Métodos Estatísticos Multivariados

Teoria, interpretação e aplicações em Python

VICTOR COSCRATO

Índice

Prefácio

Este livro é uma introdução às técnicas de análise multivariada. A ideia é construir, aos poucos, uma forma de olhar para dados com muitas variáveis ao mesmo tempo: resumir informação, enxergar estruturas, comparar grupos, classificar observações e estudar relações entre conjuntos de variáveis.

O material foi escrito para estudantes que estão tendo um primeiro contato com estatística multivariada. Sempre que possível, os conceitos são acompanhados de intuição geométrica, interpretação estatística e exemplos computacionais.

Os métodos cobertos são:

- Análise de Componentes Principais
- [Análise de Correspondência] EM BREVE!
- [Análise Fatorial] EM BREVE!
- [Análise de Agrupamento] EM BREVE!
- [Análise Discriminante] EM BREVE!
- [Análise de Correlação Canônica] EM BREVE!

O livro está organizado em duas partes principais. Primeiro, revisamos os objetos básicos que aparecem o tempo todo em análise multivariada: vetores aleatórios, matrizes de covariância, estimação, normal multivariada, decomposições matriciais e medidas de distância. Em seguida, entramos nas técnicas propriamente ditas. Cada técnica forma um capítulo completo, integrando a fundamentação teórica, exemplos práticos (manuais e computacionais) e listas de exercícios.

> **!** Importante
>
> Esse material está (e continuará por um bom tempo) sendo revisto constantemente. Certamente existem erros e partes incompletas.

Parte I

Parte I: Introdução

1 Introdução

A análise multivariada é o campo da estatística dedicado a compreender conjuntos de dados com múltiplas variáveis inter-relacionadas. Em vez de analisar cada variável isoladamente, seu foco é examinar simultaneamente as relações entre elas para extrair padrões e estruturas que de outra forma permaneceriam ocultos.

Os métodos multivariados multivariada permitem uma compreensão mais profunda e realista dos dados. Os principais objetivos são:

- **Simplificação Estrutural:** Reduzir a dimensionalidade dos dados, identificando as principais fontes de variação e eliminando redundâncias. Isso facilita a visualização e a interpretação de dados complexos, revelando a estrutura subjacente de forma mais clara.
- **Agrupamento e Classificação:** Organizar as observações em grupos homogêneos (agrupamento) ou atribuir observações a categorias predefinidas (classificação). O objetivo é identificar padrões que permitam segmentar os dados de maneira significativa.
- **Investigação de Estruturas de Dependência:** Explorar e quantificar as relações entre variáveis. Isso inclui desde a análise de correlações simples até a modelagem de interações complexas entre múltiplos conjuntos de variáveis.
- **Predição:** Construir modelos para prever o valor de uma ou mais variáveis com base em outras.
- **Inferência:** Realizar testes de hipóteses e inferências estatísticas sobre as relações em um contexto multivariado.

Nos próximos capítulos, construiremos a base teórica para atingir esses objetivos, começando pelo conceito de vetor aleatório e seus parâmetros, para depois explorarmos como as amostras de dados nos permitem estimar e analisar essas estruturas.

As técnicas de análise multivariada podem ser classificadas com base em seus objetivos e na natureza das relações entre as variáveis. Uma distinção fundamental é entre **técnicas de dependência**, que analisam a relação entre variáveis dependentes e independentes, e **técnicas de interdependência**, que exploram as relações em um único conjunto de variáveis.

- **Técnicas de Dependência:** Analisam a relação entre uma ou mais variáveis dependentes e um conjunto de variáveis independentes. O objetivo é prever ou explicar o valor das variáveis dependentes.

- **Técnicas de Interdependência:** Exploram as relações entre todas as variáveis de um conjunto, sem fazer distinção entre dependentes e independentes. O foco é entender a estrutura geral dos dados.

Além disso a escolha de uma determinada técnica depende também dos tipos de variáveis em questão.

- **Variáveis Categóricas (Qualitativas):** Representam categorias ou grupos (e.g., gênero, tipo de produto).
- **Variáveis Métricas (Quantitativas):** Representam quantidades numéricas (e.g., idade, altura, renda, temperatura).

Com o objetivo de classificar os métodos a serem apresentados nesse livro e posteriormente auxiliar na escolha da técnica mais adequada para o tratamento de um conjunto de dados, apresentamos a seguir uma tabela com algumas características de cada método e na sequência um fluxograma de decisão.

Tabela 1.1: Principais técnicas abordadas neste livro.

Técnica	Objetivo Principal	Tipo de Variável	Tipo de Análise
Componentes Principais (PCA)	Redução de dimensionalidade	Quantitativas	Interdependência
Análise Fatorial (FA)	Identificação de fatores latentes	Quantitativas	Interdependência
Análise de Agrupamento	Formação de grupos homogêneos	Quantitativas/Qualitativas	Interdependência
Análise Discriminante	Classificação de observações	Mista (Quali/Quanti)	Dependência
Correlação Canônica (CCA)	Relação entre conjuntos de variáveis	Quantitativas	Dependência
Análise de Correspondência (CA)	Relação entre variáveis categóricas	Qualitativas	Interdependência

1.1 Vetores aleatórios

Em estudos estatísticos multivariados, temos uma população de interesse (e.g., todos os estudantes de uma universidade), e um conjunto de p características que nos interessam (e.g., nota em matemática, nota em história, horas de estudo). Devido as variações, ou seja a distribuição probabilistica, que existem nessas características dentre os componentes da população, podemos tratar cada uma dessas características como uma variável aleatória. Para facilitar, podemos organizar essas variáveis em um **vetor aleatório**.

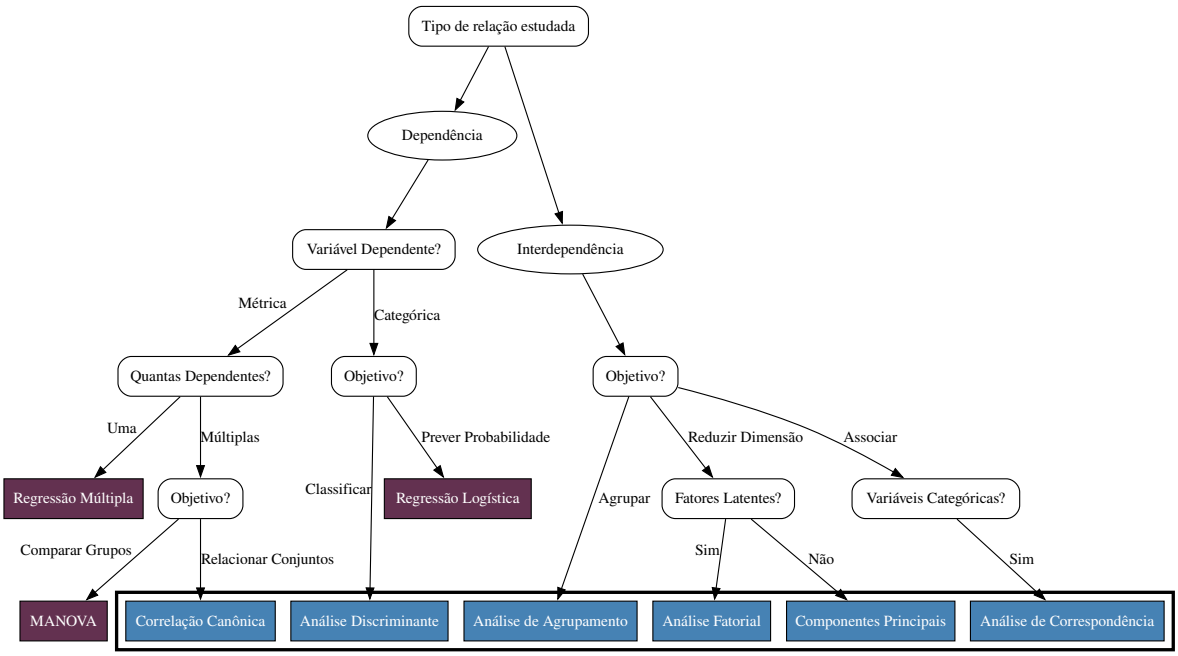

Figura 1.1: Diagrama de decisão para escolha de técnica de Análise Multivariada. Nós azuis indicam as técnicas de análise multivariada abordadas neste livro. **Importante**: Este diagrama é um guia simplificado para auxiliar na escolha da técnica mais adequada com base nas características dos dados e nos objetivos da análise. Ele não é exaustivo e serve apenas para posicionar as técnicas discutidas neste livro. A escolha final da técnica deve sempre considerar o contexto específico do problema e as características detalhadas dos dados

Definição 1.1. Um **vetor aleatório** $\boldsymbol{x}$ é um vetor-coluna cujos componentes são p variáveis aleatórias, $X_1, X_2, \ldots, X_p$.

$$\boldsymbol{x} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

A esse vetor, podemos associar uma distribuição probabilistica multivariada que governa o mecanismo gerador dos dados de uma população. Toda a teoria da análise multivariada se baseia na compreensão das propriedades e da estrutura de distribuição deste vetor.

> Cuidado com a Notação
>
> - Uma letra minúscula em negrito (e.g., $\boldsymbol{x}$) denota um **vetor aleatório**.
> - Uma letra maiúscula comum (e.g., X_j) denota uma **variável aleatória** escalar, o j-ésimo componente do vetor.
> - Mais adiante, uma letra maiúscula em negrito (e.g., $\boldsymbol{X}$) será usada para a **matriz de dados** (amostral).
> - O sobrescrito T denota **transposição**: $\boldsymbol{A}^T$ é a transposta de $\boldsymbol{A}$. Essa operação troca linhas por colunas; em particular, o vetor-coluna $\boldsymbol{x}$ da definição acima se torna o vetor-linha $\boldsymbol{x}^T = (X_1, X_2, \ldots, X_p)$.

Assim como variáveis aleatórias, os vetores aleatórios também são caracterizados por parâmetros como a média e a variância. Com o detalhe adicional que agora existem também parâmetros associados às correlações das variáveis em estudo.

Definição 1.2. O **vetor de médias populacional**, denotado por $\boldsymbol{\mu}$, é o vetor das expectativas de cada uma de suas variáveis componentes.

$$\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{x}] = \begin{pmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \vdots \\ \mathrm{E}[X_p] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

Geometricamente, $\boldsymbol{\mu}$ representa o **centróide** (centro de massa) da distribuição de probabilidade no espaço p-dimensional.

