Estatística Multivariada 2

Victor Coscrato

2025-07-31

Índice

Prefácio

Seja Bem-vindo! Este livro aborda uma variedade de técnicas de análise multivariada, essenciais para a compreensão de dados complexos em diversas áreas do conhecimento.

Este material foi elaborado especialmente para estudantes tendo um primeiro contato com técnicas estatísticas multivariadas. São cobertos os métodos a seguir:

- Análise de Componentes Principais
- Análise Fatorial
- Análise de Agrupamento
- Análise Discriminante
- Análise de Correlação Canônica
- Análise de Correspondência

O objetivo é fornecer uma base sólida e prática para a aplicação dessas técnicas. antes, temos uma breve introdução dos conceitos fundamentais que norteiam a análise multivariada e algumas definições e resultados vetores e matrizes que são importantes para o acompanhamento do livro.

Part 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 três ou mais variáveis para extrair padrões e estruturas que de outra forma permaneceriam ocultos.

Cada observação em um estudo — seja um paciente descrito por indicadores de saúde, um consumidor por hábitos de compra, ou uma empresa por métricas financeiras — pode ser representada como um **vetor de observações**. A análise multivariada nos fornece as ferramentas para entender a estrutura de dependência e interdependência dentro desses vetores.

1.1 Por que usar Análise Multivariada?

A análise multivariada é motivada pela necessidade de extrair informações significativas de conjuntos de dados complexos. Ao invés de analisar variáveis de forma isolada, essas técnicas 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.

1.2 Visão Geral das Técnicas Multivariadas

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	Relação entre conjuntos de variáveis	Quantitativas	Dependência
Análise de Correspondência	Relação entre variáveis categóricas	Qualitativas	Interdependência

Figura 1.1: Diagrama de decisão para escolha de técnica de Análise Multivariada. Nós enfatizados fundo azul escuro 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.

2 Vetores Aleatórios e Suas Características

No capítulo anterior, estabelecemos a motivação para a análise multivariada: a necessidade de entender sistemas complexos onde múltiplas variáveis interagem. Para fazer isso de maneira formal e rigorosa, precisamos primeiro definir o objeto matemático central de nosso estudo. Em vez de começar com uma tabela de dados, começamos com o conceito que gera esses dados: o **vetor aleatório**.

2.1 O Vetor Aleatório

Imagine que, para uma população de interesse (e.g., todos os estudantes de uma universidade), associamos a cada membro um conjunto de p características que nos interessam (e.g., nota em matemática, nota em história, horas de estudo). Antes de observarmos um membro específico, os valores dessas características são incertos. Podemos modelar essa incerteza tratando cada característica como uma variável aleatória.

Definição 2.1. Um **vetor aleatório** $\mathbf{x}$ é um vetor-coluna cujos componentes são p variáveis aleatórias, $X_1, X_2, \ldots, X_p$.

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

Este vetor é a representação matemática de uma "observação multivariada" em nível populacional. 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., $\mathbf{x}$) 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., $\mathbf{X}$) será usada para a **matriz de dados** (amostral).

2.2 Parâmetros Populacionais

Assim como variáveis aleatórias escalares são caracterizadas por parâmetros como a média (expectativa) e a variância, os vetores aleatórios também o são. Esses parâmetros descrevem a tendência central, a dispersão e as inter-relações das variáveis que compõem o vetor.

Definição 2.2. O **vetor de médias populacional**, denotado por $\boldsymbol{\mu}$, é o vetor das expectativas de cada uma de suas variáveis componentes.

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

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

Definição 2.3. 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, $\sigma_{jk} = \text{Cov}(X_j, X_k) = E[(X_j - \mu_j)(X_k - \mu_k)]$.

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

- **Diagonal (σ_{jj}):** As **variâncias**, $\text{Var}(X_j)$, 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 $\text{Cov}(X_j, X_k) = \text{Cov}(X_k, X_j)$, o que implica $\sigma_{jk} = \sigma_{kj}$.

