UFSM

CIÊNCIA e NATURA

ISSN 2179-460X

OPEN ACCESS

Ci. e Nat., Santa Maria, v. 47, esp. 1, e90532, 2025 • https://doi.org/10.5902/2179460X90532

Submissão: 30/04/24 • Aprovação: 09/07/24 • Publicação: 14/02/25

IV Jornada de Matematica e Matematica aplicada UFSM

PCA: uma ferramenta matemática para a análise dos COREDEs agropecuários do Rio Grande do Sul

PCA: a mathematical tool for the analysis of agricultural and livestock COREDEs in Rio Grande do Sul

Rafael Pentiado Poerschke [I], **João Roberto Lazzarin** [II], **Fernando Colman Tura** [II]

[I] Universidade Federal de Santa Maria, RS, Brasil

RESUMO

Este estudo teve como objetivo principal a redução da matriz original de dados dos Conselhos Regionais de Desenvolvimento (COREDEs) agropecuários do Rio Grande do Sul, utilizando informações do Censo Agropecuário de 2017. Iniciamos com um rigoroso apanhado de resultados matemáticos que fundamentam a Análise de Componentes Principais (Principal Components Analysis - PCA), culminando com uma aplicação da PCA, reduzindo significativamente os dados coletados inicialmente de 15 variáveis para somente três componentes, que foram capazes de explicar cerca de 87% da variância dos dados. Essas informações podem influenciar na tomada de decisões na condução de políticas agrícolas e estratégias de desenvolvimento regional voltadas para esses COREDEs.

Palavras-chave: Análise de componentes Principais; Decomposição espectral; Censo agropecuário; COREDEs

ABSTRACT

This study aimed to reduce the original data matrix from the agricultural Regional Development Councils (COREDEs) of Rio Grande do Sul, utilizing information from the 2017 Agricultural Census. We commenced with a rigorous exposition on mathematical results underpinning Principal Components Analysis (PCA), culminating in an appliction of PCA that significantly reduced the initially collected data from 15 variables to only three components, which collectively explained approximately 87% of the data variance. These insights have the potential to influence decision-making in agricultural policies and regional development strategies targeted at these COREDEs.

Keywords: Principal component analysis; Spectral decomposition; Census of agriculture; COREDEs

1 INTRODUÇÃO

O levantamento sistemático de diversas variáveis, sobre uma ampla gama de indivíduos, empresas, municípios e/ou países, resulta em matrizes de dados de elevada dimensão. Em geral, o número de observações supera o montante de variáveis observadas, portanto, grande parte desse conjunto de dados não possui as características de uma matriz simétrica, limitando o uso de técnicas para sua decomposição. Para matrizes simétricas, a técnica da Decomposição Espectral de Matrizes (DEM) tem uma série de aplicações em diversos ramos das ciências.

No âmbito nacional, as pesquisas em periódicos de economia que compartilharam dessa técnica são limitadas. O mais célebre desses trabalhos buscou tipificar um conjunto de municípios no estado de São Paulo (SP) a partir de suas principais características sociais e econômicas (Kageyama & Leone (1999)). Essa mesma ideia aparece para o Rio Grande do Sul (RS), quando se buscou a criação de grupos homogêneos de municípios em Schneider & Waquil (2001). Por outro lado, Freitas et al. (2007) abordaram o estado do Rio Grande do Sul com a ideia de explorar a existência de padrões determinados pelo grau de utilização de insumos agrícolas modernos nos estabelecimentos rurais. O objetivo dos autores foi agrupá-los conforme sua similaridade e relação com um conjunto de variáveis latentes obtidas a partir das variáveis originais do Censo Agropecuário de 1995/96.

Em comum, ambos os trabalhos, que se depararam com um número elevado de observações, bem como de variáveis coletadas sobre esses municípios, buscaram reduzir a dimensão inicial da matriz de dados. É nesse sentido que a Análise de Componentes Principais (do inglês, PCA - *Principal Component Analysis*) pode ser útil, uma vez que consiste em uma transformação linear capaz de reduzir uma matriz de dados inicial a um novo conjunto menor, ortogonal e mais homogêneo. É possível substituir um conjunto de entrada de p variáveis originais por m componentes principais, de maneira que o número de componentes principais seja menor que o número de variáveis observadas. Essa redução de dimensão, além dos aspectos computacionais, auxilia o pesquisador ou o agente tomador de decisão a conduzir uma análise exploratória sem precisar trabalhar sobre a matriz original.

Em outras palavras, verificada a relação entre grupos de variáveis correlacionadas entre si, elas podem ser agrupadas e representadas em componentes que conservem o

máximo da informação original. A possibilidade de otimizar a interpretação de grandes conjuntos de dados em um número menor de variáveis latentes é de grande utilidade na análise econômica, uma vez que é notório o acesso a dados de corte transversal, sejam eles oriundos de levantamentos primários ou em bases de dados secundários, como é o caso dos dados de Censos.

Neste estudo, pretendemos apresentar com rigor a matemática por trás da técnica da PCA. Posteriormente, iremos ilustrar o método com uma aplicação da PCA a fim de selecionar os componentes principais de uma base de dados para o Rio Grande do Sul (RS). Nesse sentido, com a estimativa dos componentes, buscamos investigar os COREDEs[1] agropecuários gaúchos, considerando a existência e o grau de similaridade entre os municípios com base nos dados do Censo Agropecuário de 2017. Num conjunto de 127 municípios, agregados em 8 COREDEs[2] predominantemente agropecuários[3], questionamos a homogeneidade desse grupo quando decomposto via DEM. Em outras palavras, em que medida essa agregação dos municípios por contiguidade garantiria a homogeneidade dos COREDEs predominantemente agropecuários.

A pesquisa, além dessa breve introdução, traz uma discussão sobre os conhecimentos fundamentais para a operacionalização da derivação e análise estatística dos componentes principais. Na seção seguinte, apresentamos a técnica da Análise de Componentes Principais; serão posteriormente definidos os dados utilizados e os procedimentos empregados para o tratamento e estimação dos componentes. Na última seção, abordamos os principais resultados e são apresentados apontamentos para as novas etapas nas quais esta pesquisa irá avançar.

2 ANÁLISE DE COMPONENTES PRINCIPAIS: REDUÇÃO DA DIMENSÃO DE MATRIZES DE DADOS

O principal objetivo da **Análise de Componentes Principais** (**PCA**) é a redução da dimensionalidade dos dados pela obtenção de variáveis latentes. Estas novas variáveis

[1]Os Conselhos Regionais de Desenvolvimento foram criados oficialmente pela Lei 10.283/1994. Por definição, são um fórum de discussão para a promoção de políticas e ações que visam o desenvolvimento regional no RS.

[2]Os municípios e seus respectivos COREDEs podem ser encontrados no Anexo A.

[3]A tipificação dos COREDEs foi efetuada pelo projeto Rumos 2015, o qual mostrou a existência da dinâmica econômica desses agrupamentos segundo a renda gerada pelos setores de Serviços, da Indústria e da Agropecuária.

são combinações lineares das variáveis originais, escolhidas de tal forma que capturam a maior quantidade possível da variância dos dados originais.

Portanto, a PCA transforma linearmente um conjunto de p variáveis correlacionadas entre si em um novo grupo de m variáveis latentes e ortogonais (com $m \leq p$), que explicam uma parcela substancial das informações do conjunto original. Essa abordagem possibilita a geração, a seleção e a interpretação dos componentes investigados e, também, auxilia na identificação das variáveis de maior influência na formação de cada componente. Ou seja, a PCA procura explicar a estrutura da variância e covariância de um conjunto de variáveis utilizando uma parcela do todo, os componentes principais, com o mínimo de perda de informação.

2.1 Conceitos e Notações Matemáticas

Antes de prosseguir com a apresentação da Análise de Componentes Principais, é essencial estabelecer um conjunto de notações e definições relevantes em estatística e álgebra linear. Essas ferramentas serão fundamentais para uma compreensão mais profunda dos métodos e resultados apresentados nesta pesquisa.

2.1.1 Da Estatística Básica

Considere que $\mathbf{A}^{\mathsf{T}}$ denota a matriz transposta de uma matriz $\mathbf{A}$. Fixemos uma matriz $\mathbf{X} = (x_{jk})_{(n\times p)}$, que em termos práticos, cada x_{jk}, representa a k-ésima variável observada sobre o j-ésimo item, isto é, x_{jk} representa a medida da k-ésima variável observada sobre o j-ésimo item. A matriz $\mathbf{X}$ contém o conjunto de dados de todas as observações e variáveis. Nestes termos $\mathbf{x}_1 = [\, x_{11},\, x_{21},\, \ldots,\, x_{n1}]^{\mathsf{T}}$ reúne as n observações sobre a primeira variável, e assim sucessivamente para as demais p variáveis. Assim, $\mathbf{X}$ representa um vetor das p **variáveis aleatórias**: $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$, dispostos em colunas, ou seja,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{(n\times p)} . \tag{1}$$

Cada variável aleatória $\mathbf{x_k}$ fornecerá uma média. Nesta pesquisa, em específico, ela foi tomada sempre em termos amostrais. Portanto, têm-se p **médias amostrais** dadas por:

$$\overline{x}_k = \frac{1}{n}\sum_{j=1}^{n} = x_{jk} \qquad k = 1,\ 2,\ \ldots,\ p. \tag{2}$$

As médias do vetor de variáveis aleatórias $\mathbf{X}$ são reunidas no vetor de médias μ

$$\mathbb{E}[\mathbf{x}] = \mu\ = \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \ldots & \overline{x}_p \end{bmatrix}^{\mathsf{T}} \tag{3}$$

sendo que cada média é calculada por (2). A **variância amostral** é definida, para cada $\mathbf{x_k}$, por

$$s_k^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{jk} - \overline{x}_k)^2 \qquad k = 1,\ 2,\ \ldots,\ p, \tag{4}$$

e a **covariância amostral**, para cada par x_k e x_l fixados, definida por

$$s_{kl}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{jk} - \overline{x}_k)(x_{jl} - \overline{x}_l), \quad k \neq l\ . \tag{5}$$

Como isso, obtemos uma matriz de dimensão $p \times p$ dada por

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

chamada **matriz de variâncias-covariâncias**, que ao longo do texto será denotada simplesmente por $\mathbf{S}$ ou por $\mathbf{Cov}(\mathbf{S})$ ou ainda por $\mathbf{S_X}$, quando houver dúvida de quais variáveis $\mathbf{X}$ estamos tratando. É fácil ver que por (5) tem-se $s_{kk} = s_k$ e, além disso, $s_{kj} = s_{jk}$ e que, portanto, **S** é uma matriz simétrica.