Definição 1.3. A **matriz de covariâncias populacional**, denotada por $\boldsymbol{\Sigma}$, é uma matriz simétrica $p \times p$ cujo elemento (j, k) é a covariância entre a j-ésima e a k-ésima variável aleatória, $\sigma_{jk} = \mathrm{Cov}(X_j, X_k) = \mathrm{E}[(X_j - \mu_j)(X_k - \mu_k)]$.

$$\boldsymbol{\Sigma} = \operatorname{Cov}(\boldsymbol{x}) = \operatorname{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

- **Diagonal (σ_{jj}):** As **variâncias**, $\operatorname{Var}(X_j)$, medem a dispersão de cada variável.
- **Fora da Diagonal (σ_{jk}):** As **covariâncias**, medem a tendência de associação linear entre as variáveis X_j e X_k.
- **Simetria:** A matriz é simétrica, pois $\operatorname{Cov}(X_j, X_k) = \operatorname{Cov}(X_k, X_j)$, o que implica $\sigma_{jk} = \sigma_{kj}$.

Definição 1.4. A **matriz de correlações populacional**, denotada por $\boldsymbol{P}$, é uma versão reescalada da matriz de covariâncias, com elementos $\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}}$.

$$\boldsymbol{P} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

Seus elementos ρ_{jk} variam de -1 a 1, fornecendo uma medida de associação linear livre de escala.

Os operadores de esperânça e variância no caso multivariado são generalizações diretas dos resultados univariados, e possuem propriedades semelhantes. Seja $\boldsymbol{x}$ um vetor aleatório p-dimensional com média $\boldsymbol{\mu}$ e covariância $\boldsymbol{\Sigma}$. Sejam $\boldsymbol{c}$ um vetor de constantes $p \times 1$ e $\boldsymbol{A}$ uma matriz de constantes $q \times p$.

1. **Esperança de uma Combinação Linear:**

$$\operatorname{E}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}] = \boldsymbol{A}\operatorname{E}[\boldsymbol{x}] + \boldsymbol{c} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{c}$$

2. **Covariância de uma Combinação Linear:**

$$\operatorname{Cov}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}) = \boldsymbol{A}\operatorname{Cov}(\boldsymbol{x})\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$$

Na prática, não temos acesso a esses parâmetros populacionais ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{P}$), mas podemos usar dados observados para obter estimativas confiáveis deles.

Assumimos que coletamos uma **amostra aleatória** de n observações da população. Cada observação, $\boldsymbol{x}_i$ (com $i = 1, \ldots, n$), é uma **realização** independente do vetor aleatório $\boldsymbol{x}$ que definimos no capítulo anterior. A coleção de todas essas observações forma a nossa **matriz de dados**.

Definição 1.5. A **matriz de dados**, denotada por $\boldsymbol{X}$, é uma matriz de dimensão $n \times p$, onde cada linha é uma observação multivariada e cada coluna representa uma variável.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{1.1}$$

O elemento x_{ij} representa o valor da j-ésima variável para a i-ésima observação.

Com uma matriz de dados em mãos, podemos finalmente começar a aplicar nossos métodos estatísticos multivariados. O objetivo inferencial mais comum é a **estimação** dos parâmetros populacionais $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$. Nesse livro não daremos enfoque para isso, no entanto conhecer alguns estimadores usuais é importante.

Definição 1.6. O estimador usual de $\boldsymbol{\mu}$ é o **vetor de médias amostral**, $\bar{\boldsymbol{x}}$, cujos componentes $\bar{x}_j$ são a média das observações para a j-ésima variável.

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad \text{resultando em} \quad \bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

Definição 1.7. O estimador usual de $\boldsymbol{\Sigma}$ é a **matriz de covariâncias amostral**, $\boldsymbol{S}$. Seus elementos são a variância amostral (s_{jj}) e a covariância amostral (s_{jk}).

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

A matriz resultante é:

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

> **Nota**
>
> Lembrando que a divisão por $n-1$ (ao invés de n) torna s_{jk} um estimador não-viesado σ_{jk}, ou seja, $\mathrm{E}[s_{jk}] = \sigma_{jk}$.

Definição 1.8. O estimador usual de $\boldsymbol{P}$ é a **matriz de correlações amostral**, $\boldsymbol{R}$, cujos elementos r_{jk} são obtidos padronizando a covariância amostral.

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

A matriz resultante é:

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

A matriz $\boldsymbol{R}$ é uma matriz simétrica com 1s na diagonal.

Em resumo: No **nível populacional**, temos parâmetros teóricos e não observáveis ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$). No **nível amostral**, temos dados observáveis na matriz $\boldsymbol{X}$, a partir da qual calculamos estatísticas ($\bar{\boldsymbol{x}}$, $\boldsymbol{S}$) que **estimam** esses parâmetros.

A maior parte das técnicas que veremos neste livro opera sobre as matrizes de covariâncias ou correlações amostrais ($\boldsymbol{S}$ ou $\boldsymbol{R}$) para fazer descobertas sobre as relações existentes entre as variáveis componentes de um vetor aleatório.

2 Álgebra Matricial

Com os conceitos estatísticos fundamentais estabelecidos, voltamos nossa atenção para as ferramentas matemáticas necessárias para manipular objetos multivariados. Neste capítulo, revisaremos conceitos-chave, como formas quadráticas, matrizes positiva-definidas e a decomposição espectral, que são a base para muitos métodos multivariados que veremos.

Definição 2.1. Uma forma quadrática é uma função polinomial que contém apenas termos de grau dois. Dado um vetor aleatório $\boldsymbol{x}$ de dimensão $p \times 1$ e uma matriz numérica simétrica $\boldsymbol{A}$ de dimensão $p \times p$, o produto definido por:

$$Q(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} x_i x_j$$

é uma forma quadrática.

Definição 2.2. Uma matriz simétrica $\boldsymbol{A}$ é dita **positiva-definida** se $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ para todo vetor não nulo $\boldsymbol{x}$.

Proposição 2.1 (Propriedades de Matrizes Positiva-Definidas)**.** *Se $\boldsymbol{A}$ é positiva-definida, então:*

- *todos os seus autovalores são estritamente positivos ($\lambda_i > 0$);*
- *$\boldsymbol{A}$ é invertível;*
- $\det(\boldsymbol{A}) > 0$.

Definição 2.3. Uma matriz simétrica $\boldsymbol{A}$ é dita **positiva-semidefinida** se $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$ para todos os vetores não-nulos $\boldsymbol{x}$.

Definição 2.4. Uma matriz quadrada $\boldsymbol{A}$ é dita **idempotente** se

$$\boldsymbol{A}^2 = \boldsymbol{A}.$$

Isso significa que aplicar duas vezes a transformação definida por $\boldsymbol{A}$ produz o mesmo resultado que aplicá-la uma única vez. Quando $\boldsymbol{A}$ também é simétrica, essa transformação é uma projeção ortogonal sobre o espaço gerado pelas colunas de $\boldsymbol{A}$.

Matrizes de covariância ($\boldsymbol{\Sigma}$) e correlação ($\boldsymbol{R}$) são, por natureza, positiva-semidefinidas.

Teorema 2.1 (Decomposição Espectral)**.** *Toda matriz simétrica $\boldsymbol{A}$ de dimensão $p \times p$ pode ser escrita como:*

$$\boldsymbol{A} = \boldsymbol{E}\Lambda\boldsymbol{E}^T = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$$

onde $\lambda_1, \ldots, \lambda_p$ são os autovalores de $\boldsymbol{A}$, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ são os autovetores ortonormais correspondentes, Λ é a matriz diagonal formada pelos autovalores λ_i, e $\boldsymbol{E}$ é a matriz ortogonal ($\boldsymbol{E}^T\boldsymbol{E} = \boldsymbol{E}\boldsymbol{E}^T = \boldsymbol{I}$) cujas colunas são os autovetores $\boldsymbol{e}_i$.

Essa decomposição tem uma interpretação geométrica importante. Cada autovetor $\boldsymbol{e}_i$ define uma **direção invariante** da transformação linear associada a $\boldsymbol{A}$, pois $\boldsymbol{A}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$. O autovalor correspondente λ_i indica o fator pelo qual essa direção é escalada. Como os autovetores são ortonormais, eles formam um novo sistema de eixos no qual a matriz $\boldsymbol{A}$ é representada pela matriz diagonal Λ.

Para compreender essa mudança de coordenadas, considere um vetor $\boldsymbol{x}$ escrito na base formada pelos autovetores:

$$\boldsymbol{x} = y_1\boldsymbol{e}_1 + y_2\boldsymbol{e}_2 + \cdots + y_p\boldsymbol{e}_p.$$

Os valores $y_1, \ldots, y_p$ são simplesmente as coordenadas de $\boldsymbol{x}$ nesse novo sistema de eixos. Reunindo essas coordenadas no vetor $\boldsymbol{y} = (y_1, \ldots, y_p)^T$, podemos escrever a mesma relação de forma compacta como $\boldsymbol{x} = \boldsymbol{E}\boldsymbol{y}$. Como $\boldsymbol{E}$ é ortogonal, também temos $\boldsymbol{y} = \boldsymbol{E}^T\boldsymbol{x}$.

Substituindo $\boldsymbol{x} = \boldsymbol{E}\boldsymbol{y}$ na forma quadrática e usando $\boldsymbol{A} = \boldsymbol{E}\Lambda\boldsymbol{E}^T$, obtemos:

$$\begin{aligned}
\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= (\boldsymbol{E}\boldsymbol{y})^T(\boldsymbol{E}\Lambda\boldsymbol{E}^T)(\boldsymbol{E}\boldsymbol{y}) \\
&= \boldsymbol{y}^T\Lambda\boldsymbol{y} \\
&= \sum_{i=1}^{p} \lambda_i y_i^2.
\end{aligned}$$

Assim, a mudança de $\boldsymbol{x}$ para $\boldsymbol{y}$ não altera o vetor representado: ela apenas troca o sistema de coordenadas. Nessa nova base, cada autovalor λ_i descreve a contribuição da direção $\boldsymbol{e}_i$ para a forma quadrática. Quando $\boldsymbol{A}$ é positiva-semidefinida, todos os autovalores são não negativos e podem ser ordenados como $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Veremos durante o livro o caso especial em que $\boldsymbol{A}$ é uma matriz de covariância. Nesse contexto, os autovetores indicam direções de projeção não correlacionadas, e os autovalores representam as variâncias nessas direções.

Exemplo 2.1. Vamos decompor a seguinte matriz simétrica $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

1. **Autovalores**: Resolvendo a equação característica $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, encontramos $\lambda_1 = 3$ e $\lambda_2 = 1$.
2. **Autovetores**:
   - Para $\lambda_1 = 3$: O autovetor correspondente é $\boldsymbol{e}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$.
   - Para $\lambda_2 = 1$: O autovetor correspondente é $\boldsymbol{e}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$.

