requisitada por *my.eigen$vectors,* irá listar os autovetores. Esses encontram-se listados no quadro (8) a seguir.

6: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```
#############################################################
# Decomposição dos autovalores de R
#############################################################
my.eigen = eigen(my.cov)
my.eigen

#############################################################
# Decomposição dos autovetores de R
#############################################################
loadings = my.eigen$vectors
loadings
```

```
> my.eigen = eigen(my.cov)
> my.eigen
eigen() decomposition
#values
 [1] 9.69608665 2.49547604 1.41159589 1.02696236 0.79277407 0.59258459 0.27787907
  ↪  0.235057886 0.17256200
[10] 0.10445061 0.07829778 0.06083829 0.02096250 0.01680054 0.01002757 0.00535983
 ↪  0.00228425
```

Dos 17 autovetores estimados, apenas os quatro primeiros são maiores que a unidade, mas o que isso indica? Ao se perguntar quais e quantos serão os componentes mais importantes dos 17 estimados, esse resultado seja um indício daqueles autovalores que con-

servam a maior parte da variância do conjunto original. Mas antes dessa discussão, explora-se outro resultado interessante:

Resultado 2 : A soma dos autovalores é igual a p e ao traço de $\mathbf{Z}$

Sejam λ_j os autovalores de $\boldsymbol{Z}$ e $\sum_{j=1}^{n} \lambda_j = p$ sua soma. Considerando uma matriz qualquer $\boldsymbol{Z}$, reescrita $\mathbf{Z}$ na forma canônica de Jordan, i.e. $\boldsymbol{Z} = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^{-1}$ onde $\boldsymbol{S}$ é a matriz dos autovetores generalizados e $\boldsymbol{\Lambda}$ é uma matriz triangular superior com a diagonal formada pelos autovalores de $\boldsymbol{Z}$. Perceba que a função traço nos permite comutar os elementos dentro dela. Tem-se então

$$\text{Tr}(\boldsymbol{Z}) = \text{Tr}(\boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^{-1}) = \text{Tr}(\boldsymbol{\Lambda}\boldsymbol{S}^{-1}\boldsymbol{S}) = \text{Tr}(\boldsymbol{\Lambda}) = p$$

No caso particular das matrizes objeto da técnica, tem-se

$$\sum_{i=1}^{p} Var(X_i) = \text{Tr}(\boldsymbol{\Sigma}) = \text{Tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} Var(Y_i).$$

Vale lembrar também que a soma dos autovalores é igual ao posto da matriz de correlações, isto é, ao número máximo de colunas (no nosso caso variáveis) independentes em uma matriz. Finalmente, o leitor poderá também verificar que o produtório dos autovalores é igual ao determinante da matriz de correlações. Ainda, acrescenta-se que a covariância de $\mathbf{Y}$ será dada por

$$Cov(\mathbf{Y}) = \mathbf{W}\,\text{Cov}(\mathbf{X})\mathbf{W}^{-1} = \mathbf{W}\,\text{Cov}(\mathbf{X})\mathbf{W}^{T}.$$

O Resultado 2 acima pode ser verificado com alguns comandos no R. Portanto, observe que o traço da matriz de variâncias-covariâncias estimada sobre os dados padronizados é igual a soma dos autovalores extraídos. Vale lembrar da álgebra, que o traço de uma matriz nada mais é que a soma dos elementos da diagonal principal. O quadro (7) apresenta essa verificação pelo R.

Os autovalores representam o total de variância que pode ser explicada por um dado componente principal. Na prática, eles podem ser negativos ou positivos, mas, no caso de estarem explicando a variância, eles devem ser positivos. Se os autovalores forem maior que zero, se está no caminho certo.

Como a ACP ordena os autovalores em ordem decrescente, o primeiro autovalor explicará a maior parte da variância do conjunto - cerca de 57% da variância do conjunto. Espera-se que o segundo autovalor tenha um poder de explicação menor, e, à medida que se afasta dos primeiros autovalores, essa capacidade de explicação vai diminuindo. Sendo assim, como se pode ver no quadro (7) a seguir, os três primeiros autovalores são responsáveis por 80% da variância total, e, ao se tomar os cinco primeiros, esse percentual ultrapassa os 90%.

7: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```
#####################################################################
# A soma dos autovalores representa a variância total dos dados

sum(my.eigen$values)

var(df_st[,1]) + var(df_st[,2]) + var(df_st[,3]) + var(df_st[,4]) + var(df_st[,5]) +
  ↪  var(df_st[,6]) + var(df_st[,7]) + var(df_st[,8]) + var(df_st[,9]) +
  ↪  var(df_st[,10]) + var(df_st[,11]) + var(df_st[,12]) + var(df_st[,13]) +
  ↪  var(df_st[,14]) + var(df_st[,15]) + var(df_st[,16]) + var(df_st[,17])

# Com um pacote extra podemos ver a proporção da variância explica por cada autovalor

library("factoextra")
pca <- prcomp(data,  scale = TRUE)

get_eig(pca)
```

```
> sum(my.eigen$values)
[1] 17

> var(df_st[,1]) + var(df_st[,2]) + var(df_st[,3]) + var(df_st[,4]) + var(df_st[,5])
  ↪  + var(df_st[,6]) +
var(df_st[,7]) + var(df_st[,8]) + var(df_st[,9]) + var(df_st[,10]) + var(df_st[,11])
  ↪  + var(df_st[,12]) + var(df_st[,13]) + var(df_st[,14]) + var(df_st[,15]) +
  ↪  var(df_st[,16]) + var(df_st[,17])
[1] 17