Quando as variáveis aleatórias são de dimensões muito distintas, é preferencial que se use, ao invés de $\mathbf{S}$, uma matriz padronizada para se calcular os componentes

principais. Essa é a chamada matriz de **matriz de correlação amostral**[4], que é dada por:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & \cdots & \cdots & 1 \end{bmatrix}_{(p\times p)}$$

tal que cada **correlação amostral** é calculada por

$$r_{kl} = \frac{s_{kl}}{s_k s_l}, \qquad k, l = 1, 2, \ldots, p. \tag{6}$$

Note que o coeficiente de correlação varia no intervalo $-1 \leq r_{kl} \leq 1$. Um coeficiente de correlação próximo a 1 indica uma forte correlação positiva, o que significa que as variáveis tendem a aumentar ou diminuir juntas. Por outro lado, um coeficiente próximo a -1 sugere uma forte correlação negativa, indicando que as variáveis tendem a se mover em direções opostas. Um coeficiente de correlação próximo a 0 indica uma correlação fraca ou inexistente entre as variáveis, o que significa que elas não têm relação linear discernível entre si.

2.1.2 Decomposição Espectral de Matrizes

Iniciamos com um pouco de notação da álgebra linear. Uma matriz $\mathbf{A} \in \mathbf{M}_{p\times p}(\mathbb{R})$ possui um **autovalor** $\lambda \in \mathbb{R}$ se, e somente se, existe um vetor $\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix} \in \mathbb{R}^p$, não nulo, tal que $\mathbf{Au} = \lambda\mathbf{u}$. Nesse caso, dizemos que $\mathbf{u}$ é um autovetor associado ao autovalor λ. Pela equação $\mathbf{Au} = \lambda\mathbf{u}$ é fácil ver que o problema de calcular autovalores de uma matriz se reduz ao de encontrar as raízes reais para o polinômio $p(\lambda) = det(\mathbf{A} - \lambda\mathbf{I})$. Este polinômio é conhecido na literatura como sendo o **polinômio característico** de $\mathbf{A}$.

Fica evidente que matrizes diagonais, matrizes de variância-covariância e de correlação são exemplos de matrizes simétricas. Porém, pode-se construir matrizes simétricas usando matrizes ortogonais reais, que são matrizes quadradas com

[4]Se o caso for de uma padronização de média zero e variância constante (1), temos: $z_{ik} = \frac{x_{ik} - \overline{x}_k}{s(\mathbf{x}_k)}$, tal que i é a i-ésima linha da j-ésima coluna e $s(\mathbf{x}_k) = \sqrt{s_k^2}$ é o desvio-padrão da k-ésima variável $\mathbf{x_k}$.

entradas reais que são inversíveis e que cuja inversa é a sua transposta. Nestes casos, se $\mathbf{Q}_{(n\times n)}$ é ortogonal e $\mathbf{\Lambda}_{(n\times n)}$ é uma matriz diagonal, então $\mathbf{A}_{(n\times n)}$ dada por

$$\mathbf{A} \coloneqq \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{T}}$$

é simétrica, pois $\mathbf{A}^{\mathsf{T}} = \left(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{T}}\right)^{\mathsf{T}} = \left(\mathbf{Q}^{\mathsf{T}}\right)^{\mathsf{T}} \mathbf{\Lambda}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{T}} = \mathbf{A}$.

Para finalizar, diremos que uma matriz simétrica é **positiva semi-definida** quando seus autovalores são todos não negativos. As matrizes de variância-covariância e de correlação são exemplos de matrizes semi-definidas (para detalhes, ver Johnson & Wichern (2002)).

Nas pesquisas em economia, as matrizes de dados em geral são não quadradas, e o mesmo vale neste estudo, uma vez que analisaremos dados de 127 municípios com 17 variáveis cada. Contudo, a técnica da PCA se baseia na decomposição das matrizes de variância-covariância e de correlação, que são simétricas. Sendo assim, a obtenção de resultados matemáticos se concentra nestes tipos de matrizes. Nesse sentido, enunciamos agora o teorema que sustentará a análise feita nas seções finais deste artigo.

Teorema 1. ***(Teorema da Decomposição Espectral)*** *Seja $\mathbf{A}$ uma matriz simétrica $p \times p$. Então, existem:*

- *Um conjunto de p autovetores linearmente independentes $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ associados a $\mathbf{A}$.*
- *Um conjunto correspondente de p autovalores reais (distintos ou não): $\lambda_1, \lambda_2, \ldots, \lambda_p$,*

tal que a matriz $\mathbf{A}$ pode ser decomposta como:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{T}} = \lambda_1\mathbf{v}_1\mathbf{v}_1^{\mathsf{T}} + \lambda_2\mathbf{v}_2\mathbf{v}_2^{\mathsf{T}} + \cdots + \lambda_p\mathbf{v}_p\mathbf{v}_p^{\mathsf{T}}.$$

onde:

- *$\mathbf{Q}$ é a matriz ortogonal cujas colunas são os autovetores normalizados $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$.*
- *$\mathbf{\Lambda}$ é uma matriz diagonal cujos elementos da diagonal são os autovalores $\lambda_1, \lambda_2, \ldots, \lambda_p$ de $\mathbf{A}$.*

A prova[5] deste teorema é baseada nos seguintes resultados:

1. Toda matriz simétrica com entradas reais possui todos os seus autovalores reais;
2. Toda matriz simétrica com entradas reais é diagonalizável, isto é, possui uma base de autovetores para o espaço $\mathbb{R}^p$;
3. Utilizando-se o Processo de Gram-Schimidt, caso necessário, esta base de autovetores pode ser ortogonalizada e, após normalização destes autovetores, obtém-se uma base $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ ortonormal para $\mathbb{R}^p$;
4. Com estes vetores, constrói-se uma matriz ortogonal $\mathbf{Q} = \begin{bmatrix} \vdots & & \vdots \\ \mathbf{v}_1 & \cdots & \mathbf{v}_p \\ \vdots & & \vdots \end{bmatrix}_{(p \times p)}$ cujas colunas são os autovetores da base ortonormal;
5. Estes vetores $\mathbf{v}_1, \ldots, \mathbf{v}_p,$ juntamente com a matriz diagonal $\mathbf{\Lambda}$ formada pelos autovalores correspondentes dispostos na diagonal, é que irão satisfazer o teorema.

Desde que as matrizes de variância-covariância amostral ($\mathbf{S}$) e de correlações ($\mathbf{R}$) se enquadram nas hipóteses deste teorema, concluímos que estas possuem autovalores reais, geram bases de autovetores ortogonais e, portanto, são decompostas pelo método da Decomposição Espectral, o que será de grande valia para as próximas seções.

3 ANÁLISE DE COMPONENTES PRINCIPAIS: UMA AVALIAÇÃO DO MÉTODO

Segundo Mardia et al. (1979), uma combinação linear adequada de variáveis promove muitas vezes mais informação que o conjunto original de variáveis, melhor ainda, se a dimensão destes novos dados for reduzida. Ainda, segundo os autores, as transformações lineares podem simplificar a estrutura da matriz de covariância-correlação, tornando sua interpretação mais direta e descomplicada.

Em termos da álgebra linear, as componentes principais são provenientes dos autovetores ortogonais que definem novos eixos no espaço de características dos

[5]Uma demonstração completa pode ser encontrada em Strang (2019)Strang (2019).

dados. Esses autovetores (variáveis latentes) são derivados diretamente da matriz de variância-covariância ou da matriz de correlações provenientes da matriz de dados (1). Resumidamente, os componentes principais são formados por combinações lineares dos dados originais, cujos escalares são as coordenadas dos autovetores correspondentes aos maiores autovalores. Vamos precisar melhor estas ideias nesta seção.

Considere a i-ésima combinação linear

$$\mathbf{w}_r = \mathbf{u}^{\mathsf{T}}\mathbf{X} = u_1\mathbf{x}_{r1} + \cdots + u_p\mathbf{x}_{rp}, \qquad r = 1, 2, \ldots, n, \tag{7}$$

onde $\mathbf{u} = [u_1, \ldots, u_p]^{\mathsf{T}}$ é dado. Utilizando (2), a média $\overline{w}$ de $\mathbf{w}$ é dada por

$$\overline{w} = \frac{1}{n}\mathbf{u}^{\mathsf{T}}\sum_{r=1}^{n}\mathbf{x}_r = \mathbf{u}^{\mathsf{T}}\overline{\mathbf{x}}, \tag{8}$$

assim, baseado em (4) e (5) podemos calcular a variância de $\mathbf{w}_r = [w_1, \ldots, w_n]$, isto é

$$s_w^2 = \frac{1}{n}\sum_{r=1}^{n}(\mathbf{w}_r - \overline{w})^2 = \frac{1}{n}\sum_{r=1}^{n}\mathbf{u}^{\mathsf{T}}(\mathbf{x}_r - \overline{\mathbf{x}})(\mathbf{x}_r - \overline{\mathbf{x}})^{\mathsf{T}}\mathbf{u} = \mathbf{u}^{\mathsf{T}}\mathbf{S}\mathbf{u}. \tag{9}$$

Tomando cada $\mathbf{x}_i$ como um dos vetores que compõe o vetor de variáveis aleatórias correlacionadas $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$, que possui matriz de variâncias-covariâncias $\mathbf{S}$, pela Expressão (7) eles serão combinados com as coordenadas do vetor $\mathbf{u}$. Portanto, $\mathbf{X}$ será transformado em uma nova matriz $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p]$, que pretendemos que seja composto por variáveis não correlacionadas.

Definimos cada novo $\mathbf{w}_i$ em função linear dos autovetores de $\mathbf{S}$, combinados com os vetores que compõe $\mathbf{X}$ do seguinte modo

$$\mathbf{w}_i = \mathbf{u}_i^{\mathsf{T}}\mathbf{X} = u_{i1}\mathbf{x}_1 + u_{i2}\mathbf{x}_2 + \cdots + u_{ip}\mathbf{x}_p, \qquad \forall i = 1, 2, \ldots, p. \tag{10}$$

Por exemplo, $\mathbf{u}_1$ é um vetor dado por $\mathbf{u}_1 = [u_{11}, u_{12}, \ldots, u_{1p}]$. Portanto, o primeiro componente principal será a combinação linear

$$\mathbf{w}_1 = \mathbf{u}_1^{\mathsf{T}}\mathbf{X} = u_{11}\mathbf{x}_1 + u_{12}\mathbf{x}_2 + \cdots + u_{1p}\mathbf{x}_p = \sum_{i=1}^{p}\mathbf{u}_{1i}\mathbf{x}_i.$$

Assim, o primeiro componente principal $\mathbf{w}_1^{\mathsf{T}} = [w_{11},\, w_{12}, \ldots,\, w_{1n}]$ tem coordenadas dadas por

$$
\begin{aligned}
w_{11} &= u_{11}x_{11} + u_{12}x_{12} + \cdots + u_{1p}x_{1p} \\
w_{12} &= u_{11}x_{21} + u_{12}x_{22} + \cdots + u_{1p}x_{2p} \\
\vdots \;&= \qquad\qquad\qquad\qquad \vdots \\
w_{1n} &= u_{11}x_{n1} + u_{12}x_{n2} + \cdots + u_{1p}x_{np}\,.
\end{aligned}
$$

Queremos verificar que uma parcela dos componentes combinados por (10) serão responsáveis por uma parcela significativa da variância original dos dados, uma vez que ela será distribuída de forma não-crescente a medida que os $\mathbf{w}_p$ componentes principais são estimados e haverá uma parte residual de menor variância (Jöreskog (1979)).