A decomposição é $\boldsymbol{A} = \boldsymbol{E}\Lambda\boldsymbol{E}^T$, com:

$$\boldsymbol{E} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

Assim, $\boldsymbol{A}$ multiplica por 3 a direção de $\boldsymbol{e}_1$ e por 1 a direção de $\boldsymbol{e}_2$.

Definição 2.5 (Decomposição em Valores Singulares)**.** Toda matriz $\boldsymbol{A}$ de dimensão $I \times J$ pode ser escrita como:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

onde $\boldsymbol{U}$ é uma matriz ortogonal $I \times I$ cujas colunas são os **vetores singulares à esquerda**, $\boldsymbol{V}$ é uma matriz ortogonal $J \times J$ cujas colunas são os **vetores singulares à direita**, e $\boldsymbol{\Lambda}$ é uma matriz retangular $I \times J$ que contém os **valores singulares** σ_k em sua diagonal principal, ordenados de modo que $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, e zeros nas demais posições. Os valores singulares são as raízes quadradas dos autovalores não-nulos de $\boldsymbol{A}^T\boldsymbol{A}$ e $\boldsymbol{A}\boldsymbol{A}^T$.

Enquanto a decomposição espectral se aplica a matrizes simétricas, a SVD é um resultado mais geral, possibilitando a decomposição de qualquer matriz. Por isso, ela é uma das fatorações mais importantes da álgebra linear, com aplicações em estatística e aprendizado de máquina, incluindo a Análise de Componentes Principais e a Análise de Correspondência.

Para uma matriz simétrica e positiva-semidefinida $\boldsymbol{A}$, como uma matriz de covariância, a SVD e a decomposição espectral são essencialmente a mesma coisa. Seus valores singulares são seus autovalores, e seus vetores singulares à esquerda e à direita são seus autovetores ($\boldsymbol{U} = \boldsymbol{V} = \boldsymbol{E}$).

A grande utilidade da SVD vem do fato de que ela fornece a **melhor aproximação de baixo posto** de uma matriz. O Teorema de Eckart-Young mostra que, se truncarmos a decomposição para usar

apenas os M maiores valores singulares, a matriz resultante $\boldsymbol{A}_M$ é a melhor aproximação de posto M da matriz original $\boldsymbol{A}$.

$$\boldsymbol{A} \approx \boldsymbol{A}_M = \boldsymbol{U}_M \boldsymbol{\Lambda}_M \boldsymbol{V}_M^T = \sum_{k=1}^{M} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^T$$

Isso fornece uma maneira ótima de capturar parte da estrutura de uma matriz usando um número menor de dimensões.

3 A Distribuição Normal Multivariada

O vetor de médias e a matriz de covariâncias descrevem o centro e a variabilidade de um vetor aleatório, mas não determinam sozinhos toda a sua distribuição. A distribuição Normal Multivariada (NMV) é um modelo de referência particularmente útil quando a nuvem de dados tem forma aproximadamente elíptica e quando queremos usar resultados inferenciais clássicos. Ela não deve ser tomada como descrição automática de qualquer conjunto de dados.

Um vetor aleatório $\boldsymbol{x}$ de dimensão p segue uma distribuição Normal Multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias positiva-definida $\boldsymbol{\Sigma}$, denotada por $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, se tem função de densidade

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

O termo no expoente é a distância de Mahalanobis ao quadrado entre $\boldsymbol{x}$ e o centro da distribuição:

$$d_M^2(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}).$$

Ele considera simultaneamente a escala de cada variável e sua associação com as demais. Pontos com maior distância de Mahalanobis ao centro recebem menor densidade. A inversa e o determinante presentes na densidade existem porque, neste capítulo, $\boldsymbol{\Sigma}$ é positiva-definida.

3.1 Propriedades da Distribuição Normal Multivariada

As propriedades a seguir explicam por que esse modelo aparece com frequência na teoria multivariada.

1. **Transformações lineares.** Se $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então, para uma matriz constante $\boldsymbol{A}$ e um vetor constante $\boldsymbol{b}$,
$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \sim N_q(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T).$$
Em particular, toda combinação linear $\boldsymbol{a}^T\boldsymbol{x}$ tem distribuição normal univariada.

2. **Distribuições marginais.** Qualquer subconjunto das componentes de um vetor NMV também tem distribuição normal multivariada. Por exemplo, se $\boldsymbol{x} = (X_1, X_2, X_3)^T$ é NMV, então $(X_1, X_3)^T$ também é NMV.

3. **Covariância zero e independência.** Em geral, não correlação não implica independência. Para vetores conjuntamente normais, porém, dois subconjuntos são independentes quando toda a matriz de covariâncias cruzadas entre eles é nula.

3.2 Visualizando a Normal Bivariada

No caso bivariado, os conjuntos de pontos com a mesma densidade satisfazem

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2.$$

São elipses centradas em $\boldsymbol{\mu}$. Os autovetores de $\boldsymbol{\Sigma}$ apontam para seus eixos; os comprimentos dos semi-eixos são proporcionais às raízes quadradas dos autovalores. A Figura 3.1 mostra três possibilidades.

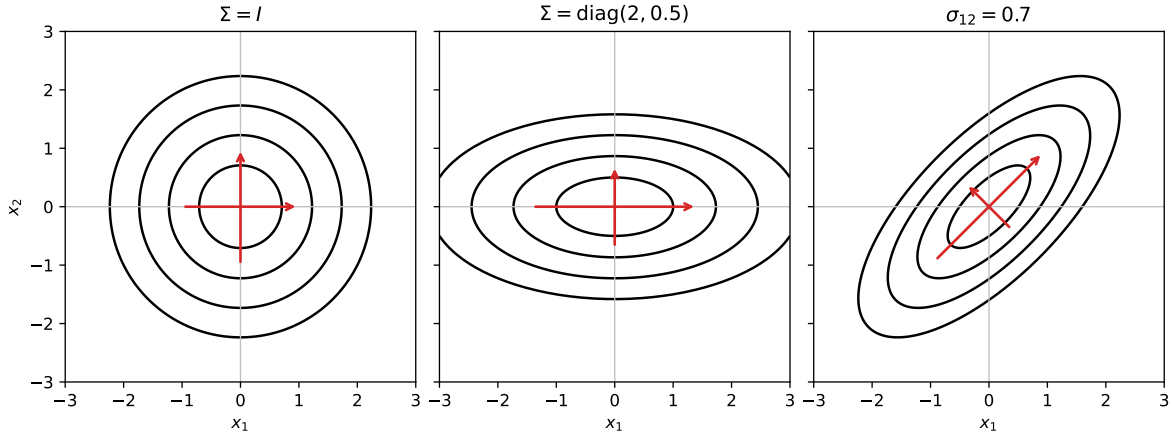

Figura 3.1: Contornos de densidade de normais bivariadas. Os vetores vermelhos indicam os autovetores da matriz de covariâncias.

Quando a covariância é zero e as variâncias são iguais, os contornos são círculos. Com variâncias diferentes e covariância zero, eles permanecem alinhados aos eixos originais. Quando há covariância, os eixos principais se rotacionam.

A normalidade multivariada será uma hipótese importante na análise discriminante e nos modelos de mistura. Já a ACP e o agrupamento hierárquico podem ser calculados sem assumir esse modelo; nele, a normalidade muda principalmente a interpretação inferencial, não a definição do método.

Parte II

Parte II: Métodos multivariados

4 Análise de Componentes Principais

A Análise de Componentes Principais (ACP) é uma técnica para estudar e reorganizar a variabilidade de um conjunto de variáveis. A ACP constrói um novo sistema de coordenadas para os dados, possivelmente simplificando as análises. Os novos eixos são combinações lineares das variáveis originais, são mutuamente ortogonais e aparecem em ordem decrescente de variância.

Definição 4.1. Dado um vetor aleatório p-dimensional $\boldsymbol{x} = (X_1, \ldots, X_p)^T$, com matriz de covariância $\boldsymbol{\Sigma}$, os **componentes principais (CPs)** são um novo sistema de coordenadas $Y_1, \ldots, Y_p$, tal que

$$\begin{aligned} Y_1 &= \boldsymbol{u}_1^T \boldsymbol{x} = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Y_2 &= \boldsymbol{u}_2^T \boldsymbol{x} = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ &\vdots \\ Y_p &= \boldsymbol{u}_p^T \boldsymbol{x} = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{aligned}$$

Onde $\boldsymbol{u}_i^T$ são vetores que definem as direções das novas coordenadas. Os CPs tem duas propriedades importantes:

- São ortogonais entre sí, ou seja, para $i \neq j$, $\boldsymbol{u}_i^T \boldsymbol{u}_j = 0$ e consequentemente $\mathrm{Var}\,(Y_i, Y_j) = 0$
- Suas variâncias são decrescentes, isto é, $\mathrm{Var}\,(Y_1) \geq \mathrm{Var}\,(Y_2) \geq \cdots \geq \mathrm{Var}\,(Y_p)$.

Essa mudança de coordenadas pode ter dois propósitos relacionados: **descrever a estrutura de dependência entre as variáveis** e/ou **reduzir a dimensionalidade**. Frequentemente, a ACP é uma ferramenta intermediária em uma análise. Isso porque muitas vezes a análise dos dados observados no novo sistema de coordenadas é mais conveniente. Uma característica importantíssima para tal conveniência é a independência.

> **!** Importante
>
> A variabilidade que a ACP reorganiza é medida em torno da média. Por isso, antes de construir os componentes, deslocamos a nuvem de dados para que seu centro coincida com a origem. No contexto populacional, o vetor centrado é $\boldsymbol{x}^c = \boldsymbol{x} - \boldsymbol{\mu}$; em uma amostra, cada observação é centralizada por $\boldsymbol{x}_i^c = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$. Esse deslocamento não altera a forma da nuvem nem as distâncias entre as observações, apenas seu ponto de referência. Assim, as projeções nos novos eixos descrevem desvios em relação ao centro dos dados.

Para simplificar a notação, daqui em diante $\boldsymbol{x}$ representa o vetor de variáveis já centralizado.

Exemplo 4.1. Considere dados fictícios de 160 estudantes matriculados em uma mesma disciplina. Para cada estudante, foram registradas a nota na prova teórica (X_1) e a nota em um trabalho prático (X_2), ambas na escala de 0 a 10. Aplicando a centralização descrita acima, subtraímos de cada nota a média da respectiva avaliação. Seguindo a convenção adotada no capítulo, continuaremos chamando as variáveis centralizadas de X_1 e X_2. Assim, valores positivos indicam desempenho acima da média da turma e valores negativos, desempenho abaixo da média. A Figura 4.1 (esquerda) mostra essas duas variáveis, que apresentam forte associação positiva; cada ponto representa um estudante.