```

```
> pca <- prcomp(data,  scale = TRUE)
> get_eig(pca)
        eigenvalue    variance.percent      cumulative.variance.percent
Dim.1  9.696086656      57.03580386                    57.03580
Dim.2  2.495476044      14.67927085                    71.71507
Dim.3  1.411595895       8.30350527                    80.01858
Dim.4  1.026962361       6.04095507                    86.05954
Dim.5  0.792774072       4.66337689                    90.72291
  :         :                :                           :
Dim.16 0.005359831       0.03152842                    99.98656
Dim.17 0.002284257       0.01343680                   100.00000
```

Os autovetores representam o peso de cada autovalor. Antes de se discutir os componentes principais que representam a maior parte da variância-covariância do conjunto, visualiza-se nossa matriz $\mathbf{W}_{17\times17}$.

8: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```
##########################################
#The Eigenvectors
##########################################
loadings = my.eigen$vectors
loadings
```

```
> loadings = my.eigen$vectors
> eigenvectors <- round(loadings,3)
> eigenvectors
          PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9   PC10   PC11
          ↪  PC12   PC13   PC14   PC15
GADO   -0.264  0.281  0.206 -0.053 -0.245  0.007 -0.009  0.131 -0.007  0.028 -0.095
 ↪  0.296  0.246  0.245 -0.213
POP    -0.209 -0.336  0.268  0.056  0.248  0.337  0.100  0.006 -0.482  0.467 -0.313
 ↪  -0.032  0.171 -0.013  0.050
PEA    -0.241 -0.314  0.303  0.022 -0.072  0.207  0.157  0.063  0.010 -0.308  0.066
 ↪  0.259 -0.684 -0.020 -0.185
ASSPEC -0.188 -0.324 -0.172  0.005 -0.490 -0.296 -0.082 -0.664 -0.141  0.140 -0.005
 ↪  0.126  0.004  0.019 -0.003
```

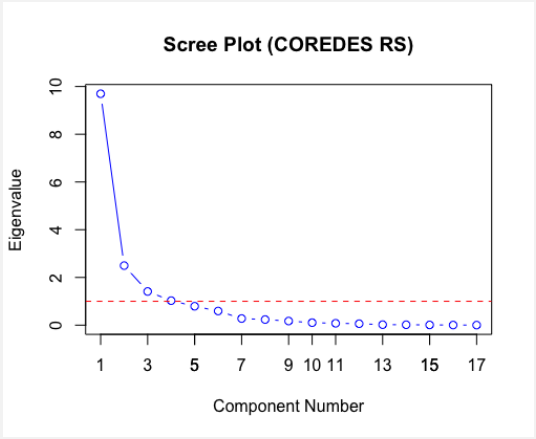
```
ASSAGR -0.157 -0.410 -0.154 -0.169  0.076 -0.517 -0.372  0.556 -0.042  0.085 -0.054
↪  0.141 -0.023 -0.021  0.009
TRATOR -0.299 -0.049 -0.209 -0.002  0.125  0.061  0.081 -0.054  0.447  0.143 -0.223
↪  -0.134  0.101 -0.121 -0.696
ADUBOS -0.179 -0.481  0.154 -0.021 -0.111  0.199  0.048  0.053  0.237 -0.413  0.253
↪  -0.243  0.515  0.133  0.150
IRRIGA -0.250  0.115  0.108 -0.242  0.300  0.271 -0.703 -0.290  0.222  0.071  0.117
↪  0.068 -0.092  0.044  0.161
RECVEG -0.288  0.117 -0.115  0.213  0.258 -0.158  0.114 -0.009 -0.203  0.052  0.414
↪  0.003  0.059  0.379 -0.170
RECANI -0.282  0.156  0.150  0.018 -0.330 -0.078 -0.082  0.152 -0.009  0.194  0.008