Definição 1. *(**Componentes Principais**) Se $\mathbf{X}$ é um vetor de p variáveis aleatórias com média μ e matriz de variância-covariância amostral $\mathbf{S_X}$, então a transformação em componentes principais é dada por*

$$\mathbf{X} \to \mathbf{W} = \mathbf{U}^{\mathsf{T}}(\mathbf{X} - \mu),$$

onde $\mathbf{U}$ é uma matriz ortogonal, que existe, conforme o Teorema da Decomposição Espectral, de tal modo que $\mathbf{U}^{\mathsf{T}}\mathbf{S}\mathbf{U} = \mathbf{\Lambda}$ é uma matriz diagonal formada pelos autovalores de $\mathbf{S_X}$: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0,$ já dispostos em ordem não-crescente, sendo todos não negativos, já que $\mathbf{S}$ (ou $\mathbf{R}$) é positiva semi-definida. Observe que o i-ésimo componente principal $\mathbf{w}_i$ é definido como o i-ésimo elemento da matriz $\mathbf{W}$, e é calculado por

$$\mathbf{w}_i = \mathbf{u}_i^{\mathsf{T}}(\mathbf{X} - \mu).$$

*Aqui, $\mathbf{u}_i$ é a i-ésima coluna de $\mathbf{U}$, e pode ser chamado de i-ésimo **vetor de pesos** do componente principal.*

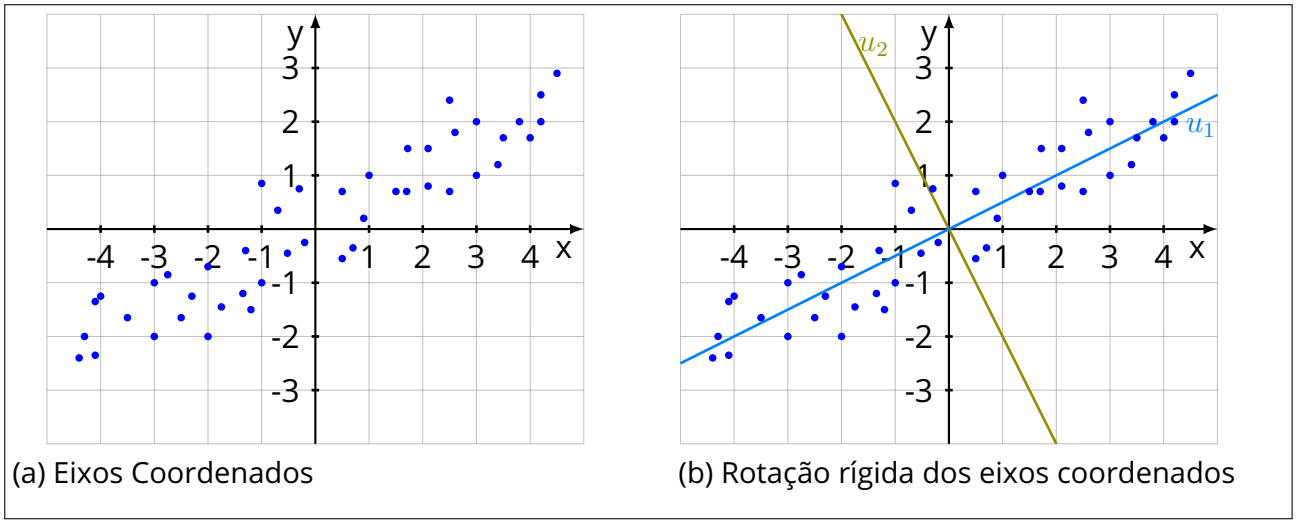
3.1 Uma abordagem geométrica dos Componentes Principais

A técnica de componentes principais é, por definição, uma combinação linear de p variáveis correlacionadas e tem como objetivo rotacionar o sistema de eixos

coordenados, posicionando os novos eixos no sentido da maior variabilidade dos dados. Assim, existe a possibilidade de se reduzir a dimensão do conjunto de dados, com a garantia de que a m-ésima parcela escolhida seja responsável por parte significativa da informação contida na matriz original.

Essa situação torna-se mais simples se ilustrada em duas dimensões. A Figura (1a) ilustra o comportamento de 50 indivíduos descritos por duas variáveis, fornecendo $\mathbf{X} = [\mathbf{x}_1 = \mathbf{x} \;\; \mathbf{x}_2 = \mathbf{y}]$.

Figura 1 – Ilustração da rotação rígida dos eixos coordenados definidos em duas variáveis correlacionadas x e y



(a) Eixos Coordenados (b) Rotação rígida dos eixos coordenados

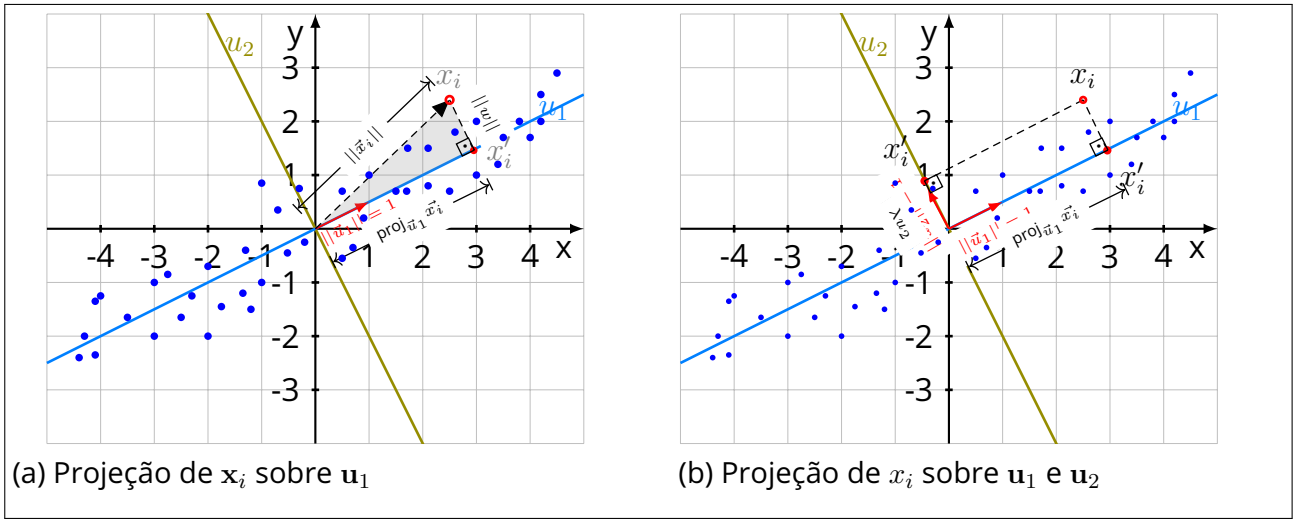
Fonte: autores (2024)

Claramente, esses indivíduos possuem entre si uma relação direta, pois a nuvem de pontos formada pela dispersão no plano apresenta um comportamento bem definido. Podemos inferir, ainda, que uma maior parcela da variação está no sentido do eixo das abscissas. Traçando as direções fornecidas pelos autovetores de $\mathbf{S_X}$, não é difícil concluir que os eixos das ordenadas e das abscissas não seriam a melhor representação desse conjunto.

Um vetor que passa no meio dessa nuvem de pontos seria preterível em relação aos eixos x e y. Podemos pensar ainda em uma rotação dos eixos coordenados, tal que $\mathbf{w}_1 = \alpha_1\mathbf{x} + \alpha_2\mathbf{y}$ e $\mathbf{w}_2 = \beta_1\mathbf{x} + \beta_2\mathbf{y}$ são resultados de uma combinação linear. Esse novo conjunto é gerado da rotação baseada nas direções $\mathbf{u}_1 = [\alpha_1, \alpha_2]$ e $\mathbf{u}_2 = [\beta_1, \beta_2]$. Observa-se que $\mathbf{u}_1$ é a direção que possui a maior parte da variância do conjunto quando comparado com $\mathbf{u}_2$. Como o objetivo da **PCA** é determinar os componentes

que concentram o máximo da variância original dos dados, estamos procurando exatamente por essa direção (variável latente).

Figura 2 – Ilustração das projeções Escalar de x_i e Vetorial de $\mathbf{x}_i$



(a) Projeção de $\mathbf{x}_i$ sobre $\mathbf{u}_1$

(b) Projeção de x_i sobre $\mathbf{u}_1$ e $\mathbf{u}_2$

Fonte: autores (2024)

Na Figura 2a, tomando o ponto x_i, essa distância é representada pelo segmento w e, dado que $||\mathbf{x}_i||^2 = ||\lambda\mathbf{u}_1||^2 + ||\mathbf{w}||^2$ [6], reduzindo w ao menor possível, equivale a determinar a direção na qual a projeção terá máxima variância (essa abordagem é similar à feita em Hotelling (1933)). Sob essa perspectiva, os pontos originais serão projetados sobre um vetor que maximiza a distância entre x'_i e a origem. De acordo com a Figura 2b, o vetor $\lambda\mathbf{u}_1$ é notavelmente a direção de maior importância, ou seja, a direção que capta a maior variabilidade dos dados originais. Pensando em redução da dimensão original, poderíamos inclusive descartar a direção $\lambda\mathbf{u}_2$. Mas o que garante que a direção $\lambda\mathbf{u}_1$ seja a melhor entre as opções?

Encontrar a direção que maximiza a variância dos dados originais corresponde a resolver um problema de otimização. Seja $\mathbf{u}_k$ a direção sobre a variância máxima e tomemos esse vetor como unitário[7] como restrição, ou seja, $||\mathbf{u}_k|| = 1 = \mathbf{u}_k^\mathsf{T}\mathbf{u}_k$. Olhando para a observação x_i, sua projeção escalar (comp) em $\mathbf{u}_k$ é dada por $||\mathbf{x}'_i||$, isto é, $||\mathbf{x}'_i|| = \text{comp}_{\mathbf{u}_k} x_i = \frac{\langle \mathbf{x}_i, \mathbf{u}_k \rangle}{||\mathbf{u}_k||} = \langle \mathbf{x_i}, \mathbf{u}_k \rangle = \mathbf{x}_i^\mathsf{T}\mathbf{u}_k$[8]. No caso da Figura 2b, entre as direções $\mathbf{u}_1$ e $\mathbf{u}_2$, não

[6]Aqui, definimos a *norma Euclidiana*. Dado um vetor (ou uma matriz coluna) $\mathbf{x}^\mathsf{T} = [x_1, \ldots, x_n] \in \mathbb{R}^n$, a sua norma $||\mathbf{x}||$ é definida por $||\mathbf{x}|| = \sqrt{x_1^2 + \cdots + x_n^2}$.