Os eixos X_1 e X_2 são perfeitamente válidos, mas não acompanham a orientação principal da nuvem. A dispersão é muito maior ao longo de uma direção diagonal: estudantes com nota acima da média na prova tendem também a ficar acima da média no trabalho, e o mesmo ocorre com desempenhos abaixo da média. Isso sugere que podemos definir um novo eixo, isto é, uma nova dimensão, para descrever cada estudante combinando as notas na prova e no trabalho. A Figura 4.1 (direita) ilustra essa nova dimensão, que vamos chamar de Y_1.

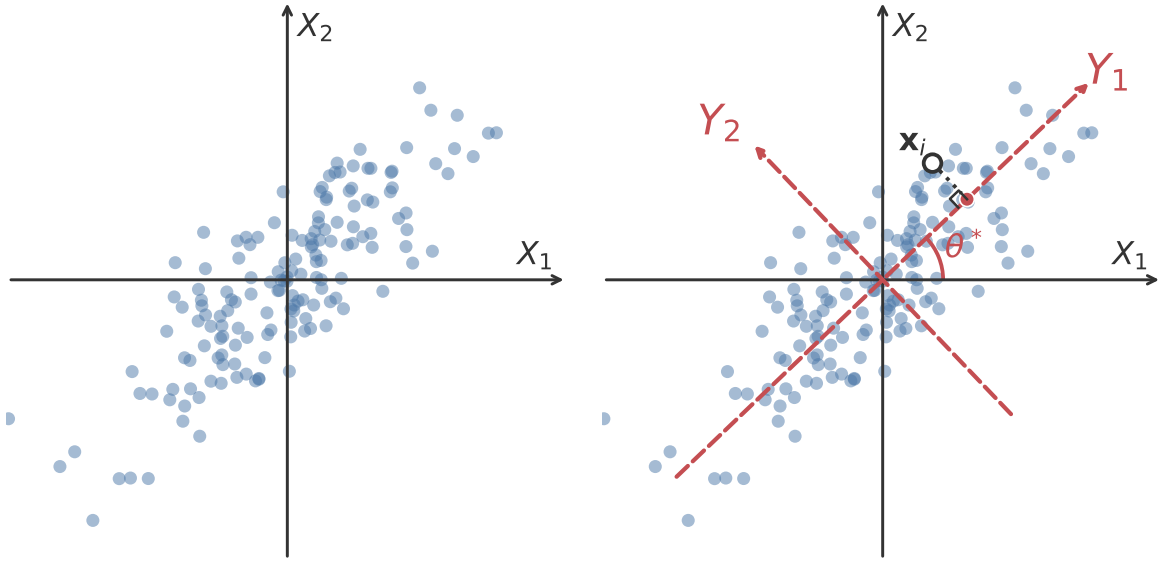

Figura 4.1: Representação geométrica dos dados nos eixos originais e nos novos eixos Y_1 e Y_2, com a projeção ortogonal de uma observação.

4.1 Variância como Medida de Informação

No Exemplo 4.1 Note que o eixo Y_1 acompanha a direção de maior alongamento da nuvem de pontos. Em estatística, a **variância** é frequentemente usada como uma medida de **informação**. Uma variável

com alta variância indica que seus valores são bem espalhados, o que nos ajuda a diferenciar as observações. Se a variância fosse zero, todos os pontos seriam idênticos, não nos fornecendo nenhuma informação sobre suas diferenças.

Definição 4.2. A **variância total** de um conjunto de dados com p variáveis é a soma das variâncias de cada variável individual. Matematicamente, se $\boldsymbol{x} = (X_1, \ldots, X_p)^T$ é o vetor de variáveis aleatórias com matriz de covariâncias $\boldsymbol{\Sigma}$, a variância total é definida como:

$$\text{Variância Total} = \sum_{j=1}^{p} \text{Var}\,(X_j) = \sum_{j=1}^{p} \sigma_{jj} = \text{tr}\,(\boldsymbol{\Sigma})$$

onde σ_{jj} é a variância da j-ésima variável e $\text{tr}\,(\boldsymbol{\Sigma})$ é o traço da matriz de covariâncias (a soma dos elementos da diagonal principal). Essa medida representa a dispersão total na nuvem de pontos, somando a variabilidade em cada uma das direções dos eixos originais.

4.2 Interpretação e solução geométrica

Seguindo o Exemplo 4.1. Propositalmente, a Figura 4.1 mostra que a busca pelo eixo mais informativo Y_1 pode ser traduzido como a busca um um ângulo ideal θ^* que define uma rotação do eixo original X_1. Para um ângulo θ qualquer, a rotação pode ser expressa pela matriz de rotação

$$\begin{bmatrix} \boldsymbol{u}_1(\theta) & \boldsymbol{u}_2(\theta) \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

cujas colunas $\boldsymbol{u}_1(\theta),\ \boldsymbol{u}_2(\theta)$ formam uma base ortonormal, isso garante que as direções Y_1 e Y_2 sejam perpendiculares entre si. As projeções dos dados originais nesse sistema são

$$\begin{aligned} Y_1 &= \boldsymbol{u}_1(\theta)^T \boldsymbol{x} = \cos(\theta) X_1 + \sin(\theta) X_2 \\ Y_2 &= \boldsymbol{u}_2(\theta)^T \boldsymbol{x} = -\sin(\theta) X_1 + \cos(\theta) X_2. \end{aligned}$$

Ainda não sabemos como obter o ângulo ideal θ^*. Podemos, entretanto, formular precisamente o que significa alinhar o primeiro eixo com a maior dispersão. A variância da coordenada Y_1 é

$$\begin{aligned} \text{Var}\,(Y_1) &= \text{Var}\left(\boldsymbol{u}_1(\theta)^T \boldsymbol{x}\right) \\ &= \boldsymbol{u}_1(\theta)^T \boldsymbol{\Sigma} \boldsymbol{u}_1(\theta) \\ &= \begin{bmatrix} \cos\theta \sin\theta \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \\ &= \sigma_{11} \cos^2\theta + 2\sigma_{12} \sin\theta \cos\theta + \sigma_{22} \sin^2\theta. \end{aligned}$$

em que $\mathbf{\Sigma}$ é a matriz de covariâncias de $(X_1, X_2)^T$. Portanto, cada rotação produz uma variância diferente para o primeiro eixo. A ACP escolhe a rotação que maximiza essa função. Nesse exemplo simples, podemos observar isso gráficamente. A Figura 4.2 ilustra a solução.

> **Nota**
>
> Em duas dimensões, existe também uma solução trigonométrica para esse problema de maximização. Essa solução contudo, não se estende de maneira conveniente a dezenas ou centenas de variáveis, por isso não prosseguiremos com ela.

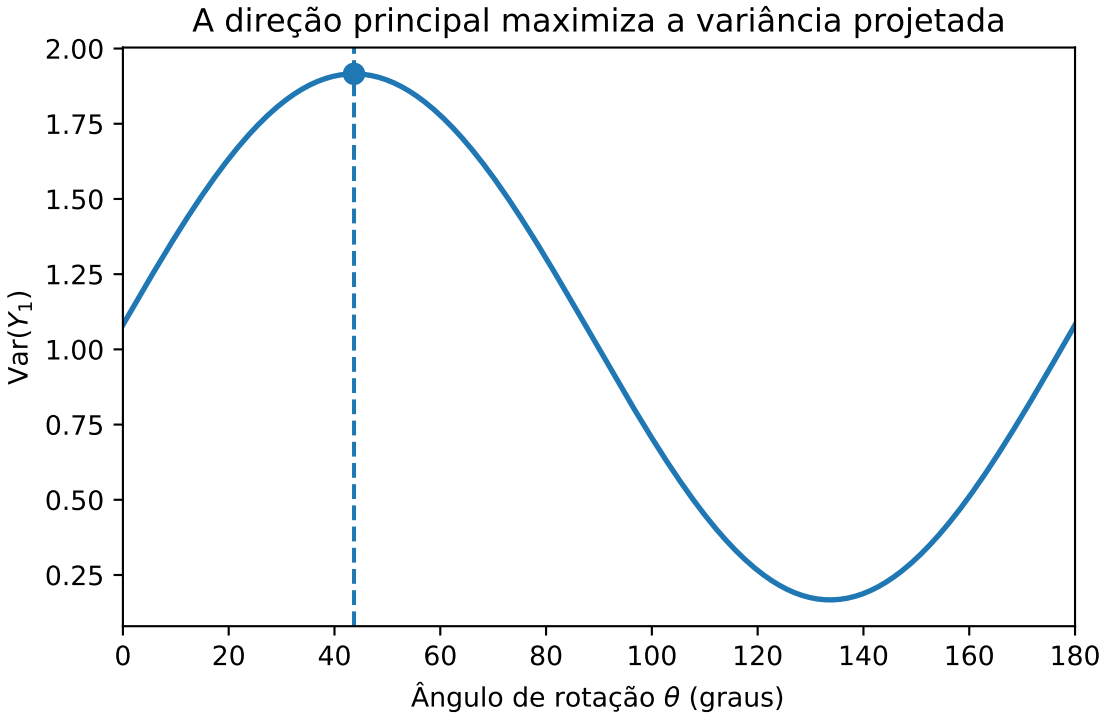

Figura 4.2: Variância da primeira coordenada rotacionada em função do ângulo. O máximo identifica a direção do primeiro componente principal.

As variâncias das notas nos eixos originais são $\text{Var}(X_1) = 1.08$ e $\text{Var}(X_2) = 1.0$. Consequentemente, a variância total é 2.08.

Uma vez fixado o ângulo ideal θ^*, O eixo Y_1 é exatamente a direção de maior variabilidade, e consequentemente, o mais informativo. Para calcular a variância dos dados nesse novo eixo consideramos as projeções ortogonais das observações sobre ele. Essas projeções são definidas exatamente pela matriz de rotação. Para um ponto $\boldsymbol{x}_i = (x_1, x_2)$ arbitrário, a projeção sobre Y_1 é dada por $\boldsymbol{y}_i = \cos(\theta)x_1 + \sin(\theta)x_2$. A Figura 4.1 (direita) ilustra a projeção. A variância dos dados projetados sobre Y_1 é

$$\begin{aligned}\text{Var}\,(Y_1) &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{Y}_1)^2 \\ &= 1.92\end{aligned}$$

Note que Y_1 possui uma variância superior a X_1 e X_2. Com relação a variância total, temos o resíduo Variância Total $-\ \text{Var}\,(Y_1) = 0.16$. É aqui que introduzimos o segundo novo eixo Y_2. Ele é perpendicular a Y_1 e captura exatamente essa variância residual $\text{Var}\,(Y_2) = 0.16$.