↪  -0.759 -0.302  0.111  0.085
FINANC -0.281  0.162 -0.071  0.149  0.218 -0.182 -0.098 -0.142 -0.273 -0.615 -0.529
↪  -0.138  0.037 -0.023  0.061
MAQCOL -0.275 -0.061 -0.317  0.100  0.174 -0.011  0.369 -0.016  0.429  0.174 -0.195
↪  0.093 -0.144  0.086  0.549
VALVEG -0.295  0.081 -0.092  0.193  0.246 -0.104  0.092 -0.059 -0.191 -0.008  0.512
↪  -0.007 -0.010 -0.338  0.018
VALANI -0.289  0.198  0.147 -0.035 -0.241  0.002  0.040  0.123  0.036  0.027  0.022
↪  0.120  0.179 -0.720  0.138
AREXPL -0.276  0.265  0.153 -0.028 -0.209 -0.021  0.072  0.131  0.043  0.031 -0.038
↪  0.312  0.081  0.328  0.159
DISTRG  0.106 -0.065  0.142  0.886 -0.092  0.011 -0.327  0.067  0.199  0.082 -0.063
↪  0.092  0.020 -0.018  0.009
RENDA  -0.095  0.017 -0.668  0.063 -0.294  0.544 -0.173  0.234 -0.250 -0.078  0.028
↪  0.046 -0.039  0.017 -0.005
```

Agora, já se define quantos componentes devem ser retidos para aplicar a análise fatorial, pois, se o objetivo da ACP é reduzir a dimensão, não faz sentido ficar com os 17 componentes estimados. Embora o pesquisador tenha liberdade para definir quantos fatores serão suficientes, existe uma regra de bolso, conhecida como Critério de Kaiser-Guttman. De acordo com a regra, serão retidos apenas os componentes nos quais a variância ultrapasse 1, pois menos que isso significaria que a variância contida seria inferior à variância original dos dados (JOLLIFFE, 2002). Iremos, então, determinar o número de autovalores que sejam maiores que a unidade, contudo, o pesquisador pode definir o número de componentes/fatores que irá utilizar. Como se vê, apenas os quatro primeiros autovalores são maiores que a

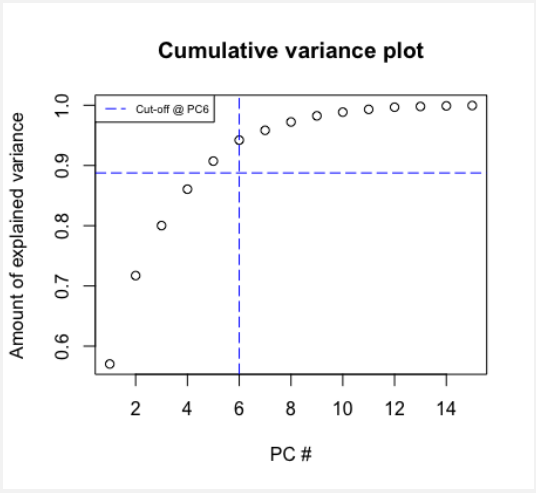
unidade e, juntos, respondem por pouco mais de 86% da variância do conjunto de dados. O quinto autovalor poderia ser considerado, embora sua contribuição seja menor que o autovalor anterior. Se lembre que o objetivo aqui é reduzir o tamanho da amostra para facilitar a interpretação dos resultados, então, dificilmente ele será considerado.

Com o número de autovalores e autovetores definidos, pode-se já tentar extrair algumas conclusões da relação dos componentes com as variáveis. Embora apareçam 15 dos 17 no quadro que segue (8), o interesse é compreender a relação das variáveis com os quatro primeiros componentes.