[7]A ideia de tratar $\mathbf{u}$ como um vetor unitário é aplicada para impor uma restrição finita ao problema de maximização.

[8]A expressão $\langle \cdot, \cdot \rangle$ motiva a definição de *produto interno usual*. Ela é uma função binária definida entre dois vetores, $\mathbf{u}^\mathsf{T} = [u_1, u_2, \ldots, u_n]$ e $\mathbf{v}^\mathsf{T} = [v_1, v_2, \ldots, v_n]$ em $\mathbb{R}^n$, que fornece um número real (escalar) como resultado do somatório das coordenadas.

é difícil de ver que a primeira é a projeção que maximiza a variância, pois $\text{comp}_{\mathbf{u}_1} x_i > \text{comp}_{\mathbf{u}_2} x_i$, e o mesmo se aplica a maioria dos demais pontos em azul da figura.

Assim, queremos encontrar $\mathbf{u}_k$ tal que $\text{Var}(\text{comp}_{\mathbf{u}_k} x_i) = \text{Var}(\mathbf{x}_i^\mathsf{T}\mathbf{u}_k) = \text{Var}(\mathbf{w}_i)$ seja máxima. Portanto, como queremos maximizar a variância de todos os pontos de $\mathbf{X}$, e não apenas de x_i, nosso problema de maximização será o de maximizar a variância

$$\text{Var}(\mathbf{X}^\mathsf{T}\mathbf{u}_k) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\mathsf{T}\mathbf{u}_k)^2, \tag{11}$$

e, restringindo $\mathbf{u}_k$ como um vetor unitário qualquer, a expressão (11) pode ser reescrita como

$$\text{Var}(\mathbf{X}^\mathsf{T}\mathbf{u}_k) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\mathsf{T}\mathbf{u}_k)^2 = \frac{\mathbf{u}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{u}}{n} = \mathbf{u}^\mathsf{T}\frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{n}\mathbf{u} = \mathbf{u}^\mathsf{T}\mathbf{S}\mathbf{u}, \tag{12}$$

tal que temos o seguinte problema de maximização de (12):

$$\begin{cases} \max\limits_{\mathbf{u}} & (\mathbf{u}^\mathsf{T}\mathbf{S}\mathbf{u}) \\ \text{sujeito a:} & ||\mathbf{u}|| = \mathbf{u}^\mathsf{T}\mathbf{u} = 1. \end{cases} \tag{13}$$

Com a aplicação do próximo teorema, é possível identificar os pontos extremos de (13). Em seu enunciado, usaremos a notação ∇ para identificar o vetor gradiente de uma função real $f : U \subset \mathbb{R}^p \to \mathbb{R}$, definido por $\nabla f(\mathbf{u}) = (\frac{\partial f}{\partial e_1}(u), \cdots, \frac{\partial f}{\partial e_p}(u))$.

Teorema 2. ***(Método dos multiplicadores de Lagrange)***[9] *Seja $U \subset \mathbb{R}^p$ um aberto e suponhamos que $f, g_0, g_1, \ldots, g_h : U \subset \mathbb{R}^p \to \mathbb{R}$ sejam funções reais de classe C^1 no aberto U. Para que um ponto $\mathbf{x} \in U$ seja um ponto de máximo ou de mínimo local de f satisfazendo as restrições $g_0(x) = 0, g_1(0) = 0, \ldots, g_h(0) = 0$ é necessário que a matriz $\mathbf{D} = (d_{ij})_{(p\times h)}$, com $d_{ij} = (\frac{\partial g_j}{\partial e_i})$, tenha posto h e que existam constantes $\alpha_0, \alpha_1, \cdots \alpha_h$ tais que*

$$\nabla f(\mathbf{x}_0) = \alpha_0\nabla g_0(\mathbf{x}_0) + \alpha_1\nabla g_1(\mathbf{x}_0) + \cdots + \alpha_h\nabla g_h(\mathbf{x}_0). \tag{14}$$

No teorema acima, as constantes $\alpha_1, \cdots, \alpha_h$ de (14) são chamadas de multiplicadores de Lagrange. Como um corolário deste teorema, provaremos que as

[9]Uma demonstração pode ser encontrada em Bartle (1983).

componentes principais de variáveis aleatórias são de fato as de máxima variância.[10] sobre a matriz de variância amostral $\mathbf{S}$.

Teorema 3. ***(Componentes Principais de Variáveis Aleatórias)*** *Assuma que posto* $(\mathbf{S}_X) = p$. *Então os* p *componentes principais de uma variável aleatória multivariada* $\mathbf{X} \in \mathbb{R}^p$, *denotados por* $\mathbf{w}_j$ *para* $j = 1, 2, \ldots, p$, *são dados por*

$$\mathbf{w}_j = \mathbf{u}_j^{\mathsf{T}} \mathbf{X},$$

onde $\mathbf{u} \in \mathbb{R}^p$ *e* $\{\mathbf{u}_j\}_{j=1}^p$ *são os* p *autovetores de* $\mathbf{S}_X$ *associados aos maiores autovalores* λ_j. *Além disso,* $\lambda_j = Var(\mathbf{w}_j)$ *para* $j = 1, 2, \ldots, p$.

Demonstração. Como visto em (13) precisamos achar $\mathbf{u}$ que maximiza o problema

$$\begin{cases} f(\mathbf{u}) = & \mathbf{u}^{\mathsf{T}} \mathbf{S} \mathbf{u} \\ \text{sujeito a:} & g_0(\mathbf{u}) = \mathbf{u}^{\mathsf{T}} \mathbf{u} - 1 = 0 \end{cases}$$

As derivadas parciais de f são

$$\frac{\partial f}{\partial e_i}(u) = \lim_{h \to 0} \frac{(u + he_i)^{\mathsf{T}} S (u + he_i) - u^{\mathsf{T}} S u}{h} = u^{\mathsf{T}} S e_i + e_i^{\mathsf{T}} S u = (e_i^{\mathsf{T}} S u)^{\mathsf{T}} + e_i^{\mathsf{T}} S u = 2 e_i^{\mathsf{T}} S u,$$

para $i = 1, \ldots, p$. Portanto, $\nabla f(\mathbf{u}) = 2\mathbf{S}\mathbf{u}$. Por sua vez, as derivadas parciais de g_0 são dadas por

$$\frac{\partial g_0}{\partial e_i}(u) = \lim_{h \to 0} \frac{(u + he_i)^{\mathsf{T}} (u + he_i) - u^{\mathsf{T}} u}{h} = e_i^{\mathsf{T}} u + u^{\mathsf{T}} e_i = 2 e_i^{\mathsf{T}} u.$$

para $i = 1, \ldots, p$, e assim,

$$\nabla g_0(\mathbf{u}) = 2\mathbf{u}.$$

Pelo Teorema de Lagrange, precisamos encontrar soluções não nulas de $\mathbf{u}$ (com isso garantimos que o posto de $\mathbf{D} = (2\mathbf{u})$ seja 1) tais que $\nabla f(\mathbf{u}) = \alpha_0 \nabla g_0(\mathbf{u})$,, ou seja, $2\mathbf{S}\mathbf{u} = 2\alpha_0 \mathbf{u} \Leftrightarrow (\mathbf{S} - \alpha_0 I)\mathbf{u} = 0$, logo, as soluções procuradas necessariamente são autovetores. Tomando $\mathbf{u} = \mathbf{u}_1$, como sendo um autovalor associado ao maior autovalor dentre todos, garantimos uma solução com a maior variância possível para

[10] Em Vidal et al. (2005) outras provas com abordagens distintas podem ser encontradas. A prova aqui desenvolvida foi baseada em Jolliffe (2002) e Vidal et al. (2005).

nosso problema. Cabe ressaltar aqui que a escolha não é única (dependendo da multiplicidade geométrica do autovalor poderemos ter várias direções independentes para escolher), porém, como estamos interessados nas direções de maior variância, qualquer um destes autovetores associados a este maior autovalor serão igualmente considerados nas próximas etapas.

Para obtermos a direção com a segunda maior variância, que seja independente da primeira, o problema de maximização passa a ser dado por

$$\begin{cases} f(\mathbf{u}) = & \mathbf{u}^{\mathsf{T}}\mathbf{S}\mathbf{u} \\ \text{sujeito a:} & \begin{cases} g_0(\mathbf{u}) = \mathbf{u}^{\mathsf{T}}\mathbf{u} - 1 = 0 \\ g_1(\mathbf{u}) = \mathbf{u}^{\mathsf{T}}\mathbf{u_1} = 0 \end{cases} \end{cases}$$

considerando que $\mathbf{D} = \begin{bmatrix} \frac{\partial g_0}{\partial e_1} & \frac{\partial g_1}{\partial e_1} \\ \frac{\partial g_0}{\partial e_2} & \frac{\partial g_1}{\partial e_2} \\ \vdots & \vdots \\ \frac{\partial g_0}{\partial e_p} & \frac{\partial g_1}{\partial e_p} \end{bmatrix} = \begin{bmatrix} \vdots & \vdots \\ \mathbf{u} & \mathbf{u}_1 \\ \vdots & \vdots \end{bmatrix}$ terá posto 2, pois queremos que $\mathbf{u}$ e $\mathbf{u}_1$ sejam linearmente independentes, precisamos ter constantes α_0, α_1 como no Teorema 2, tais que $\nabla f(\mathbf{u}) = \alpha_0 \nabla g_0(\mathbf{u}) + \alpha_1 \nabla g_1(\mathbf{u})$. Calculando os gradientes como acima, obtemos

$$2\mathbf{S}\mathbf{u} = 2\alpha_0\mathbf{u} + \alpha_1\mathbf{u}_1.$$

Aplicando $\mathbf{u} = \mathbf{u}_1^{\mathsf{T}}$, à esquerda, em ambos os lados, pelas restrições $g_0(\mathbf{u}) = 1$ e $g_1(\mathbf{u}) = 0$ e ainda, por $\mathbf{u}_1^{\mathsf{T}}\mathbf{S}\mathbf{u} = (\mathbf{u}^{\mathsf{T}}\mathbf{S}\mathbf{u}_1)^{\mathsf{T}} = (\mathbf{u}^{\mathsf{T}}\lambda_1\mathbf{u}_1)^{\mathsf{T}} = \lambda_1\mathbf{u}^{\mathsf{T}}\mathbf{u}_1 = \lambda_1.0 = 0$ obtemos, $0 = 2\alpha_0.0 + \alpha_1.1$, assim, necessariamente $\alpha_1 = 0$. Novamente obtemos $\mathbf{S}\mathbf{u} = \alpha_0\mathbf{u}$, o que implica que $\mathbf{u}$ seja um autovetor associado a um autovalor α_0. Desde que $\mathbf{u} = \mathbf{u}_2$ fornece $f(\mathbf{u}_2) = \lambda_2 \leq \lambda_1 = f(\mathbf{u}_1)$, segue-se que o autovetor $\mathbf{u} = \mathbf{u}_2$, é a solução procurada (ou ao menos uma delas, caso a multiplicidade de $\lambda_2 > 1$). Seguindo desse modo, podemos considerar os autovetores $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{i-1}$ como já obtidos de modo a satisfazerem $f(\mathbf{u}_{i-1}) = \lambda_{i-1} \leq \cdots \leq \lambda_1 = f(\mathbf{u}_1)$, e assim precisamos resolver o problema de maximização de

$$\begin{cases} f(\mathbf{u}) = & \mathbf{u}^\mathsf{T}\mathbf{S}\mathbf{u} \\ \\ \text{com restrições:} & \begin{cases} g_0(\mathbf{u}) = \mathbf{u}^\mathsf{T}\mathbf{u} - 1 = 0 \\ g_1(\mathbf{u}) = \mathbf{u}^\mathsf{T}\mathbf{u_1} = 0 \\ \quad \vdots \\ g_{i-1}(\mathbf{u}) = \mathbf{u}^\mathsf{T}\mathbf{u_{i-1}} = 0 \end{cases} \end{cases}$$