Fica claro como os novos eixos Y_1 e Y_2 definem um par de coordenadas que redistribui a informação dos dados. O primeiro eixo Y_1 captura a maior parte da variabilidade, enquanto o segundo Y_2 captura a variância residual.

> **Nota**
>
> Não é possível, e nem necessário, que o leitor verifique o calculo dessas variâncias, pois no exemplo apresentamos apenas a nuvem de pontos, mas não os dados originais.

4.3 A Solução dos componentes principais

Para encontrar o primeiro componente principal, partimos da combinação linear $Y_1 = \boldsymbol{u}_1^T\boldsymbol{x}$, em que $\boldsymbol{u}_1$ é, por enquanto, apenas uma direção candidata. Sua variância é $\text{Var}\,(Y_1) = \text{Var}\left(\boldsymbol{u}_1^T\boldsymbol{x}\right) = \boldsymbol{u}_1^T\boldsymbol{\Sigma}\boldsymbol{u}_1$, onde $\boldsymbol{\Sigma}$ é a matriz de covariâncias de $\boldsymbol{x}$.

Para evitar que a variância seja aumentada simplesmente inflando os coeficientes em $\boldsymbol{u}_1$, impomos a restrição de que seu comprimento seja unitário, $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$. Formalmente, o problema de maximização para o primeiro componente principal se torna:

$$\begin{aligned}\max_{\boldsymbol{u}_1} \quad & \boldsymbol{u}_1^T\boldsymbol{\Sigma}\boldsymbol{u}_1 \\ \text{sujeito a} \quad & \boldsymbol{u}_1^T\boldsymbol{u}_1 = 1\end{aligned}$$

Utilizando o método dos multiplicadores de Lagrange, a função a ser maximizada é:

$$L(\boldsymbol{u}_1, \lambda_1) = \boldsymbol{u}_1^T\boldsymbol{\Sigma}\boldsymbol{u}_1 - \lambda_1(\boldsymbol{u}_1^T\boldsymbol{u}_1 - 1)$$

Derivando em relação a $\boldsymbol{u}_1$ e igualando a zero, obtemos:

$$\frac{\partial L}{\partial \boldsymbol{u}_1} = 2\boldsymbol{\Sigma}\boldsymbol{u}_1 - 2\lambda_1\boldsymbol{u}_1 = 0 \implies \boldsymbol{\Sigma}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1$$

Note que chegamos à equação fundamental de autovalores e autovetores. Esta equação mostra que a direção candidata $\boldsymbol{u}_1$ que resolve o problema deve ser um autovetor da matriz de covariâncias $\boldsymbol{\Sigma}$. Para encontrar a variância, pré-multiplicamos a equação por $\boldsymbol{u}_1^T$:

$$\boldsymbol{u}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^T \boldsymbol{u}_1$$

Como $\text{Var}\,(Y_1) = \boldsymbol{u}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_1$ e a restrição é $\boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$, temos:

$$\text{Var}\,(Y_1) = \lambda_1$$

Para maximizar a variância de Y_1, devemos escolher o maior autovalor possível. Portanto, λ_1 é o **maior autovalor** de $\boldsymbol{\Sigma}$. Denotamos por $\boldsymbol{e}_1$ um autovetor unitário associado a λ_1; na solução do problema, a direção candidata se torna $\boldsymbol{u}_1 = \boldsymbol{e}_1$.

> **Nota**
>
> A matriz de covariâncias $\boldsymbol{\Sigma}$ é, por construção, simétrica e positiva semidefinida. Conforme discutido em Teorema 2.1, o **Teorema Espectral** garante que seus autovalores são reais e não negativos e que existe uma base ortonormal de autovetores.

Para obter o segundo componente, consideramos uma nova direção candidata $\boldsymbol{u}_2$ e escrevemos $Y_2 = \boldsymbol{u}_2^T \boldsymbol{x}$. Como $\boldsymbol{e}_1$ já determina a direção de maior variância, exigimos que $\boldsymbol{u}_2$ tenha comprimento unitário e seja ortogonal a $\boldsymbol{e}_1$. Essa ortogonalidade também garante que os componentes sejam não correlacionados: $\boldsymbol{\Sigma}$, $\text{Cov}\,(Y_1, Y_2) = \boldsymbol{e}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_2 = \lambda_1 \boldsymbol{e}_1^T \boldsymbol{u}_2 = 0$. Assim, o segundo problema de otimização é

$$\begin{aligned} \max_{\boldsymbol{u}_2} \quad & \boldsymbol{u}_2^T \boldsymbol{\Sigma} \boldsymbol{u}_2 \\ \text{sujeito a} \quad & \begin{cases} \boldsymbol{u}_2^T \boldsymbol{u}_2 = 1 \\ \boldsymbol{e}_1^T \boldsymbol{u}_2 = 0 \end{cases} \end{aligned}$$

A restrição de ortogonalidade introduz um segundo multiplicador de Lagrange, ϕ. A função Lagrangiana e sua derivada são

$$\begin{aligned} L(\boldsymbol{u}_2, \lambda_2, \phi) &= \boldsymbol{u}_2^T \boldsymbol{\Sigma} \boldsymbol{u}_2 - \lambda_2 (\boldsymbol{u}_2^T \boldsymbol{u}_2 - 1) - \phi \boldsymbol{e}_1^T \boldsymbol{u}_2, \\ \frac{\partial L}{\partial \boldsymbol{u}_2} &= 2\boldsymbol{\Sigma} \boldsymbol{u}_2 - 2\lambda_2 \boldsymbol{u}_2 - \phi \boldsymbol{e}_1 = \boldsymbol{0}. \end{aligned}$$

Pré-multiplicamos a derivada por $\boldsymbol{e}_1^T$. Como $\boldsymbol{e}_1^T \boldsymbol{\Sigma} = \lambda_1 \boldsymbol{e}_1^T$, $\boldsymbol{e}_1^T \boldsymbol{u}_2 = 0$ e $\boldsymbol{e}_1^T \boldsymbol{e}_1 = 1$, obtemos

$$
\begin{aligned}
0 &= 2\boldsymbol{e}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_2 - 2\lambda_2 \boldsymbol{e}_1^T \boldsymbol{u}_2 - \phi \boldsymbol{e}_1^T \boldsymbol{e}_1 \\
&= 2\lambda_1 \boldsymbol{e}_1^T \boldsymbol{u}_2 - 2\lambda_2 \boldsymbol{e}_1^T \boldsymbol{u}_2 - \phi \\
&= -\phi.
\end{aligned}
$$

Logo, $\phi = 0$, e a derivada se reduz a $\boldsymbol{\Sigma}\boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2$. A segunda direção também deve, portanto, ser um autovetor de $\boldsymbol{\Sigma}$. Entre os autovetores ortogonais a $\boldsymbol{e}_1$, a maior variância disponível é o segundo maior autovalor, λ_2. Denotando o autovetor unitário correspondente por $\boldsymbol{e}_2$, a solução é $\boldsymbol{u}_2 = \boldsymbol{e}_2$.

Este processo continua: o k-ésimo componente principal (Y_k) é definido pelo autovetor $\boldsymbol{e}_k$ associado ao k-ésimo maior autovalor λ_k, garantindo que $\operatorname{Var}(Y_k) = \lambda_k$ e que todos os componentes sejam mutuamente não correlacionados.

A decomposição espectral reúne todos esses resultados de uma só vez. Na expressão $\boldsymbol{\Sigma} = \boldsymbol{E}\Lambda\boldsymbol{E}^T$, as colunas de $\boldsymbol{E}$ são as direções dos componentes principais e os elementos da diagonal de Λ são suas variâncias. Ordenando os autovalores do maior para o menor, obtemos simultaneamente todos os componentes principais na ordem desejada.

Proposição 4.1 (Conservação da variância total)**.** *Se $Y_1, \ldots, Y_p$ são todos os componentes principais de $\boldsymbol{x}$, então a soma de suas variâncias é igual à variância total das variáveis originais:*

$$
\sum_{k=1}^{p} \operatorname{Var}(Y_k) = \sum_{j=1}^{p} \operatorname{Var}(X_j).
$$

Comprovação. Pela Definição 4.2, a variância total das variáveis originais é $\operatorname{tr}(\boldsymbol{\Sigma})$. Usando a decomposição espectral $\boldsymbol{\Sigma} = \boldsymbol{E}\Lambda\boldsymbol{E}^T$, as propriedades do traço, a ortogonalidade de $\boldsymbol{E}$ e o fato de que $\operatorname{Var}(Y_k) = \lambda_k$, obtemos

$$
\begin{aligned}
\sum_{j=1}^{p} \operatorname{Var}(X_j) &= \operatorname{tr}(\boldsymbol{\Sigma}) \\
&= \operatorname{tr}\left(\boldsymbol{E}\Lambda\boldsymbol{E}^T\right) \\
&= \operatorname{tr}\left(\Lambda\boldsymbol{E}^T\boldsymbol{E}\right) \\
&= \operatorname{tr}(\Lambda) \\
&= \sum_{k=1}^{p} \lambda_k \\
&= \sum_{k=1}^{p} \operatorname{Var}(Y_k).
\end{aligned}
$$

Portanto, a ACP não cria nem elimina variabilidade; ela apenas a redistribui entre os componentes principais. □

> **Nota**
>
> A demonstração acima, utilizando multiplicadores de Lagrange, é uma maneira moderna e elegante de conduzir a derivação do problema de maximização. Uma abordagem clássica restringe a norma de $\boldsymbol{u}_1$ através do quociente,
>
> $$\text{Var}\,(Y_1) = \max_{\boldsymbol{u}_1} \frac{\boldsymbol{u}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_1}{\boldsymbol{u}_1^T \boldsymbol{u}_1}$$
>
> Este é um problema clássico na álgebra linear. Um teorema fundamental afirma que para qualquer matriz simétrica A, o máximo da forma quadrática $\boldsymbol{x}^T A \boldsymbol{x}$, sujeito à restrição $\boldsymbol{x}^T \boldsymbol{x} = 1$, é o **maior autovalor** de A. O vetor $\boldsymbol{x}$ que atinge esse máximo é o autovetor correspondente. Como a matriz de covariâncias $\boldsymbol{\Sigma}$ é simétrica, este teorema se aplica diretamente ao nosso problema.

4.4 Componentes Principais Amostrais

A derivação anterior foi formulada em termos populacionais, supondo conhecidos a matriz de covariâncias $\boldsymbol{\Sigma}$ e seus autovalores e autovetores. O Exemplo 4.1, por outro lado, já utilizou uma amostra para construir a interpretação geométrica. Resta explicitar como a solução teórica é aplicada nesse contexto.