Vale lembrar que a soma do quadrado das cargas de cada componente deve ser igual a 1. A busca por quantos autovetores reter pode ser auxiliada pela inspeção visual. Alguns pesquisadores costumam observar um gráfico *scree plot* dos autovalores ou da explicação acumulada da variância. A curvatura desses poderá também fornecer pistas para que a definição do número de componentes seja mais assertiva.



(a) Eigenvalues de A



(b) Explicação Acumulada

Figura 2: Ilustração dos Autovalores e poder de explicação

Para finalizar a análise dos componentes, resta, ainda, desvelar a relação de cada indivíduo com o componente em questão. Nesse sentido, deve-se multiplicar a matriz ordenada dos autovetores $\mathbf{W}$ pela matriz original de dados $\mathbf{X}$. Nesse momento, se gera a matriz $\mathbf{F}$, a qual contém os *scores* fatoriais, isto é, a relação de cada observação com cada componente.

9: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```
################################################################################
#The coordinates of the individuals (observations) on the principal components.
################################################################################
pca$x         #score de cada observação
scores <- round(pca$x,2)
```

```
scores
```

```
> scores <- round(pca$x,2)
> scores
          PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10  PC11  PC12  PC13
           ↪  PC14  PC15  PC16  PC17
  [1,] -12.80  3.97 -1.52  0.17 -3.13 -1.15 -0.73  0.14  1.54 -1.44  0.52  0.30  0.03
   ↪  0.23 -0.05  0.04 -0.05
  [2,]   0.98  1.92 -0.87  0.45 -0.19 -0.35  0.05 -0.25  0.34  0.43 -0.20  0.07 -0.05
   ↪  -0.11 -0.01  0.04  0.00
  [3,]  -9.47  3.77  0.65 -3.51  1.32 -0.55 -0.15 -0.01 -1.14 -0.04  0.07 -0.37 -0.30
   ↪  -0.35 -0.13  0.16 -0.10
  [4,]   0.82  1.86  0.42  0.44 -1.00 -0.11  0.09 -0.68  0.09  0.18 -0.03  0.29 -0.10
   ↪  0.07 -0.02  0.01  0.07
  [5,]  -1.21  2.83 -0.77  0.91 -1.73 -0.18  0.31  0.11 -0.29  0.19 -0.17  0.29  0.06
   ↪  0.14 -0.06 -0.17  0.10
  [6,]  -5.81  2.48 -1.48  1.45 -1.37 -0.01  0.58  0.24  0.00  0.13 -0.26  0.28  0.32
   ↪  -0.01 -0.22  0.27  0.02
  [7,]  -7.11  3.83 -3.28  2.01 -4.47  0.41  0.64  1.24 -0.21  0.79 -0.23  0.06 -0.12
   ↪  -0.03  0.12 -0.27  0.05
  [8,]  -7.19  1.47  1.01 -1.94  2.75 -0.28 -0.24 -0.76 -0.29 -0.54 -0.88  0.15 -0.09
   ↪  0.26 -0.20 -0.21  0.06
  [9,] -10.60  3.54 -1.79  1.43 -0.15  0.64  0.32  0.25 -0.33 -0.13 -0.77 -0.08 -0.12
   ↪  -0.22  0.00  0.31  0.05
 [ reached getOption("max.print") -- omitted 130 rows ]
```