Novamente aqui a matriz $\mathbf{D} = \begin{bmatrix} \frac{\partial g_0}{\partial e_1} & \cdots & \frac{\partial g_{i-1}}{\partial e_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_0}{\partial e_p} & \cdots & \frac{\partial g_{i-1}}{\partial e_p} \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \mathbf{u} & \mathbf{u}_1 & \ldots & \mathbf{u}_{i-1} \\ \vdots & \vdots & & \vdots \end{bmatrix}$ terá posto $i-1$, por estarmos procurando soluções $\mathbf{u}$ que sejam linearmente independentes com os já independentes vetores $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$.

Assim, estamos em busca de constantes $\alpha_0, \ldots, \alpha_{i-1}$, tais que a equação do Teorema de Lagrange tenha solução. Calculando os gradientes de $f, g_0, \ldots, g_{i-1}$, como acima, concluímos que a solução $\mathbf{u}$ procurada deve satisfazer:

$$2\mathbf{S}\mathbf{u} = 2\alpha_0\mathbf{u} + \alpha_1\mathbf{u}_1 + \cdots + \alpha_{i-1}\mathbf{u}_{i-1}. \tag{15}$$

Para cada $k = 1, \ldots, i-1$, multiplica-se à esquerda, em ambos os lados da Expressão (15) por $\mathbf{u}_k^\mathsf{T}$ obtemos

$$\mathbf{u}_k^\mathsf{T} S\mathbf{u} = 2\alpha_0 g_k(\mathbf{u}) + \alpha_1\mathbf{u}_k^\mathsf{T}\mathbf{u}_1 + \cdots + \alpha_k g_0(\mathbf{u}_k) + \cdots + \alpha_{i-1}\mathbf{u}_k^\mathsf{T}\mathbf{u}_{i-1},$$

de onde se conclui que necessariamente $\alpha_k = 0$, para todo $k = 1, \cdots, i-1$. Resta na Expressão 15 que $\mathbf{S}\mathbf{u} = \alpha_0\mathbf{u}$, tal que o autovetor $\mathbf{u} = \mathbf{u}_i$ associado ao autovalor λ_i, é a solução. Desde que isso ocorre para todo $i = 2, \cdots p$, conclui-se que uma base ortonormal de autovetores $\{\mathbf{u}_1, \cdots \mathbf{u}_p\}$ de $\mathbf{S_X}$, dispostos de forma que seus autovalores associados estejam em ordem não crescente fornecerão as direções de máxima varição nas restrições impostas. □

Podemos considerar os autovetores cujos autovalores sejam ”significativos”(não nulos, por exemplo, ou maiores que um certo valor que se considere adequado a um dado modelo preterido).

Neste sentido, podemos redefinir as componentes principais por meio do seguinte corolário.

Corolário 1. ***(Redefinindo os Componentes Principais de Variáveis Aleatórias)*** *Seja* $m \leq p$*. Assuma que posto* $(\mathbf{S}_X) \geq m$*. Então os primeiros* m *componentes principais de uma variável aleatória multivariada* $\mathbf{X} \in \mathbb{R}^p$ *são dados por* $\mathbf{w}_j = \mathbf{u}_j^{\mathsf{T}}\mathbf{X}$*, onde* $\mathbf{u} \in \mathbb{R}^p$ *e* $\{\mathbf{u}_j\}_{j=1}^m$ *são os* p *autovetores de* $\mathbf{S}_X$ *associados aos maiores autovalores* $\lambda_j > 0$*.*

Pelo Corolário 1, a equação (10) será truncada em um certo número de termos, como $m < p$, tal que

$$\mathbf{w}_j = u_{j1}\mathbf{x}_1 + u_{j2}\mathbf{x}_2 + \cdots + u_{jm}\mathbf{x}_m + \mathbf{v}_j,$$

onde, $\mathbf{v}$

$$\mathbf{v}_j = u_{j\,m+1}\mathbf{x}_{m+1} + u_{j\,m+1}\mathbf{x}_{m+1} + \cdots + u_{jp}\mathbf{x}_p,$$

é interpretado como o resíduo de menor variância (Jöreskog (1979)).

Uma observação importante que podemos obter dos componentes principais é que para $k \neq j$, temos,

$$\text{cov}(\mathbf{u}_k^{\mathsf{T}}\mathbf{X}, \mathbf{u}_j^{\mathsf{T}}\mathbf{X}) = \mathbf{u}_k^{\mathsf{T}}\mathbf{S}_X\mathbf{u}_j = \mathbf{u}_j^{\mathsf{T}}\mathbf{S}_X\mathbf{u}_k = \mathbf{u}_j^{\mathsf{T}}\lambda_k\mathbf{u}_k^{\mathsf{T}} = \lambda_k\mathbf{u}_j^{\mathsf{T}}\mathbf{u}_k = \lambda_k\mathbf{u}_k^{\mathsf{T}}\mathbf{u}_j = 0.$$

o que indica a ortogonalidade entre os componentes principais, isto é, ausência de correlação entre os mesmos.

4 ANÁLISE DOS RESULTADOS

No âmbito nacional, as pesquisas em economia que compartilharam dessa técnica utilizaram o método de componentes principais com um meio e não como um fim. Em Kageyama & Leone (1999), buscou-se tipificar um conjunto de municípios no estado de São Paulo (SP) a partir de suas principais características sociais e econômicas. Segundo as autoras, esse conjunto de municípios paulistas possuía cinco regiões relativamente homogêneas: rural muito pobre, rural pobre, intermediária, urbano em expansão e urbano denso. Partindo dos componentes principais, essas

tipificações foram descritas em termos de renda, população e produção agrícola locais.

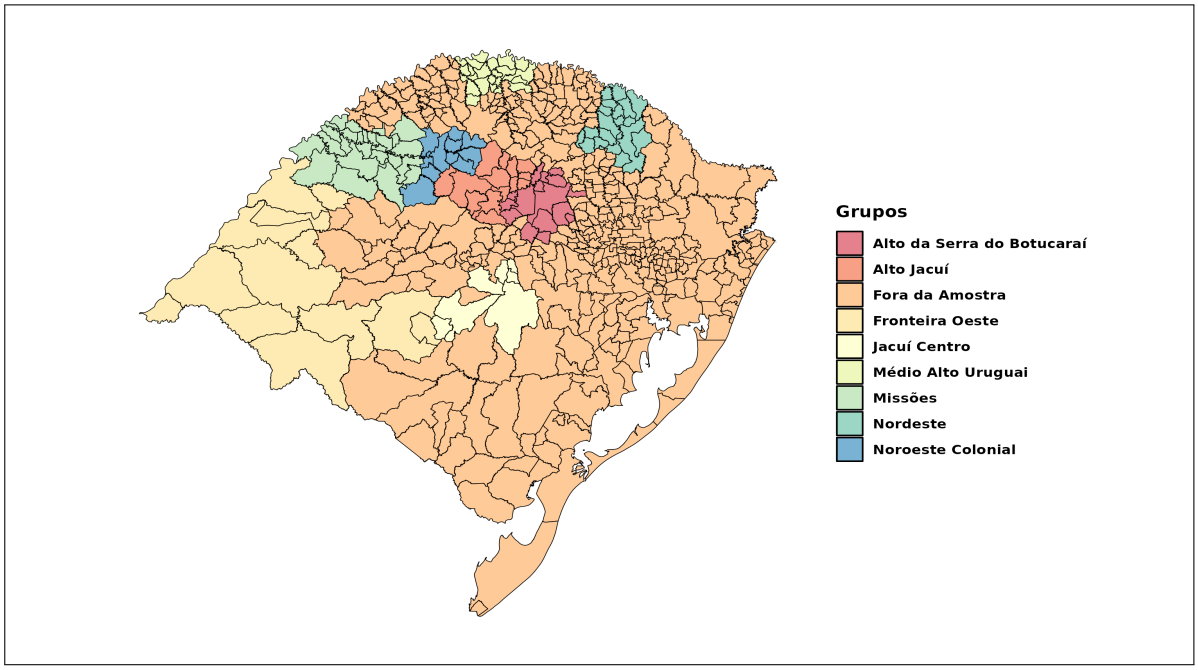
Essa mesma ideia foi igualmente aplicada para o Rio Grande do Sul. Schneider & Waquil (2001) buscaram a criação de grupos homogêneos entre os municípios do estado. Os autores classificaram o Rio Grande do Sul em cinco grupos de municípios, e acabaram por descartar a hipótese de que o estado podia ser dividido em duas partes, isto é, entre uma metade sul mais atrasada e o norte desenvolvido. Por outro lado, Freitas et al. (2007) abordaram o estado do Rio Grande do Sul com a ideia de explorar a existência de padrões determinados pelo grau de utilização de insumos agrícolas modernos nos estabelecimentos rurais para o Censo Agropecuário de 1995/96. O objetivo principal do artigo era agrupá-los conforme suas similaridades e relação com um conjunto de variáveis latentes obtidas a partir das variáveis originais utilizando componentes principais. Devido ao número de municípios e suas vocações, Poerschke & Junior (2020) optaram usar a base de dados do Censo Agropecuário 1995/96 apenas para os municípios parte dos COREDEs com dinâmica econômica ligada à agricultura e pecuária. A atualização do Censo Agropecuário pelo IBGE, e a inexistência de pesquisas sobre o assunto, motivou a aplicação do PCA deste artigo.