Em uma amostra, centralizamos as observações e substituímos $\boldsymbol{\Sigma}$ pela matriz de covariâncias amostral $\boldsymbol{S}$. Os componentes principais são obtidos da mesma maneira, e as quantidades resultantes são estimativas de seus análogos populacionais:

- O k-ésimo autovalor amostral, $\hat{\lambda}_k$, é uma estimativa de λ_k.
- O k-ésimo autovetor amostral, $\hat{\boldsymbol{e}}_k$, é uma estimativa de $\boldsymbol{e}_k$.
- O k-ésimo componente principal amostral, $\hat{Y}_k = \hat{\boldsymbol{e}}_k^T \boldsymbol{x}$, é uma estimativa de Y_k.

A teoria e a interpretação permanecem as mesmas. Para simplificar a notação, ao longo deste capítulo, omitimos o acento circunflexo ($\hat{}$), mas é importante lembrar que, na aplicação prática, estamos sempre lidando com estimativas amostrais.

Outra questão importante é a escala das variáveis. Retome os dados do Exemplo 4.1 e imagine que, além das duas notas, registramos o **tempo de estudo no semestre**, medido em horas. As notas são expressas em pontos e o tempo de estudo em horas, de modo que suas variâncias têm unidades diferentes e não são diretamente comparáveis. Se por acaso o tempo fosse registrado em minutos ou até segundos, sua variância seria muito maior.

> **Importante**
>
> Para evitar que as escalas tenham esse efeito, padronizamos as variáveis dividindo cada uma por seu desvio padrão. Todas passam a ter variância igual a 1. Devemos notar que essa padronização é exatamente aquela que transforma a matriz de covariâncias $\boldsymbol{S}$ em uma matriz de correlações $\boldsymbol{R}$. Ou seja, para evitar efeitos de escala, aplicamos a decomposição sobre a matriz de covariâncias $\boldsymbol{S}$.

Exemplo 4.2 (Um Cálculo Manual Completo)**.** Até aqui, a decomposição espectral resolveu o problema de ACP em termos abstratos, com $\boldsymbol{\Sigma}$ e seus autovalores e autovetores. Mas como isso se traduz quando temos dados reais em mãos? Vamos resolver um problema pequeno inteiramente à mão, com uma amostra cujos números foram escolhidos para deixar as contas simples.

Suponha que observamos $n = 5$ pares de valores já centralizados de duas variáveis, X_1 e X_2, organizados na matriz de dados $\boldsymbol{X}$ (uma observação por linha, como em Equação 1.1):

$$\boldsymbol{X} = \begin{bmatrix} -2 & -2 \\ -1 & 1 \\ 0 & 0 \\ 1 & -1 \\ 2 & 2 \end{bmatrix}$$

Matriz de covariâncias amostral. Como os dados já estão centralizados, a soma de cada coluna de $\boldsymbol{X}$ é zero. As variâncias e a covariância amostrais são:

$$s_{11} = \frac{\sum x_{i1}^2}{n-1} = \frac{10}{4} = 2.5, \qquad s_{22} = \frac{\sum x_{i2}^2}{n-1} = \frac{10}{4} = 2.5, \qquad s_{12} = \frac{\sum x_{i1}x_{i2}}{n-1} = \frac{6}{4} = 1.5$$

$$\boldsymbol{S} = \begin{bmatrix} 2.5 & 1.5 \\ & 2.5 \end{bmatrix}$$

> **Nota**
>
> Neste exemplo, vamos assumir que X_1 e X_2 tem escalas comparáveis, e por isso, nao padronizaremos os dados, ou seja, vamos aplicar a decomposição diretamente em $\boldsymbol{\Sigma}$.

Autovalores e Autovetores. Resolvemos $\det(\boldsymbol{S} - \lambda \boldsymbol{I}) = 0$:

$$(2.5 - \lambda)^2 - 1.5^2 = 0 \quad \Longrightarrow \quad 2.5 - \lambda = \pm 1.5 \quad \Longrightarrow \quad \lambda_1 = 4, \quad \lambda_2 = 1$$

Note que $\lambda_1 + \lambda_2 = 5 = \operatorname{tr}(\boldsymbol{S})$, como esperado.

Para cada autovalor, substituímos λ_k no sistema $(\boldsymbol{S} - \lambda_k \boldsymbol{I})\boldsymbol{e}_k = \boldsymbol{0}$. Os sistemas resultantes e uma solução unitária para cada um são

$$\begin{aligned}
\lambda_1 = 4: \quad & \begin{bmatrix} -1.5 & 1.5 \\ 1.5 & -1.5 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\
& e_{21} = e_{11} \quad \Longrightarrow \quad \boldsymbol{e}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \\
\lambda_2 = 1: \quad & \begin{bmatrix} 1.5 & 1.5 \\ 1.5 & 1.5 \end{bmatrix} \begin{bmatrix} e_{12} \\ e_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\
& e_{22} = -e_{12} \quad \Longrightarrow \quad \boldsymbol{e}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.
\end{aligned}$$

Os vetores com sinais opostos também seriam soluções válidas, pois representam os mesmos eixos com orientação invertida.

Expressões dos componentes principais. Substituindo os autovetores na equação $Y_k = \boldsymbol{e}_k^T \boldsymbol{x}$, obtemos as combinações lineares explícitas dos dois componentes:

$$\begin{aligned}
Y_1 &= \frac{1}{\sqrt{2}} X_1 + \frac{1}{\sqrt{2}} X_2 \approx 0.707 X_1 + 0.707 X_2 \\
Y_2 &= \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_2 \approx 0.707 X_1 - 0.707 X_2
\end{aligned}$$

em que X_1 e X_2 representam as variáveis centralizadas.

Banco de dados transformado. Reunindo os autovetores em $\boldsymbol{E} = [\boldsymbol{e}_1 \ \boldsymbol{e}_2]$, obtemos a matriz de dados expressa no sistema de coordenadas dos componentes principais:

$$\begin{aligned}
\boldsymbol{T} &= \boldsymbol{X}\boldsymbol{E} \\
&= \boldsymbol{X} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\
&= \frac{1}{\sqrt{2}} \begin{bmatrix} -4 & 0 \\ 0 & -2 \\ 0 & 0 \\ 0 & 2 \\ 4 & 0 \end{bmatrix}.
\end{aligned}$$

As linhas de $\boldsymbol{T}$ continuam representando as cinco observações, enquanto suas colunas agora correspondem a Y_1 e Y_2. O elemento t_{ik} é chamado **escore** da observação i no componente k.

4.5 Interpretando os Componentes Principais

Os coeficientes e_{jk} do autovetor $\boldsymbol{e}_k$, frequentemente chamados de **cargas** (*loadings*), indicam o peso de cada variável original X_j na formação do componente Y_k. A magnitude de e_{jk} reflete a importância relativa da variável j no componente k. Infelizmente, a interpretação direta desses coeficientes não é simples.

Uma medida mais interpretável é a **correlação entre os componentes principais e as variáveis originais**, $\mathrm{Corr}\,(Y_k, X_j)$. Ela nos diz o quão "alinhado" um componente está com cada variável original, numa escala padronizada de -1 a 1.

Proposição 4.2 (Correlação entre componentes principais e variáveis originais)**.** *Se $Y_k = \boldsymbol{e}_k^T \boldsymbol{x}$ é o k-ésimo componente principal, associado ao autovalor $\lambda_k > 0$, e $\sigma_{jj} = \mathrm{Var}\,(X_j) > 0$, então*

$$\mathrm{Corr}\,(Y_k, X_j) = \frac{e_{jk}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}},$$

em que e_{jk} é a carga da variável X_j no autovetor $\boldsymbol{e}_k$.

Comprovação. Escrevendo o componente como $Y_k = \sum_{\ell=1}^{p} e_{\ell k} X_\ell$, podemos calcular diretamente sua covariância com X_j. Na derivação abaixo, a terceira igualdade usa a simetria da matriz de covariâncias, enquanto a quarta corresponde à j-ésima equação do sistema $\boldsymbol{\Sigma e}_k = \lambda_k \boldsymbol{e}_k$. Para passar da covariância à correlação, usamos $\mathrm{Var}\,(Y_k) = \lambda_k$ e $\mathrm{Var}\,(X_j) = \sigma_{jj}$:

$$\begin{aligned}
\mathrm{Cov}\,(Y_k, X_j) &= \sum_{\ell=1}^{p} e_{\ell k}\,\mathrm{Cov}\,(X_\ell, X_j) \\
&= \sum_{\ell=1}^{p} e_{\ell k}\sigma_{\ell j} \\
&= \sum_{\ell=1}^{p} \sigma_{j\ell} e_{\ell k} \\
&= \lambda_k e_{jk}, \\
\mathrm{Corr}\,(Y_k, X_j) &= \frac{\mathrm{Cov}\,(Y_k, X_j)}{\sqrt{\mathrm{Var}\,(Y_k)}\sqrt{\mathrm{Var}\,(X_j)}} \\
&= \frac{e_{jk}\lambda_k}{\sqrt{\lambda_k}\sqrt{\sigma_{jj}}} \\
&= \frac{e_{jk}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}}.
\end{aligned}$$

□

Note que a correlação reescala a carga e_{jk} pelo desvio padrão do componente, $\sqrt{\lambda_k}$, e pelo desvio padrão da variável original, $\sqrt{\sigma_{jj}}$. Isso facilita a comparação entre componentes. Quando a ACP é realizada sobre a **matriz de correlação** (ou seja, com dados padronizados), as variâncias σ_{jj} são todas iguais a 1. Nesse caso, a fórmula simplifica para $\mathrm{Corr}\,(Y_k, X_j) = e_{jk}\sqrt{\lambda_k}$.

Para facilitar ainda mais a interpretação, existem representações visuais úteis para a ACP. A mais comum é o **biplot**. Ele combina um gráfico de dispersão das observações projetadas nos componentes principais com os autovetores, permitindo visualizar tanto as direções dos componentes quanto as cargas das observações. Vamos ilustrar e interpretar um biplot no Exemplo 4.3, mais tarde nesse capítulo.

4.6 Escolhendo o Número de Componentes

A ACP expressa os dados em um sistema de eixos rotacionado. A Proposição 4.1 mostra que esse processo não perde nenhuma informação. Agora, quando o interesse é na **redução de dimensionalidade**, precisamos aceitar descartar parte da informação.