Definição 2.4. A **matriz de correlações populacional**, denotada por $\mathbf{P}$, é uma versão reescalada da matriz de covariâncias, com elementos $\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}}$.

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

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

2.3 Propriedades de μ e Σ

As propriedades de combinações lineares são generalizações diretas dos resultados univariados. Seja $\mathbf{x}$ um vetor aleatório p-dimensional com média $\boldsymbol{\mu}$ e covariância $\boldsymbol{\Sigma}$. Sejam $\mathbf{c}$ um vetor de constantes $p \times 1$ e $\mathbf{A}$ uma matriz de constantes $q \times p$.

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

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

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

$$\text{Cov}[\mathbf{A}\mathbf{x} + \mathbf{c}] = \mathbf{A}\text{Cov}[\mathbf{x}]\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$$

No próximo capítulo, veremos como, na prática, não temos acesso a esses parâmetros populacionais ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{P}$), mas podemos usar dados observados para obter estimativas confiáveis deles.

3 Amostra e Estimação de Parâmetros

No capítulo anterior, introduzimos os conceitos teóricos que descrevem uma população multivariada: o vetor de médias $\boldsymbol{\mu}$ e a matriz de covariâncias $\boldsymbol{\Sigma}$. Esses parâmetros são construções ideais que existem no nível populacional. Na prática, quase nunca temos acesso a toda a população para calculá-los diretamente.

O nosso trabalho como estatísticos e analistas de dados é fazer inferências sobre esses parâmetros desconhecidos com base em um conjunto limitado de dados. Fazemos isso através da **amostragem**.

3.1 Da População à Amostra

Assumimos que coletamos uma **amostra aleatória** de n observações da população. Cada observação, $\mathbf{x}_i$ (com $i = 1, \dots, n$), é uma **realização** independente do vetor aleatório $\mathbf{x}$ que definimos no capítulo anterior.

A coleção de todas essas observações forma o nosso conjunto de dados. É aqui que, finalmente, introduzimos a **matriz de dados**, $\mathbf{X}$, uma estrutura central em toda a análise multivariada aplicada.

A matriz $\mathbf{X}$ é uma matriz de dimensão $n \times p$, onde cada linha é uma observação multivariada e cada coluna representa uma variável.

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{3.1}$$

O elemento x_{ij} representa o valor da j-ésima variável para a i-ésima observação. Com esta matriz em mãos, nosso objetivo é calcular quantidades que sirvam como boas **estimativas** para os parâmetros populacionais $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$.

3.2 Estimadores Amostrais

As quantidades que calculamos a partir da amostra são chamadas de **estatísticas amostrais** ou **estimadores**, e são as contrapartes amostrais dos parâmetros populacionais.

Definição 3.1. O estimador de $\boldsymbol{\mu}$ é o **vetor de médias amostral**, $\bar{\mathbf{x}}$, cujos componentes $\bar{x}_j$ são a média das observações para a j-ésima variável.

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

Definição 3.2. O estimador de $\boldsymbol{\Sigma}$ é a **matriz de covariâncias amostral**, $\mathbf{S}$. 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:

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

> **i** Por que dividir por $n-1$?
>
> A divisão por $n-1$ (graus de liberdade) em vez de n é feita para garantir que s_{jk} seja um estimador não-viesado de σ_{jk}, ou seja, $E[s_{jk}] = \sigma_{jk}$.

Definição 3.3. O estimador de $\mathbf{P}$ é a **matriz de correlações amostral**, $\mathbf{R}$, 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:

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

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

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

A maior parte das técnicas que veremos neste livro opera sobre as matrizes $\mathbf{S}$ ou $\mathbf{R}$ para fazer inferências sobre a estrutura da população.

4 A Distribuição Normal Multivariada

Até agora, discutimos os parâmetros de um vetor aleatório ($\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$) sem assumir uma forma específica para sua distribuição de probabilidade. No entanto, para desenvolvermos uma teoria de inferência estatística robusta e compreendermos o funcionamento de muitas técnicas clássicas, precisamos de um modelo de distribuição de referência. Na análise multivariada, esse papel é desempenhado pela **distribuição Normal Multivariada (NMV)**.