A tipificação dos COREDEs estabelecida pelo projeto Rumos 2015 identificou a existência de dinâmicas econômicas distintas. A questão da distribuição desigual das atividades econômicas, constitui a problemática das desigualdades regionais e seu enfrentamento será sempre um dos eixos norteadores da ação administrativa de todo Governo. Os COREDEs gaúchos foram classificados a partir de sua estrutura produtiva, levando em consideração a participação do setor de Serviços, da Indústria e da Agropecuária na formação da renda de cada espaço territorial político no estado. Os COREDEs com predominância do setor agropecuário aglutinam 127 municípios e representam 26% do total de municípios do estado do Rio Grande do Sul.

Entre os COREDEs predominantemente agropecuários, destacados na Figura 3 que segue, estão: Alto da Serra do Botucaraí (16 municípios); Alto Jacuí (14); Fronteira Oeste (13); Jacuí Centro (7); Médio Alto Uruguai (22); Missões (25); Nordeste (19) e Noroeste Colonial (11). Inicialmente, a análise exploratória dos 8 COREDEs agropecuários levou em consideração 17 variáveis baseadas em Freitas et al. (2007). Os autores utilizaram dados do Censo Agropecuário 1995/96 para estudar o estado

como um todo. Para o mesmo período e com base de dados semelhantes, Poerschke & Junior (2020) utilizaram um recorte voltado apenas aos municípios com dinâmica econômica ligada à agropecuária.

Figura 3 – COREDEs Agropecuários do Rio Grande do Sul



Fonte: Elaboração própria com dados do IBGE (2018)

Quadro 1 – Variáveis utilizadas*

Sigla	Nome da Variável	Referência	Unidade de Medida	Fonte
fin_veg	Financiamento (Prod. Vegetal)	Tabela 6895	N. de Estabelecimentos	IBGE
fin_pec	Financiamento (Prod. Pecuária)	Tabela 6895	N. de Estabelecimentos	IBGE
ass_veg	Assistência Técnica (Prod. Vegetal)	Tabela 6844	N. de Estabelecimentos	IBGE
ass_pec	Assistência Técnica (Prod. Pecuária)	Tabela 6844	N. de Estabelecimentos	IBGE
colhe	Colheitadeiras	Tabela 6874	Unidades	IBGE
trat	Tratores	Tabela 6869	Unidades	IBGE
gado	Rebanho Bovino	Tabela 6907	Rebanho	IBGE
pea	População Economicamente Ativa	Tabela 6887	Pessoas	IBGE
pop	População Residente	Tabela 6579	Pessoas	IBGE
rec_veg	Receitas com Lavouras	Tabela 6897	Mil R$	IBGE
val_pec	Valor da Produção Pecuária	Tabela 6898	Mil R$	IBGE
val_veg	Valor da Produção Vegetal	Tabela 6897	Mil R$	IBGE
irriga	Irrigação	Tabela 6857	N. de Estabelecimentos	IBGE
adubo	Adubação	Tabela 6847	N. de Estabelecimentos	IBGE
area_rela	Área Explorada/Área Total	15761**	Área (km^2)	IBGE
area_exp	Área Total Explorada	Tabela 6878	Área (ha)	IBGE
idese	IDESE	Bloco Renda	Numero Índice	FEE***

* - Os dados são referentes ao Censo Agropecuário 2017, exceto pela Área Total dos Municípios e IDESE Bloco Renda
** - Áreas Territoriais (Instituto Brasileiro de Geografia e Estatística)
*** - Fundação de Economia e Estatística (FEE)

Esses trabalhos foram importantes como um guia, pois são passíveis de fornecer um comparativo em dois tempos, isto é, em certa medida é possível

estabelecer retratos da agropecuária no estado por meio dos dados censitários coletados pelo IBGE ao longo do tempo. Mesmo que os objetivos e os dados não sejam exatamente os mesmos, é importante comparar os resultados e assim ilustrar a aplicação do método de componentes principais.

O tratamento inicial dos dados seguiu os mesmos passos. Com os dados coletados, as variáveis foram centralizadas e padronizadas. Após algumas rodadas de decomposições da matriz de correlações, algumas variáveis da amostra foram descartadas, uma vez que ficaram isoladas nos autovetores estimados. Ao fim e ao cabo, restaram 15 variáveis, sendo todas coletadas em IBGE (2018) e listadas no Quadro 1[11]. O Quadro 1 que segue resume as variáveis consideradas e seus códigos que doravante serão utilizados, bem como apresenta as referências de suas tabelas/fonte no banco de dados do IBGE para o Censo Agropecuário 2017. Ainda, as variáveis Área Relativa, que mede a relação entre área explorada e a área total do município, bem como Índice de Desenvolvimento Socioeconômico (IDESE) dos municípios do Rio Grande do Sul, foram descartadas. Pelo princípio da parcimônia, a adição das mesmas não corroborou com acréscimo significativo de informação à pesquisa.

Após, estimamos a matriz de variância-covariâncias ($\mathbf{S}$) para o conjunto restante de dados centralizados e padronizados[12]. O próximo passo, dado que as matrizes supracitadas são simétricas, foi a estimação dos autovalores e seus autovetores associados pela Decomposição Espectral de Matrizes sobre a matriz de correlações ($\mathbf{R}$).

Observe que a média de uma variável aleatória $\mathbf{x}_k$ é dada por (2), tal que $\mathbb{E}[\mathbf{x}] = \mu$. Tomando o desvio da média de uma variável $\mathbf{x}_k$ temos $\mathbf{x}_k^* = \mathbf{x}_k - \mu_k$. Com base em (5), podemos calcular o desvio-padrão de $\mathbf{x}_k$, pois se a variância

$$s_{kk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2$$

[11]Os dados originais podem ser encontrados no seguinte repositório: https://github.com/faecors/ArtigoPCA. Os dados utilizados na pesquisa refletem a proporção do município em relação ao total da amostra, seja a variável medida em unidades, pessoas, unidades monetárias, etc.

[12]Quanto ao polinômio característico da matriz de variâncias-covariâncias não padronizados, os autovalores possíveis apresentaram valores menores que a unidade. Esse comportamento indica a necessidade da padronização do conjunto, visto que, conforme o Quadro 1, as escalas de medidas dos dados são muito diferentes.

ou $\text{var}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \mu)^2]$ o desvio-padrão será $\sigma(\mathbf{x}) = \sqrt{\text{var}(\mathbf{x})} = \sqrt{\mathbb{E}[(\mathbf{x} - \mu)^2]}$. De outro lado, a correlação é calculada pela expressão (6), considerando a $s_{xy} = \text{cov}(\mathbf{x}, \mathbf{y}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))]$, temos

$$r_{xy} = \frac{\mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))]}{\sigma_x \sigma_y}. \tag{16}$$

Mas observe que a variância, quando estimada sobre variáveis padronizadas, passa a ser dada por

$$s_{xy} = \mathbb{E}\left[\left(\frac{\mathbf{x} - \mathbb{E}(\mathbf{x})}{\sigma_x} - 0\right) \times \left(\frac{\mathbf{y} - \mathbb{E}(\mathbf{x})}{\sigma_x} - 0\right)\right]$$

$$= \frac{\mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))]}{\sigma_x \sigma_y},$$

o que é igual à expressão da correlação acima (16), isto é, dada uma matriz $\mathbf{X}$, centralizada e padronizada, quando estimamos sua matriz de variâncias-covariâncias $\mathbf{S_X}$, ela equivale à matriz de correlações $\mathbf{R_X}$.

Tabela 1 – Correlação das Variáveis com os Autovetores (Matriz $\mathbf{R}_X$)

Autovalores	$\lambda_1^R = 8,98$	$\lambda_2^R = 2,62$	$\lambda_3^R = 1,47$
	Autovetor 1	**Autovetor 2**	**Autovetor 3**
val_veg	0,277	0,000	0,421
fin_veg	0,143	-0,507	-0,008
rec_veg	0,277	0,019	0,412
ass_veg	0,149	-0,509	-0,201
fin_pec	0,239	0,147	-0,418
val_pec	0,289	0,245	-0,135
gado	0,277	0,303	-0,119
ass_pec	0,284	0,203	-0,280
adubo	0,214	-0,390	-0,328
colhe	0,267	-0,169	0,353
trat	0,308	-0,124	0,169
pea	0,311	-0,087	-0,209
pop	0,280	0,020	0,141
area_exp	0,297	0,249	0,028
irriga	0,174	0,053	-0,042
area_rela	—	—	—
idese	—	—	—

Fonte: autores (2024)

Os autovetores estimados pelos maiores autovalores da matriz de correlações $\mathbf{R_X}$ estão dispostos na Tabela 1. Juntos, os três primeiros autovalores responderam

por cerca de 87% da variância do conjunto original de dados[13]. Esse resultado está em linha com Freitas et al. (2007) e Poerschke & Junior (2020). Os primeiros, com dados do Censo Agropecuário de 1995/96, obtiveram cinco autovetores, cuja variância explicada acumulada chegou a 82% do total do conjunto. Restringindo a análise para os COREDEs agropecuários, Poerschke & Junior (2020) estimaram quatro autovetores, que reuniram por volta de 83% da variância do conjunto original.

Passando para a relação estabelecida entre as variáveis originais com os autovetores, com base na Tabela 1, é possível verificar que todo o conjunto das variáveis tem uma relação diretamente proporcional com o **Autovetor 1**. As variáveis Número de Tratores (0,308) e População Economicamente Ativa (0,311) apresentam as maiores magnitudes. Ainda, as variáveis Área Explorada (0,297) e População (0,280) demonstram correlação com o Autovetor 1 em valores medianos.

As variáveis financeiras relacionadas à agricultura denotam maior correlação com o Autovetor 3. As maiores correlações entre variáveis e autovetores verificam-se no Valor da Produção Vegetal (0,421), as Receitas da Produção Vegetal (0,412), seguidos do Número de Colheitadeiras (0,353) e Número de Tratores (0,169). É importante ressaltar que as variáveis relacionadas à produção de gado, Financiamento da Pecuária (-0,418), Assistência Técnica à Pecuária (-0,280), Valor da Produção Pecuária (-0,135) e Rebanho Bovino (-0,119) obtiveram sinais opostos. Esse comportamento indica que o **Autovetor 3** representa, em sua maioria, a variabilidade das **variáveis financeiras** que mantêm relação com a produção agrícola e, em especial, com as intensivas no uso de máquinas e implementos agrícolas.

As variáveis relacionadas à produção pecuária tiveram maior relação com o Autovetor 2. As variáveis com relação direta foram o Rebanho Bovino (0,303), o Valor da Produção Pecuária (0,245) e o Financiamento da Pecuária (0,147). Já as variáveis financeiras ligadas à produção agrícola obtiveram magnitude significativa, mas com sinais opostos. Isso indica que o **Autovetor 2** tem uma relação direta com as variáveis ligadas à produção **pecuária**.