Uma nota importante antes de se prosseguir: existem diversos pacotes de R que fazem as mesmas operações, que, em sua maioria, diferem nos comandos para apresentar os dados da ACP, o que, em certa medida, pode confundir o pesquisador. Algumas vezes, os sinais dos autovetores são invertidos, contudo, o pacote *Procedures for Psychological, Psychometric, and Personality Research (PSYCH)* apresenta valores para os componentes principais diferentes dos demais. Nesse sentido, os autovetores são alterados como na igualdade (19) que segue, isto é, o software utiliza os autovalores para produzir as cargas fatorais, que será o assunto estudo a seguir.

10: Código no R - Exemplo com os COREDES agropecuários do RS (1994/95)

```r
library(psych)
library(GPArotation)

#a table of loadings (or correlations between variables and PCs).
pca.unrotated <- principal(data, nfactors=17, rotate="none")
print(pca.unrotated$loadings[,])
```

```
> print(pca.unrotated$loadings[,])
           PC1.    PC2     PC3     PC4    PC5     PC6     PC7     PC8     PC9
gado     0.8227 -0.4441 -0.2450 -0.0533  0.2179  0.0054  0.0049  0.0636 -0.0027
pop      0.6502  0.5311 -0.3188  0.0572 -0.2208  0.2595 -0.0526  0.0029 -0.2001
pea      0.7496  0.4965 -0.3595  0.0226  0.0637  0.1595 -0.0826  0.0304  0.0042
asspec   0.5859  0.5110  0.2041  0.0049  0.4359 -0.2275  0.0429 -0.3218 -0.0586
assagr   0.4897  0.6484  0.1832 -0.1710 -0.0674 -0.3983  0.1958  0.2697 -0.0173
trator   0.9316  0.0781  0.2485 -0.0024 -0.1114  0.0467 -0.0426 -0.0262  0.1856
adubos   0.5581  0.7603 -0.1826 -0.0209  0.0987  0.1528 -0.0254  0.0258  0.0982
irriga   0.7793 -0.1818 -0.1286 -0.2456 -0.2669  0.2086  0.3705 -0.1407  0.0921
recveg   0.8974 -0.1841  0.1360  0.2162 -0.2291 -0.1215 -0.0601 -0.0042 -0.0842
recani   0.8766 -0.2469 -0.1784  0.0186  0.2934 -0.0599  0.0430  0.0737 -0.0037
financ   0.8744 -0.2552  0.0841  0.1512 -0.1940 -0.1402  0.0515 -0.0687 -0.1132
magcol   0.8571  0.0959  0.3763  0.1016 -0.1546 -0.0086 -0.1946 -0.0075  0.1782
valveg   0.9198 -0.1280  0.1089  0.1956 -0.2194 -0.0802 -0.0486 -0.0286 -0.0791
valani   0.8992 -0.3135 -0.1748 -0.0352  0.2145  0.0017 -0.0208  0.0597  0.0150
arexpl   0.8595 -0.4189 -0.1822 -0.0283  0.1860 -0.0160 -0.0377  0.0634  0.0178
distrg  -0.3308  0.1021 -0.1692  0.8974  0.0818  0.0080  0.1725  0.0324  0.0826
renda    0.2968 -0.0270  0.7939  0.0642  0.2619  0.4187  0.0910  0.1132 -0.1038
```

Matematicamente, o que o software faz é $\sqrt{\lambda_i}\vec{e}_i$, o que nos prepara mais para uma Análise Fatorial do que ACP. E esse dado gerado pode ser interpretado como a correlação de cada variável com o respectivo componente. Ainda, disso decorre que a soma do quadrado dessas cargas irá resultar no autovalor que gerou o componente. Facilmente, pode-se mostrar com a primeira coluna do quadro (10)

$$(0,822731)^2 + (0,650256)^2 + \cdots + (0,296853)^2 = \underbrace{9,696087}_{\text{primeiro autovalor}}$$

1.1.4 Descomposição Espectral no R: abordagem matricial

A função que computa os autovalores já é de conhecimento do leitor, portanto, com o comando $eigen$ temos acesso aos autovetores e autovalores por meio da decomposição espectral sobre matrizes reais ou complexas. Lembre que eles podem ser extraídos das matrizes de correlação ou variância-covariância, e se esse última estiver padronizada, termos a matriz de correlações. Assim, o comando que segue é uma função nativa do R - não precisa de nenhum pacote auxiliar.

1: Código no R - Linhas de comando

```
e.cor<-eigen(cor(X))# p/ matriz de correlação (i.e., com dados padronizados)
e.cov<-eigen(cov(X))# p/ matriz de variância-covariância
```

O objeto criado pelos comandos (i.e., e.cor ou e.cov) será do tipo lista, e conterá dois elementos: *values* e *vectors*. *Values* são um vetor que contém p autovalores da matriz de dados, ordenados conforme o volume de variância que é explicada. *Vectors* será uma matriz $k \times k$ que contém os autovetores. Ambos podem ser alocados em objetos, tal como segue:

2: Código no R - Linhas de comando

```
eigenvalues<-e.cor$values #ou eigenvalues<-e.cov$values p/ dados padronizados
eigenvectors<-e.cor$vectors
```

Contudo, a função *eigen* apenas irá fornecer os autovalores e os respectivos autovetores, os escores, bem como a proporção de variância explicada, a variância cumulativa explicada e o desvio-padrão das novas variáveis pode ser calculados a partir da criação de dois novos objetos. Nesse ponto, vamos nos deter apenas na matriz de correlação, portanto, partindo de uma matriz de dados $\mathbf{X}_{n\times k}$ temos:

3: Código no R - Linhas de comando

```
X2=scale(X) #Nova matriz de dados, agora padronizados, média 0 e variância constante
scores<-X2%*%eigenvectors #Novas variáveis
total.var<-sum(diag(cov(X2))) #Calculando a variância total da matriz de dados
 ↪ padronizada, lembre que esse é. traço da matriz, o qual deverá ser igual ao
 ↪ número de variáveis totais utilizadas
prop.var<-rep(NA,ncol(X));cum.var<-rep(NA,ncol(X)) #cria vetores vazios
```

```
#Calculando a porporção da variância explicada e variância explicada acumulada
for(i in 1:ncol(X)){prop.var[i]<-var(scores[,i])/total.var}
for(i in 1:ncol(X)){cum.var[i]<-sum(prop.var[1,i])
sdev=sqrt(eigenvectors) #desvio-padrão dos componentes
```

Para facilitar esse processo, um *script* pode ser executado, tal que ele segue a sintaxe

4: Código no R - Linhas de comando

```
pc_data1<-PC(X,method="eigen",scaled=T,graph=F,rm.na=T,print.results=T
```

A função que segue irá extrair os componentes principais usando a decomposição espectral ou a decomposição do valor singular. A saída dessa função consiste de um objeto que irá conter os autovalores, autovetores, escores, um resumo da saída e os desvios-padrão dos novos componentes. Por *default* a função utilizada a matriz de correlações (i.e., scaled=T) para a decomposição espectral, removendo todas as observações com valores faltantes ou perdidos, e imprime um resumo dos resultados. Para utilizar a decomposição do vlaor singular, o método (method) deverá ser alterado para "svd". Quando especifica-se *graph=T*, a função irá incluir dois gráficos: (1) um *screeplot* e (2) um *biplot* dos dois primeiros componentes. Caso a função identifique valores faltantes no banco de dados de entrada, o usuário tem duas opções: (1) rm.na=T (default) para remover todas as observações com dados faltantes ou (2) rm.na=F para substituir no valor faltante a média da variável.

5: Código no R - Linhas de comando

```
PC<-function(X,method="eigen",scaled=T,graph=F,rm.na=T,print.results=T){
 if (any(is.na(X))){
 tmp<-X
 if(rm.na==T){X<-na.omit(data.frame(X));X<-as.matrix(X)}
 else{X[is.na(X)] = matrix(apply(X, 2, mean, na.rm = TRUE),
 ncol = ncol(X), nrow = nrow(X), byrow = TRUE)[is.na(X)]}}
 else{tmp<-X}
 if(method=="eigen"){
 if(scaled==1){X1=cor(X);X2=scale(X)}
 else{X1=cov(X);X2=scale(X,scale=F)}
 total.var<-sum(diag(cov(X2)))
 values<-eigen(X1)$values;vectors<-eigen(X1)$vectors;sdev=sqrt(values)}
```

```
if(method=="svd"){
if(sum(scaled,center)>1){X2<-scale(X)}else{
if(scaled==1){X2=scale(X,center=F)}else{
if(center==1){X2=scale(X,scale=F)}else{X2=X}}}
total.var<-sum(diag(cov(X2)))
var<-nrow(X2)-1
vectors<-svd(X2)$v;sdev=svd(X2)$d/sqrt(var);values<-sdev*sdev}
prop.var<-rep(NA,ncol(X));cum.var<-rep(NA,ncol(X));scores<-X2%*%vectors
namex<-as.character(1:ncol(X));scorenames<-rep(NA,ncol(X))
for(i in 1:(ncol(X))){
scorenames[i]<-do.call(paste,c("PC",as.list(namex[i]),sep=""))}
colnames(scores)<-scorenames
rownames(vectors)<-colnames(X);colnames(vectors)<-scorenames
for(i in 1:ncol(X)){prop.var[i]<-var(scores[,i])/total.var}
for(i in 1:ncol(X)){cum.var[i]<-sum(prop.var[1:i])}
importance<-t(matrix(c(sdev,prop.var,cum.var),ncol=3))
importance<-as.table(importance)
colnames(importance)<-scorenames
rownames(importance)<-c("Standard Deviation","Proportion of Variance","Cumulative
Proportion")
z<-list(values=values,vectors=vectors,scores=scores,importance=importance
,sdev=sdev)
if(graph==1){
biplot(scores[,1:2],vectors[,1:2], main="Biplot of Data",xlab=do.call
(paste,c("PC1 (",as.list(round(z$importance[2,1]*100,2)),"%)",sep=""))
,ylab=do.call(paste,c("PC2
(",as.list(round(z$importance[2,2]*100,2)),"%)",sep="")), cex=0.7)
windows()
screeplot(z,type='l',main='Screeplot of Components')
abline(1,0,col='red',lty=2)}
if(print.results==T){
if(method=="eigen"){print("PCA Analysis Using Spectral Decomposition")}
if(method=="svd"){print("PCA Analyis Using Singular Value Decomposition")}
if (any(is.na(tmp))){
if(rm.na==T){print("Warning:One or more rows of data were omitted from
analysis")}
if(rm.na==F){print("Warning: Mean of the variable was used for Missing
```

```
 values")}}
 print(importance)}
 z<-list(values=values,vectors=vectors,scores=scores,importance=importance
 ,sdev=sdev)
}#End Function
```