A NMV é uma generalização da distribuição normal (Gaussiana) para o caso de p variáveis. Ela é, de longe, a distribuição mais importante da análise multivariada, por várias razões: 1. Muitos fenômenos naturais podem ser aproximados pela NMV. 2. O Teorema do Limite Central, em sua forma multivariada, garante que a média de vetores aleatórios de (quase) qualquer distribuição tende a se comportar como uma NMV para amostras grandes. 3. Suas propriedades matemáticas são extremamente convenientes e bem compreendidas, o que facilita muito a derivação de resultados teóricos.

4.1 A Função de Densidade

Um vetor aleatório $\mathbf{x}$ de dimensão p segue uma distribuição Normal Multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$ (positiva definida), denotado por $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, se sua função de densidade de probabilidade (FDP) for dada por:

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

Onde: - $|\boldsymbol{\Sigma}|$ é o determinante da matriz de covariâncias. - $\boldsymbol{\Sigma}^{-1}$ é a inversa da matriz de covariâncias.

Apesar de parecer intimidante, a estrutura da FDP é bastante lógica. O termo no expoente, $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, é uma forma quadrática que mede a "distância" do ponto $\mathbf{x}$ ao centro $\boldsymbol{\mu}$, ponderada pela estrutura de covariância $\boldsymbol{\Sigma}$. Essa distância é chamada de **distância de Mahalanobis**. Quanto maior essa distância, menor o valor da função de densidade, o que faz todo o sentido.

4.2 Propriedades da Distribuição Normal Multivariada

A popularidade da NMV vem de suas propriedades elegantes:

1. **Combinações Lineares:** Se $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então qualquer combinação linear de suas componentes, $\mathbf{a}'\mathbf{x} = a_1X_1 + \cdots + a_pX_p$, segue uma distribuição normal univariada. Mais geralmente, se $\mathbf{A}$ é uma matriz de constantes, então $\mathbf{Ax}$ segue uma distribuição Normal Multivariada.
2. **Distribuições Marginais:** Qualquer subconjunto de variáveis de um vetor Normal Multivariado também segue uma distribuição Normal Multivariada. Por exemplo, se $\mathbf{x} = [X_1, X_2, X_3]'$ é NMV, então o vetor $[X_1, X_3]'$ também é NMV.
3. **Independência e Covariância Zero:** Para a maioria das distribuições, covariância zero não implica independência. No entanto, para a NMV, essa implicação é verdadeira. Se um subconjunto de variáveis em um vetor NMV tem covariância zero com outro subconjunto, então esses dois subconjuntos são **independentes**. Esta é uma propriedade extremamente poderosa.

4.3 Visualizando a Normal Bivariada

Para ganhar intuição, é útil visualizar o caso bivariado ($p = 2$). A função de densidade forma uma superfície em forma de sino no espaço 3D. Os contornos de densidade constante, quando projetados no plano (x_1, x_2), formam elipses.

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

A forma e a orientação dessas elipses são inteiramente determinadas pela matriz de covariâncias $\boldsymbol{\Sigma}$.

- **Se $\sigma_{12} = 0$ e $\sigma_{11} = \sigma_{22}$:** As variáveis são não correlacionadas (e, portanto, independentes) e têm a mesma variância. Os contornos são **círculos**.
- **Se $\sigma_{12} = 0$ e $\sigma_{11} \neq \sigma_{22}$:** As variáveis são não correlacionadas, mas com variâncias diferentes. Os contornos são **elipses alinhadas com os eixos** de coordenadas.
- **Se $\sigma_{12} \neq 0$:** As variáveis são correlacionadas. Os contornos são **elipses rotacionadas**. A direção do eixo principal da elipse é determinada pelos autovetores de $\boldsymbol{\Sigma}$, e o comprimento dos eixos é determinado pelos autovalores.

Essa conexão entre a álgebra da matriz $\boldsymbol{\Sigma}$ e a geometria da distribuição de dados é um dos temas mais importantes da análise multivariada e será a base para a técnica de Componentes Principais, que exploraremos mais adiante.