[13]A proporção explicada da variância original é a soma dos autovalores dos componentes retidos dividido pelo traço da matriz no qual os autovalores foram extraídos: Total Explicado $= \frac{13,07}{15} = 0,8715$. Cabe ressaltar que outras variáveis, tais como IDESE e a parcela da Área Explorada em relação à área total do município, foram inseridas em outros modelos. Contudo, as diversas combinações com a adição, ou combinação dessas, às demais, culminou em resultados inferiores de variância explicada, embora o número de autovetores fosse sempre o mesmo em todos os casos testados.

Nesse sentido, os resultados são animadores, mas carecem de um aprofundamento da análise. Por exemplo, uma rotação ortogonal dos autovetores poderia ajudar. A utilização do método Varimax, que procura maximizar a dispersão das cargas dentro dos autovetores, seria uma opção, pois, carregando um número menor de variáveis altamente correlacionadas em cada autovetor, resultaria em grupos de autovetores mais interpretáveis. O Varimax procura evitar a presença de muitas variáveis significativas em um único fator, como ocorreu no Autovetor 1. Isso pode possibilitar uma maior discussão da relação do conjunto de variáveis com um autovetor.

Agora que foram determinados que três autovetores são suficientes para explicar a maior parte da variância do conjunto, podemos usá-los para outras análises possíveis. Por exemplo, podemos tentar compactar as linhas das matrizes utilizando algum método de agrupamento baseado em três autovetores e não em 15 variáveis.

Se tomarmos a ideia de que estamos testando a similaridade, o método de Ward[14] pode ser uma opção, uma vez que ele visa gerar grupos o mais homogêneo possível. Com isso, estaríamos testando a hipótese lançada na introdução do texto, isto é, seriam esses grupos homogêneos? Então, usando os três autovetores[15], e estimando a distância euclidiana entre esses municípios, corroborando a hipótese, existem pelo menos quatro grupos que podem ser formados como conjunto de municípios parte dos COREDEs predominantemente agropecuários.

Esses grupos diferem em magnitude, sendo dois de tamanho reduzido quando comparados aos demais. Por exemplo, o Grupo 2[16] reúne Alegrete (3), Cachoeira do Sul (15), Rosário do Sul (87), Santana do Livramento (95), São Gabriel (100) e Uruguaiana (121). Todos os municípios fazem parte do COREDE Fronteira Oeste, exceto Cachoeira do Sul, que pertence ao COREDE Jacuí Centro. Os municípios ali agrupados possuem uma relação forte com o Autovetor 1, bem como se relacionam, em sua maioria, de maneira igualmente direta com o Autovetor 2. Para os dados de 1995/96, Poerschke e Moreira Junior (2020) encontraram um padrão de agrupamento semelhante, sendo

[14]Como o objetivo não é discutir a técnica de agrupamentos, apenas procedemos a um método aglomerativo que pode ser visto em Ward (1963).

[15]Os municípios e suas novas coordenadas segundo os componentes principais estão no Anexo B.

[16]A lista completa dos agrupamentos está no Anexo B. Em virtude do espaço demandado, os novos eixos rotacionados foram omitidos. Contudo, caso interesse ao leitor, eles estão disponíveis para acesso em: https://github.com/faecors/ArtigoPCA/blob/main/Componentes.txt.

o menor dos grupos, formado por Alegrete, Cachoeira do Sul, Itaqui, Rosário do Sul, Santana do Livramento, São Borja, São Gabriel e Uruguaiana. Esse grupo se manteve ao longo tempo, pois é muito semelhante ao Grupo 2 aqui estimado - São Borja e Itaqui aparecem no Grupo 4.

Quadro 2 - Escores dos municípios do Grupo 2

Código	Município	Componente 1	Componente 2	Componente 3
3	Alegrete	15,39	5,23	-2,64
15	Cachoeira do Sul	10,94	-3,24	1,47
87	Rosário do Sul	6,60	3,46	-1,47
95	Santana do Livramento	11,57	4,71	-4,82
100	São Gabriel	9,98	0,79	1,46
121	Uruguaiana	9,59	2,78	3,53

Fonte: autores (2024)

O Grupo 3 é composto por 34 municípios, e reúne Água Santa (1), Ajuricaba (2), Alpestre (4), Augusto Pestana (7), Barracão (8), Barros Cassal (10), Bossoroca (13), Cacique Doble (16), Caiçara (18), Cerro Largo (24), Entre-Ijuís (32), Erval Seco (33), Fontoura Xavier (36), Frederico Westphalen (38), Guarani das Missões (43), Ibiaçá (44), Ibiraiaras (45), Lagoão (55), Machadinho (58), Não-Me-Toque (64), Novo Cabrais (68), Palmitinho (71), Panambi (72), Paraíso do Sul (73), Planalto (78), Porto Xavier (79), Roque Gonzales (86), Santo Antônio das Missões (97), São José do Ouro (103), São Paulo das Missões (107), Seberi (110), Tapejara (114), Vicente Dutra (122) e Victor Graeff (22). Em termos de similaridade, é menos coeso, uma vez que abriga municípios de COREDEs diferentes. Em contraste ao Grupo 2, o Grupo 3 possui uma relação inversa com os Autovetores 1 e 2.

Note que no Quadro 3, quando observamos as médias dos escores entre os municípios de cada grupo, podemos inferir que os Grupos 1, 2 e 4 se relacionam de maneira positiva com o Autovetor 2, que mede a atividade pecuária. Em especial, o Grupo 4 (2,86) e o Grupo 2 (2,29) obtiveram valores significativos quando comparados aos demais. Esses mesmos grupos obtiveram os maiores escores em relação ao Autovetor 1. De outro lado, o Grupo 3 foi o único que apresentou relação inversa com o Autovetor 2 (Pecuária).

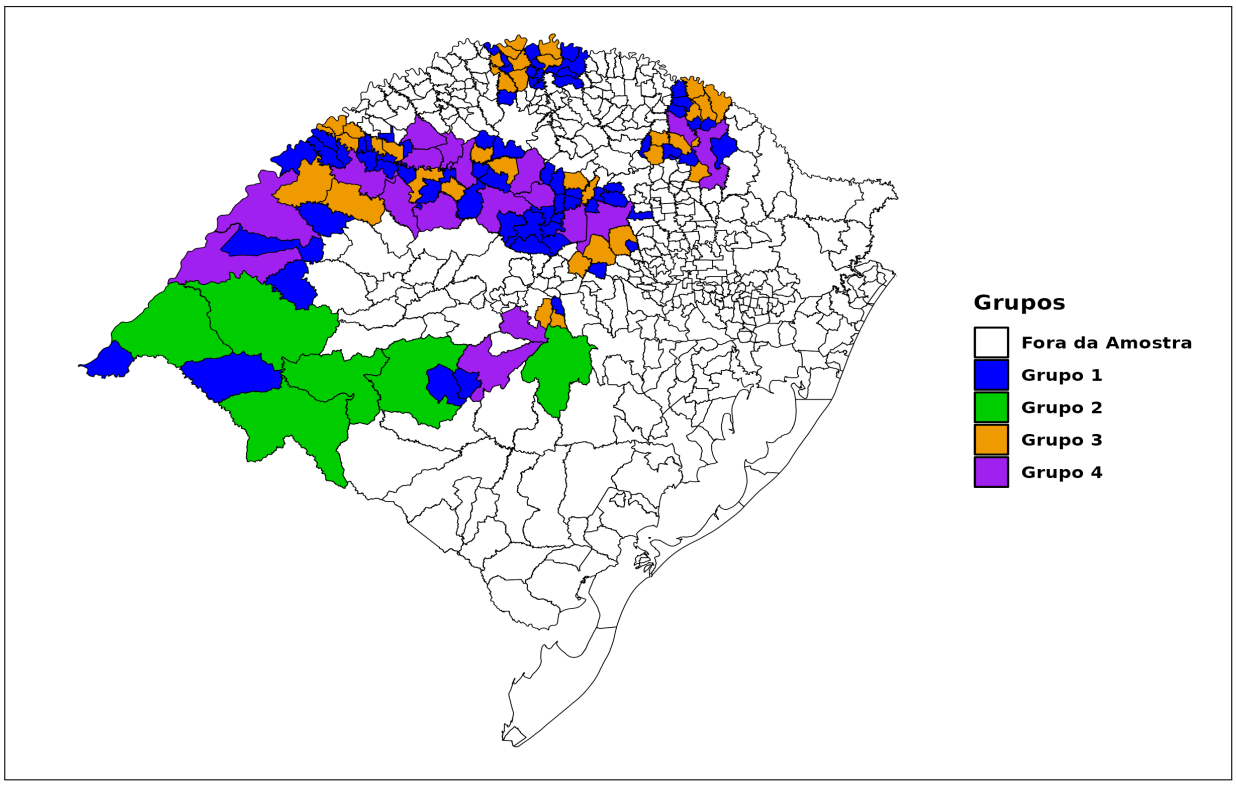
Quadro 3 - Média dos Escores dos Municípios de cada Grupo

Municípios Agrupados	Grupo	Autovetor 1	Autovetor 2	Autovetor 3
69	1	-1,62	0,71	0,15
6	2	10,68	2,29	-0,41
34	3	-011	-0,91	-0,75
18	4	2,81	2,86	-1,75

Fonte: autores (2024)

O Grupo 4 reúne outros 18 municípios: Catuípe (22), Cruz Alta (29), Espumoso (34), Giruá (40), Ibirubá (47), Ijuí (48), Itaqui (52), Jóia (54), Lagoa Vermelha (56), Restinga Seca (82), Sananduva (91), Santa Bárbara do Sul (93), Santo Ângelo (96), São Borja (99), São Luiz Gonzaga (104), São Miguel das Missões (105), São Sepé (113) e Soledade (113).

Figura 4 – Agrupamento dos Municípios



Fonte: Elaboração própria com base nos autovetores estimados

Com relação à Figura (4), ela representa a formatação espacial da aglomeração dos municípios em seus respectivos grupos. O COREDE Jacuí Centro foi o COREDE com maior diversidade de municípios, agregando municípios que pertencem aos quatro grupos gerados. Quando comparado aos resultados de Poerschke e Moreira Junior (2020), ressaltamos que o mesmo comportamento foi verificado para os dados do Censo Agropecuário de 1995/96. Dada a heterogeneidade dos municípios parte do

COREDE Jacuí Centro, esse resultado aponta para a necessidade de uma atenção especial do gestor público, pois toda ação voltada para o COREDE deveria levar em conta a disparidade dos municípios ali agregados por contiguidade.

5 CONSIDERAÇÕES FINAIS

Além do rigor matemático para mostrar como funciona a técnica de componentes principais, esse estudo forneceu uma visão aprofundada dos Conselhos Regionais de Desenvolvimento (COREDEs) agropecuários do Rio Grande do Sul. Foi possível discutir os detalhes do método para então executar a implementação deste em um estudo de caso. Como resultado concreto, reduzimos a dimensionalidade da matriz original de dados em três componentes principais, que explicaram aproximadamente 87% da variância dos dados.