Vale salientar que a ACP é um método puramente algébrico e resume o comportamento dos dados, isto é, a ACP produz componentes que são uma combinação linear dos dados observados, sem modelo causal – erro observado (TIMM, 2002, p. 497). Por outro lado, a Análise Fatorial é um modelo estatístico que intenta descrever a estrutura das relações da covariância entre muitas variáveis em termos de um conjunto menor, mas não observável diretamente (latentes) chamadas de fatores (JOHNSON and WICHERN, 1992).

Referências

Carboni, I. d. F. C. (2003). *Lógica de programação*. Cengage Learning Editores.

de Freitas, C. A., Paz, M. V., and Nicola, D. S. (2007). Analisando a modernização da agropecuária gaúcha: uma aplicação de analise fatorial e cluster. *Análise Econômica*, 25(47).

Favero, P. L. and Belfiore, P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e STATA*. 1. ed. – Rio de Janeiro: Elsevier.

Firme, V. d. A. C. and Vasconcelos, C. R. F. (2015). Identificação de nichos de mercado para países exportadores: uma análise multivariada para o ano de 2011. *Análise Econômica*, 33(64).

Gorsuch, R. L. (1993). *Applied factor analysis in the natural sciences*. Cambridge University Press, Cambridge.

JOHNSON, R. A. and WICHERN, D. W. (1992). *Applied multivariate statistical analysis*, volume 3. Prentice hall Upper Saddle River, NJ.

JOLLIFFE, I. T. (2002). Principal components in regression analysis. *Principal component analysis*, pages 167–198.

Kageyama, A. and Leone, E. T. (1999). Uma tipologia dos municípios paulistas com base em indicadores sociodemográficos. *Campinas: UNICAMP/IE*.

Mingoti, S. A. and da Silva, A. F. (1997). Um exemplo de aplicação de técnicas de estatística multivariada na construção de índices de preços. *Nova Economia*, 7(2).

Morrison, D. F., Marshall, L. C., and Sahlin, H. L. (1976). *Multivariate statistical methods*. McGraw-Hill New York.

Poerschke, R. P. (2007). Análise multivariada de dados socioeconômicos: um retrato da modernização agrícola no rio grande do sul. page 130 p.

Reyment, R. A. and Jvreskog, K. (1993). *Applied factor analysis in the natural sciences*. Cambridge University Press.

Schneider, S. and Waquil, P. D. (2001). Caracterização socioeconômica dos municípios gaúchos e desigualdades regionais. *Revista de Economia e Sociologia Rural*, 39(3):117–142.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology*, 15:201–93.

TIMM, N. H. (2002). *Applied multivariate analysis*. Springer, New York.

rafaelpoerschke