Na prática, escolhemos reter $q < p$ componentes principais, descartando $Y_{q+1}, \ldots, Y_p$. As propriedades dos componentes principais garante que esse descarte é tal que a informação preservada é maximizada.

Para que essa redução seja adequada, é preciso escolher q cuidadosamente. Essa escolha é uma balança: poucos componentes tornam a análise mais simples e fácil de interpretar, enquanto mais componentes preservam mais da variância original. Existem na literatura diversos critérios para escolha de q, cada um com suas particularidades. Os mais comuns são:

O **Critério da Variância Explicada Acumulada** é o mais comum deles. Calculamos a proporção da variância total explicada por cada componente e acumulamos essa proporção.

$$\text{Proporção da Variância por } CP_k = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}$$

Em seguida, escolhemos o menor número de componentes q cuja variância explicada acumulada atinja um limiar satisfatório, geralmente entre 70% e 90%. A escolha do limiar depende do contexto da análise.

Já o **Critério do Autovalor (ou Critério de Kaiser)** sugere reter apenas os componentes cujos autovalores (λ_k) são maiores que 1. A intuição por trás dessa regra é mais clara quando a ACP é aplicada sobre a matriz de correlação. Nesse caso, as variáveis originais são padronizadas para ter variância 1. Um componente com autovalor (variância) menor que 1 está, portanto, explicando menos variabilidade do que uma única variável original. Reter tal componente não traria uma "economia" de informação, tornando-o um candidato à exclusão.

Uma terceira alternativa, mais visual, é o **Gráfico do cotovelo (ou Scree Plot)**: A Figura 4.3 ilustra esse método. O gráfico mostra a variância explicada acumulada em função do número de componentes principais, com um "cotovelo" indicando o ponto de corte. A ideia é reter os componentes que aparecem antes do cotovelo, pois eles são os que contribuem mais significativamente para a variância total.

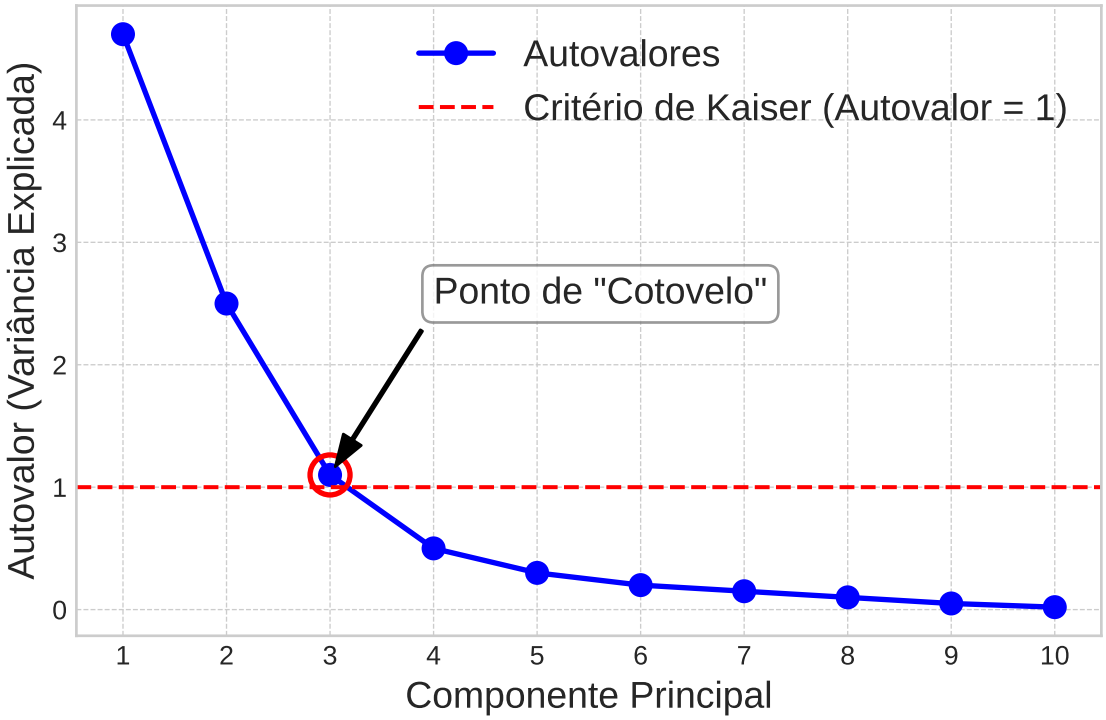

Figura 4.3: Exemplo de um Scree Plot. O 'cotovelo' em k=3 sugere a retenção de 3 componentes.

4.7 Exemplo Prático

Exemplo 4.3 (Medidas corporais dos pinguins)**.** Vamos exemplificar o uso da ACP no conjunto de dados Palmer Penguins, ele contem medidas corporais de pinguins: comprimento e profundidade do bico, comprimento da nadadeira e massa corporal. As três primeiras estão em milímetros; a última, em gramas. Por isso, uma ACP baseada diretamente na matriz de covariâncias daria peso excessivo à massa corporal. Começamos removendo os valores ausentes e padronizando cada variável.

Tabela 4.1: Matriz de correlações das medidas corporais dos pinguins.

	Comp. do bico	Prof. do bico	Comp. da nadadeira	Massa corporal
Comp. do bico	1.00	-0.24	0.66	0.60

Tabela 4.1: Matriz de correlações das medidas corporais dos pinguins.

	Comp. do bico	Prof. do bico	Comp. da nadadeira	Massa corporal
Prof. do bico	-0.24	1.00	-0.58	-0.47
Comp. da nadadeira	0.66	-0.58	1.00	0.87
Massa corporal	0.60	-0.47	0.87	1.00

O conjunto de dados tem 342 observações completas. Note que a variância total é $\mathrm{tr}\,(\boldsymbol{R}) = 1 + 1 + 1 + 1 = 4$, pois todas as variáveis padronizadas têm variância 1. A matriz também antecipa parte da estrutura que a ACP encontrará. Comprimento da nadadeira e massa corporal, por exemplo, têm correlação 0.87 isso indica que essas variáveis carregam informação parecida.

Agora fazemos a decomposição espectral de $\boldsymbol{R}$. As colunas da matriz `E` são os autovetores e definem as combinações lineares; a matriz `T` contém os escores das observações. A Tabela 4.2 apresenta a variância explicada, enquanto a Figura 4.4 mostra os mesmos autovalores graficamente.

> **Nota**
>
> Existem bibliotecas que facilitam a aplicação da ACP, como `sklearn.decomposition.PCA`. Agora, como os calculos são extremamente simples, vamos apenas usar a decomposição espectral fornecida pelo NumPy.

O primeiro componente tem variância 2.75 e concentra 68.8% da variância total. Os dois primeiros, juntos, concentram 88.2%. Se o interesse do problema fosse a redução da dimensionalidade, seria razoável manter dois componentes, isso porque eles preservam 88,2% da variância e permitem analisar os dados em um espaço bidimensional. Aqui, vamos focar na interpretação, vamos interpretar CP1 e CP2.

As correlações tornam a interpretação direta. O primeiro componente cresce com o comprimento da nadadeira, a massa corporal e o comprimento do bico, mas diminui com a profundidade do bico. Ele separa, portanto, pinguins maiores e mais pesados, com nadadeiras longas e bicos mais compridos e menos profundos, de pinguins com o perfil oposto.

O segundo componente está ligado principalmente às duas medidas do bico. Comprimento da nadadeira e massa corporal têm correlações próximas de zero com esse eixo. Assim, Y_2 descreve uma variação conjunta no comprimento e na profundidade do bico.

O círculo de correlações, mostrado na Figura 4.5, é outra ferramente interpretativa visual. Ele permite visualizar as correlações entre as variáveis em um plano bidimensional. Note que, variáveis correlacionadas, como comprimento da nadadeira e massa corporal, tem vetores parecidos na representação. Enquanto isso, os vetores de profundidade do bico e comprimento da nadadeira são opostos, indicando uma correlação negativa. Variáveis pouco correlacionadas, como profundidade do bico e comprimento do bico, formam um angulo quase perpendicular entre eles.

Tabela 4.2: Variância explicada pelos componentes principais.

	Autovalor	Proporção (%)	Acumulada (%
CP1	2.75	68.84	68.84
CP2	0.77	19.31	88.16
CP3	0.37	9.13	97.29
CP4	0.11	2.71	100.00

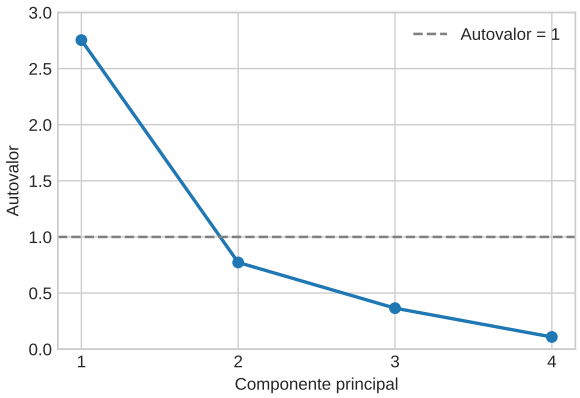

Figura 4.4: Gráfico do cotovelo da ACP dos pinguins.

Tabela 4.3: Correlações entre as medidas corporais e os dois primeiros componentes.

	CP1	CP2
Comp. do bico	0.755	0.525
Prof. do bico	-0.664	0.701
Comp. da nadadeira	0.956	0.002
Massa corporal	0.910	0.074

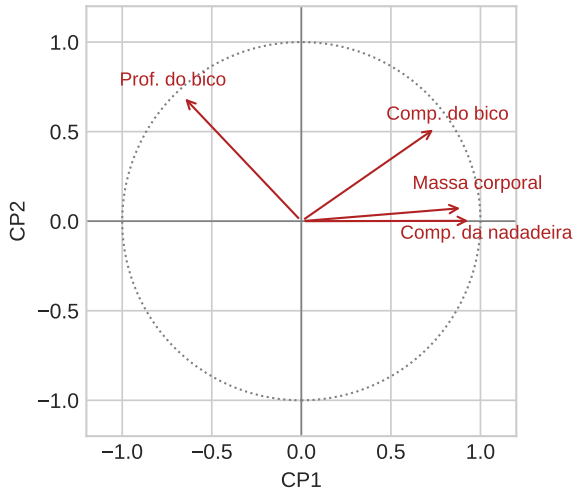

Figura 4.5: Círculo de correlações das medidas corporais dos pinguins.

O banco de dados original contém também as espécies dos pinguins. Elas são três: Adelie, Chinstrap e Gentoo. Vamos construir um biplot, colorindo os pontos de acordo com a espécie. No biplot da Figura 4.6, os pontos mostram os escores brutos nos dois primeiros componentes e as setas indicam a direção em que cada medida aumenta. A espécie não entrou no cálculo da ACP; ela aparece apenas como cor para ajudar na leitura.