Esses resultados, quando comparados do ponto de vista da decomposição, estão em linha com Freitas et al. (2007). Os autores, que realizaram uma análise semelhante para todo o estado do Rio Grande do Sul com dados do Censo Agropecuário de 1995/96, encontraram cinco autovetores, cuja variância explicada acumulada chegou a 82% do total do conjunto. Usando o mesmo conjunto de dados, mas restringindo a análise para os COREDEs agropecuários, Poerschke & Junior (2020) estimaram cinco autovetores, que reuniram por volta de 83% da variância do conjunto original.

A identificação de quatro agrupamentos potenciais de municípios dentro dos COREDEs tem implicações práticas significativas. Essa segmentação pode servir como uma ferramenta estratégica para políticas agrícolas e de desenvolvimento regional, permitindo a adaptação de estratégias específicas às características distintas de cada grupo. Isso implica que iniciativas voltadas para, por exemplo, os municípios de relação mais forte com a produção pecuária ou intensivos em capital, podem ser mais efetivas.

No entanto, é importante reconhecer que esses são resultados preliminares, e a pesquisa deverá ser aprofundada ainda mais, conforme foi sugerido. Técnicas multivariadas acessórias devem revelar um retrato mais detalhado das variáveis e suas relações com os componentes. Esses componentes, quando rotacionados e transformados em fatores, tornarão a interpretação dos autovetores ainda mais clara.

Especificamente, para trabalhos futuros, sugerimos a utilização de uma rotação da matriz dos componentes a fim de separar um pouco mais os pesos das variáveis em cada um dos vetores, bem como a utilização dos dados do Censo Agropecuário 2007 para traçar um paralelo com os resultados aqui apresentados.

REFERÊNCIAS

Bartle, R. G. (1983). *Elementos de análise real*. Campus.

Freitas, C. A., Paz, M. V., & Nicola, D. S. (2007). Analisando a modernização da agropecuária gaúcha: uma aplicação de análise fatorial e cluster. *Análise Econômica*, *25*(47), 121–149.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417–441.

IBGE (2018). *Censo Agropecuário 2017: resultados definitivos*. IBGE.

Johnson, R. A. & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Jöreskog, K. G. (1979). Basic Ideas of Factor and Component Analysis. In Joreskog, K.G. and Sorbom, D. (Eds.), *Advances in Factor Analysis and Structural Equation Models* (pp. 5-20). University Press of America.

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

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Academic Press.

Poerschke, R. P. & Junior, F. d. J. M. (2020). Análise multivariada de dados socieconômicos: um retrato da modernização agropecuária nos coredes do rio grande do sul. *Ciência e Natura*, *42*, e13–e13.

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

Strang, G. (2019). *Linear algebra and learning from data*. SIAM.

Vidal, R., Ma, Y., & Sastry, S. (2005). Generalized principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1–15.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Contribuições dos autores

1 – Rafael Pentiado Poerschke (Corresponding Author)
Bachelor of Economics
https://orcid.org/0000-0001-9618-8535 • rafael.poerschke@gmail.com
Contribuição: Conceptualization; Methodology; Data Analysis; Writing – Original Draft Preparation

2 – Jõao Roberto Lazzarin
Mathematician
https://orcid.org/0000-0001-9527-0430 • joao.lazzarin@ufsm.br
Contribuição: Literature Review, Writing – Review & Editing

3 – Fernando Colman Tura
Mathematician
https://orcid.org/0000-0001-5423-7191 • fernando.tura@ufsm.br
Contribuição: Methodology, Writing – Editing

Como citar este artigo

Poerschke, R. P., Lazzarin, J. R., & Tura, F. C. (2025). PCA: uma ferramenta matemática para a análise dos COREDEs agropecuários do Rio Grande do Sul. *Ciência e Natura*, Santa Maria, v. 47, spe. 1, e90532. DOI: https://doi.org/10.5902/2179460X90532.

Anexos

Anexo A - Composição dos COREDEs

Alto da Serra do Botucaraí: (5) Alto Alegre, (10) Barros Cassal, (19) Campos Borges, (34) Espumoso, (36) Fontoura Xavier, (42) Gramado Xavier, (46) Ibirapuitã, (51) Itapuca, (53) Jacuizinho, (55) Lagoão, (63) Mormaço, (65) Nicolau Vergueiro, (102) São José do Herval, (113) Soledade, (117) Tio Hugo, (123) Victor Graeff.

Alto Jacuí: (11) Boa Vista do Cadeado, (12) Boa Vista do Incra, (25) Colorado, (29) Cruz Alta, (37) Fortaleza dos Valos, (47) Ibirubá, (56) Lagoa dos Três Cantos, (64) Não-Me-Toque, (81) Quinze de Novembro, (88) Saldanha Marinho, (89) Salto do Jacuí, (92) Santa Bárbara do Sul, (111) Selbach, (115) Tapera.

Fronteira Oeste: (3) Alegrete, (9) Barra do Quaraí, (50) Itacurubi, (52) Itaqui, (59) Maçambará, (60) Manoel Viana, (80) Quaraí, (87) Rosário do Sul, (94) Santa Margarida do Sul, (95) Santana do Livramento, (99) São Borja, (100) São Gabriel, (121) Uruguaiana.

Jacuí Centro: (15) Cachoeira do Sul, (23) Cerro Branco, (68) Novo Cabrais, (73) Paraíso do Sul, (82) Restinga Seca, (109) São Sepé, (125) Vila Nova do Sul.

Médio Alto Uruguai: (4) Alpestre, (6) Ametista do Sul, (18) Caiçara, (28) Cristal do Sul, (31) Dois Irmãos das Missões, (33) Erval Seco, (38) Frederico Westphalen, (41) Gramado dos Loureiros, (49) Iraí, (66) Nonoai, (69) Novo Tiradentes, (71) Palmitinho, (75) Pinhal, (76) Pinheirinho do Vale, (78) Planalto, (83) Rio dos Índios, (84) Rodeio Bonito, (110) Seberi, (116) Taquaruçu do Sul, (118) Trindade do Sul, (122) Vicente Dutra, (126) Vista Alegre.

Missões: (13) Bossoroca, (17) Caibaté, (24) Cerro Largo, (30) Dezesseis de Novembro, (32) Entre-Ijuís, (35) Eugênio de Castro, (39) Garruchos, (40) Giruá, (43) Guarani das Missões, (61) Mato Queimado, (77) Pirapó, (79) Porto Xavier, (85) Rolador, (86) Roque Gonzales, (90) Salvador das Missões, (96) Santo Ângelo, (97) Santo Antônio das Missões, (104) São Luiz Gonzaga, (105) São Miguel das Missões, (106) São Nicolau, (107) São Paulo das Missões, (108) São Pedro do Butiá, (112) Sete de Setembro, (120) Ubiretama, (127) Vitória das Missões.

Noroeste: (1) Água Santa, (8) Barracão, (16) Cacique Doble, (20) Capão Bonito do Sul, (21) Caseiros, (44) Ibiaçá, (45) Ibiraiaras, (57) Lagoa Vermelha, (58) Machadinho, (62) Maximiliano de Almeida, (70) Paim Filho, (91) Sananduva, (93) Santa Cecília do Sul, (98) Santo Expedito do Sul, (101) São João da Urtiga, (103) São José do Ouro, (114) Tapejara, (119) Tupanci do Sul, (124) Vila Lângaro.

Noroeste Colonial: (2) Ajuricaba, (7) Augusto Pestana, (14) Bozano, (22) Catuípe, (26) Condor, (27) Coronel Barros, (48) Ijuí, (54) Jóia, (67) Nova Ramada, (72) Panambi, (74) Pejuçara.

Anexo B - Resultado do agrupamento segundo os novos eixos rotacionados

Grupo 1 (69): Alto Alegre (5), Ametista do Sul (6), Barra do Quaraí (9), Boa Vista do Cadeado (11), Boa Vista do Incra (12), Bozano (14), Caibaté (17), Campos Borges (19), Capão Bonito do Sul (20), Caseiros (21), Cerro Branco (23), Colorado (25), Condor (26), Coronel Barros (27), Cristal do Sul (28), Dezesseis de Novembro (30), Dois Irmãos das Missões (31), Eugênio de Castro (35), Fortaleza dos Valos Garruchos (36), Gramado dos Loureiros (41), Gramado Xavier (42), Ibirapuitã (46), Iraí (49), Itacurubi (50), Itapuca (51), Jacuizinho (53), Lagoa dos Três Cantos (56), Maçambará (59), Manoel Viana (60), Mato Queimado (61), Maximiliano de Almeida (62), Mormaço (63), Nicolau Vergueiro (65), Nonoai (66), Nova Ramada (67), Novo Tiradentes (69), Paim Filho (70), Pejuçara (74), Pinhal (75), Pinheirinho do Vale Pirapó (76), Quaraí (80), Quinze de Novembro (81), Rio dos Índios (83), Rodeio Bonito (84), Rolador (85), Saldanha Marinho (88), Salto do Jacuí (89), Salvador das Missões (90), Santa Cecília do Sul (93), Santa Margarida do Sul (94), Santo Expedito do Sul (98), São João da Urtiga (101), São José do Herval (102), São Nicolau (106), São Pedro do Butiá (108), Selbach (111), Sete de Setembro (112), Tapera (114), Taquaruçu do Sul (116), Tio Hugo (117), Trindade do Sul (118), Tupanci do Sul (119), Ubiretama (120), Vila Lângaro (124), Vila Nova do Sul (125), Vista Alegre (126), Vitória das Missões (127).

Grupo 2 (6): Alegrete (3), Cachoeira do Sul (15), Rosário do Sul (87), Santana do Livramento (95), São Gabriel (100), Uruguaiana (121).

O Grupo 3 (34): Água Santa (1), Ajuricaba (2), Alpestre (4), Augusto Pestana (7), Barracão (8), Barros Cassal (10), Bossoroca (13), Cacique Doble (16), Caiçara (18), Cerro Largo (24), Entre-Ijuís (32), Erval Seco (33), Fontoura Xavier (36), Frederico Westphalen (38), Guarani das Missões (43), Ibiaçá (44), Ibiraiaras (45), Lagoão (55), Machadinho (58), Não-Me-Toque (64), Novo Cabrais (68), Palmitinho (71), Panambi (72), Paraíso do Sul (73), Planalto (78), Porto Xavier (79), Roque Gonzales (86), Santo Antônio das Missões (97), São José do Ouro (103), São Paulo das Missões (107), Seberi (110), Tapejara (114), Vicente Dutra (122), Victor Graeff (22).

Grupo 4 (18): Catuípe (22), Cruz Alta (29), Espumoso (34), Giruá (40), Ibirubá (47), Ijuí (48), Itaqui (52), Jóia (54), Lagoa Vermelha (56), Restinga Seca (82), Sananduva (91), Santa Bárbara do Sul (93), Santo Ângelo (96), São Borja (99), São Luiz Gonzaga (104), São Miguel das Missões (105), São Sepé (113) e Soledade (113).