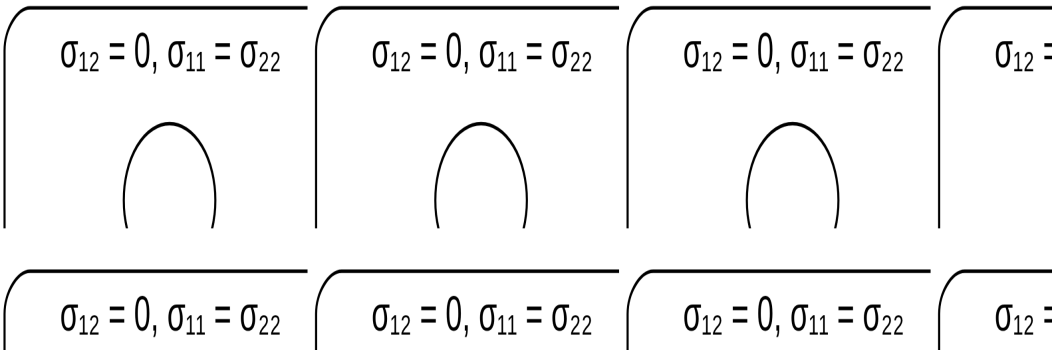

Figura 4.1: Contornos de densidade para uma distribuição Normal Bivariada, ilustrando o efeito da matriz de covariância.

5 Fundamentos Matemáticos: Álgebra Matricial

Com os conceitos estatísticos fundamentais estabelecidos, voltamos nossa atenção para as ferramentas matemáticas necessárias para manipular esses objetos. A linguagem da análise multivariada é a álgebra linear.

Neste capítulo, revisaremos conceitos-chave — formas quadráticas, matrizes positiva-definidas e a decomposição espectral — que são a base para muitas das técnicas que veremos, como a Análise de Componentes Principais (PCA).

5.1 Formas Quadráticas

Uma forma quadrática é uma função polinomial de várias variáveis que contém apenas termos de grau dois. Para um vetor $\mathbf{x}$ de dimensão $p \times 1$ e uma matriz simétrica $\mathbf{A}$ de dimensão $p \times p$, a forma quadrática é expressa como:

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

Um exemplo fundamental que já encontramos é a distância de Mahalanobis ao quadrado, $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, que aparece no expoente da distribuição normal multivariada. Esta forma quadrática define as elipses de contorno de densidade constante da distribuição.

5.2 Matrizes positiva-Definidas

O conceito de positividade para um número escalar é estendido para matrizes através das formas quadráticas. Uma matriz simétrica $\mathbf{A}$ é dita:

- **positiva-definida** se $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ para todos os vetores não-nulos $\mathbf{x}$.
- **positiva-semidefinida** se $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ para todos os vetores não-nulos $\mathbf{x}$.

Propriedades de uma matriz positiva-definida: - Todos os seus autovalores são estritamente positivos ($\lambda_i > 0$). - A matriz é invertível (não-singular). - Seu determinante é positivo.

Matrizes de covariância ($\mathbf{\Sigma}$) e correlação ($\mathbf{R}$) são, por natureza, positiva-semidefinidas. Para que a função de densidade da normal multivariada seja bem definida e a matriz $\mathbf{\Sigma}$ seja invertível, exigimos que ela seja **positiva-definida**. Isso implica que nenhuma variável no vetor aleatório é uma combinação linear perfeita de outras (ou seja, não há redundância linear total nos dados).

5.3 Decomposição Espectral

A decomposição espectral (ou de autovalores) é uma fatoração de uma matriz simétrica em seus autovalores e autovetores. Ela revela a estrutura fundamental da transformação linear representada pela matriz.