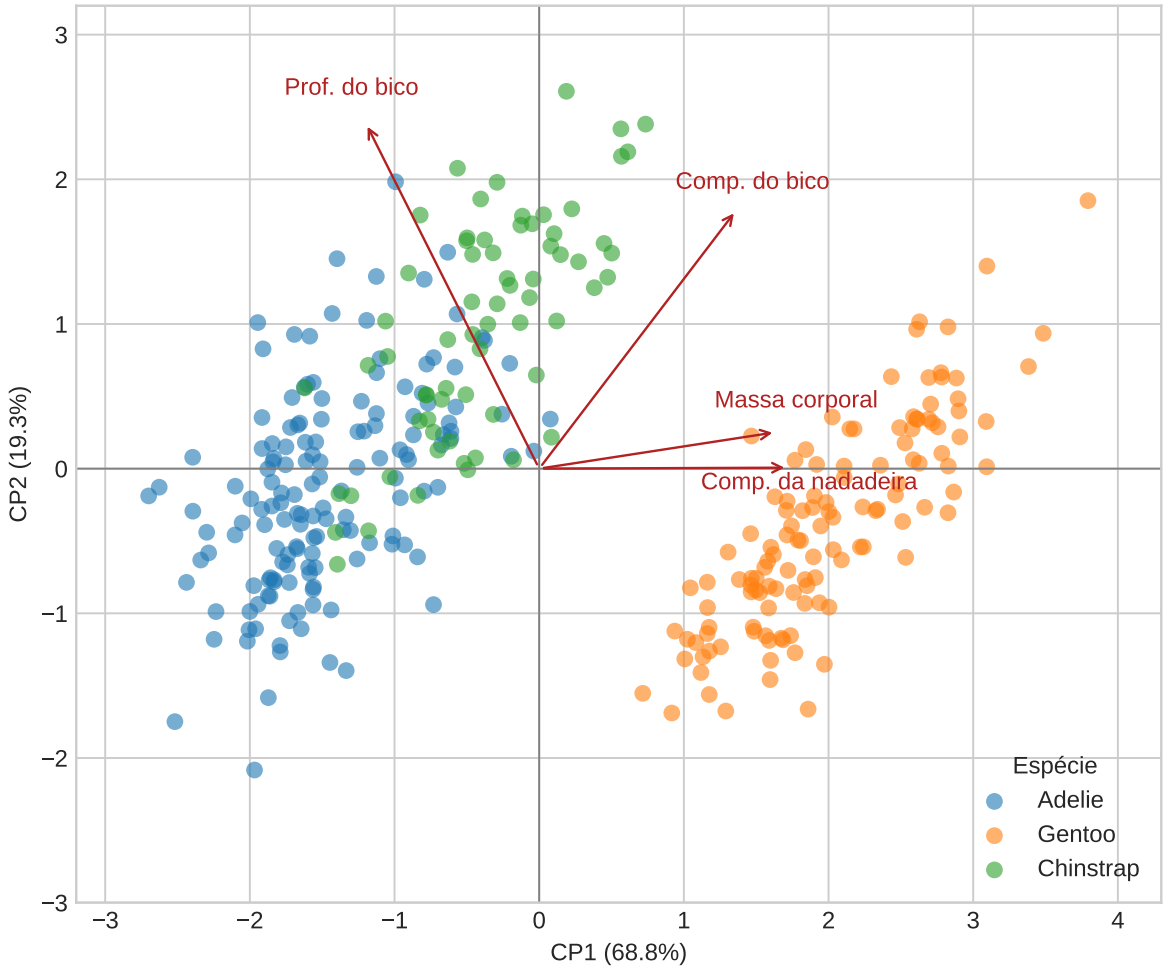

Figura 4.6: Biplot das medidas corporais dos pinguins, com os escores brutos nos dois primeiros componentes.

Os Gentoo aparecem principalmente à direita, na direção das setas de massa corporal e comprimento da nadadeira, indicando que esses pinguins são mais pesados e maiores. Os Adelie se concentram à esquerda, ilustrando que esses pinguins são mais leves e menores.

Os Chinstrap ocupam a parte superior, associada a valores altos nas duas medidas do bico, mostrando que esses pinguin tem os bicos maiores e mais profundos.

Como os dois eixos preservam 88,2% da variância, podemos concluir que essas interpretações são robustas. Em casos onde os componentes descartam mais informação, as interpretações devem ser mais cautelosas.

4.8 Exercícios

Exercício 4.1. Seja $\boldsymbol{x}$ um vetor aleatório de dimensão p com matriz de covariâncias $\boldsymbol{\Sigma} = \mathrm{Cov}\,(\boldsymbol{x})$, real e simétrica. Os componentes principais de $\boldsymbol{x}$ são definidos pela transformação linear $\boldsymbol{y} = \boldsymbol{E}^T\boldsymbol{x}$, em que $\boldsymbol{E}$ é a matriz ortogonal $p \times p$ cujas colunas são os autovetores normalizados de $\boldsymbol{\Sigma}$.

Mostre que os componentes principais $Y_1, \ldots, Y_p$ são não correlacionados entre si e têm variâncias $\lambda_1, \ldots, \lambda_p$, respectivamente. Ou seja, mostre que a matriz de covariâncias de $\boldsymbol{y}$ é a matriz diagonal contendo os autovalores de $\boldsymbol{\Sigma}$:

$$\mathrm{Cov}\,(\boldsymbol{y}) = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}$$

Exercício 4.2. Um vendedor de ferramentas automotivas coletou dados a respeito de um determinado produto de seu portfólio: X_1: Qualidade das peças; X_2: Facilidade de utilização; X_3: Qualidade da embalagem. Após analisar os dados, ele calculou a seguinte matriz de covariância amostral:

$$\boldsymbol{S} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

a) Apenas olhando para essa matriz, o que podemos esperar de uma ACP nesse caso?

b) Realize a decomposição da matriz para obter os componentes principais. Depois, compare os resultados com as presunções feitas no item **(a)**.

Exercício 4.3. Dois grupos realizaram uma ACP sobre a mesma matriz de dados. Para um dos componentes, o primeiro grupo obteve o autovetor $\boldsymbol{e}_k$ e os escores t_{ik}. O segundo encontrou $-\boldsymbol{e}_k$ e escores com os sinais invertidos. Os gráficos produzidos pelos grupos aparecem refletidos, e cada grupo acredita que o outro cometeu um erro.

a) Mostre que, se $\boldsymbol{e}_k$ é um autovetor de $\boldsymbol{\Sigma}$ associado a λ_k, então $-\boldsymbol{e}_k$ também é.

b) Compare a variância explicada, as cargas correlacionais e os escores obtidos pelos dois grupos. O que muda e o que permanece igual?

c) Afinal, os grupos encontraram componentes diferentes?

Exercício 4.4 (Correlação e Variância Explicada)**.** Considere um vetor aleatório bidimensional $\boldsymbol{x} = (X_1, X_2)^T$, em que X_1 e X_2 são padronizadas e $\mathrm{Corr}\,(X_1, X_2) = \rho$, com $0 < \rho \leq 1$.

Mostre que a proporção da variância total explicada pelo primeiro componente principal varia linearmente com ρ. Determine essa função e interprete seu comportamento quando $\rho \to 0^+$ e quando $\rho = 1$.

Exercício 4.5 (Autovalores Repetidos)**.** Dois pesquisadores aplicaram ACP à matriz de covariâncias

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

O primeiro escolheu $(1, 0, 0)^T$ e $(0, 1, 0)^T$ como os autovetores associados ao autovalor 4. O segundo escolheu, para algum ângulo ϕ,

$$\boldsymbol{v}_1 = \begin{bmatrix} \cos\phi \\ \sin\phi \\ 0 \end{bmatrix} \qquad \text{e} \qquad \boldsymbol{v}_2 = \begin{bmatrix} -\sin\phi \\ \cos\phi \\ 0 \end{bmatrix}.$$

a) Verifique que $\boldsymbol{v}_1$ e $\boldsymbol{v}_2$ são autovetores unitários e ortogonais associados ao autovalor 4.

b) Algum dos pesquisadores está errado? Explique como a repetição dos autovalores afeta a ACP.

Exercício 4.6 (Projeção e Reconstrução)**.** Uma equipe pretende comprimir uma matriz de dados centralizados $\boldsymbol{X}$, de dimensão $n \times p$, mantendo apenas os primeiros q componentes principais. Para isso, os primeiros q autovetores são reunidos na matriz $\boldsymbol{E}_q$, e os escores correspondentes são dados por $\boldsymbol{T}_q = \boldsymbol{X}\boldsymbol{E}_q$.

a) Mostre como reconstruir uma aproximação $\widehat{\boldsymbol{X}}$ dos dados originais a partir de $\boldsymbol{T}_q$ e $\boldsymbol{E}_q$. Em seguida, escreva essa reconstrução diretamente em função de $\boldsymbol{X}$.

b) Defina $\boldsymbol{P}_q = \boldsymbol{E}_q\boldsymbol{E}_q^T$ e mostre que essa matriz é simétrica e idempotente, no sentido da Definição 2.4. Use esse resultado para interpretar geometricamente a reconstrução.

c) Mostre que o resíduo $\boldsymbol{X} - \widehat{\boldsymbol{X}}$ é ortogonal às direções mantidas. O que essa propriedade diz sobre a informação que foi descartada?

d) Mostre que o erro quadrático de reconstrução satisfaz

$$\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \widehat{x}_{ij})^2 = \sum_{k=q+1}^{p}\lambda_k.$$

Use essa relação para explicar por que os primeiros q componentes fornecem a melhor aproximação de dimensão q.

Exercício 4.7 (Composição Química de Vinhos)**.** O conjunto Wine, da UCI Machine Learning Repository, reúne 178 vinhos produzidos na mesma região da Itália a partir de três cultivares. Para cada vinho, foram registradas 13 medidas de sua composição química.

O que uma ACP permite concluir a respeito desses vinhos? Faça uma análise completa usando as medidas químicas como variáveis ativas. Justifique as decisões tomadas, interprete os componentes e use gráficos, incluindo um biplot, para sustentar suas conclusões. O cultivar não deve participar do cálculo dos componentes, mas pode ser usado depois para ajudar a interpretar os escores.

Leituras recomendadas

As leituras abaixo não são pré-requisitos para acompanhar o livro. Elas são boas portas de entrada quando você quiser ver uma apresentação mais completa, outra notação ou uma discussão mais longa de algum método.