Toda matriz simétrica $\mathbf{A}$ de dimensão $p \times p$ pode ser reescrita como:

$$\mathbf{A} = \mathbf{E}\Lambda\mathbf{E}'$$

Onde:

- $\lambda_1, \ldots, \lambda_p$ são os **autovalores** de $\mathbf{A}$.
- $\mathbf{e}_1, \ldots, \mathbf{e}_p$ são os **autovetores** ortonormais correspondentes.
- Λ é a matriz diagonal com os autovalores λ_i na diagonal.
- $\mathbf{E}$ é a matriz ortogonal cujas colunas são os autovetores $\mathbf{e}_i$.

Exemplo 5.1. Vamos decompor a seguinte matriz de covariâncias $\mathbf{S}$:

$$\mathbf{S} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

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

A decomposição é $\mathbf{S} = \mathbf{E}\Lambda\mathbf{E}'$, com:

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

Isso nos diz que a maior variância dos dados (igual a 3) está na direção do vetor $(1, 1)$, enquanto a variância na direção ortogonal $(1, -1)$ é menor (igual a 1).

6 Medidas de Distância e Similaridade

Um conceito fundamental que permeia quase todas as técnicas de análise multivariada é a medição da "proximidade" ou "distância" entre observações. Seja para agrupar dados semelhantes, classificar uma nova observação ou entender a estrutura de um conjunto de dados, essas medidas determinam uma forma quantitativa para expressar o quão perto ou longe duas observações estão uma da outra no espaço p-dimensional.

6.1 Distâncias vs. Dissimilaridades

Formalmente, uma função $d(\cdot, \cdot)$ é considerada uma **métrica de distância** se satisfaz as seguintes propriedades para quaisquer pontos $\mathbf{x}, \mathbf{y}, \mathbf{z}$:

1. **Não-negatividade:** $d(\mathbf{x}, \mathbf{y}) \geq 0$
2. **Identidade:** $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$
3. **Simetria:** $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
4. **Desigualdade Triangular:** $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$

No entanto, em muitos contextos práticos, utilizamos medidas que não satisfazem todas essas propriedades, mas que ainda são extremamente úteis para quantificar o quão diferentes dois objetos são. Usamos o termo mais geral **medida de dissimilaridade** para nos referirmos a qualquer função que indique o grau de diferença entre dois pontos, onde valores pequenos indicam semelhança e valores grandes indicam diferença.

Um exemplo clássico de uma medida de dissimilaridade que não é uma métrica de distância estrita é a **distância Euclidiana quadrática**, $d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})$. Ela viola a propriedade da desigualdade triangular, mas pode ser usada em algoritmos como o K-médias e o método de Ward por suas convenientes propriedades computacionais (evitar o cálculo da raiz quadrada economiza tempo).

Nas seções a seguir, apresentamos algumas das medidas de dissimilaridade e distância mais populares. A escolha da medida ideal é um campo vasto e depende fundamentalmente da natureza dos dados e do objetivo da análise.

6.2 Medidas para Dados Contínuos

Definição 6.1. A **Distância Euclidiana** é a métrica de distância mais comum e corresponde à noção intuitiva de distância em linha reta entre dois pontos.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}$$

Definição 6.2. A **Distância de Manhattan** (ou *City-Block*) calcula a distância como a soma das diferenças absolutas entre as coordenadas dos pontos. É como se deslocar entre dois pontos em uma cidade, movendo-se apenas ao longo das ruas (horizontais e verticais).

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Esta medida é, em geral, mais robusta a *outliers* do que a distância Euclidiana.

Definição 6.3. A **Distância de Minkowski** é uma generalização tanto da Euclidiana quanto da de Manhattan.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^m \right)^{1/m}$$

- Se $m = 2$, temos a distância Euclidiana.
- Se $m = 1$, temos a distância de Manhattan.

Quanto maior o valor de m, mais peso é dado às maiores diferenças entre as coordenadas.

> ⚠ Limitação das Distâncias Comuns
>
> As distâncias Euclidiana, de Manhattan e de Minkowski são sensíveis às escalas das variáveis. Se uma variável tiver uma magnitude muito maior que as outras, ela dominará o cálculo da distância. Por isso, é prática comum **padronizar** as variáveis (subtrair a média e dividir pelo desvio padrão) antes de calcular a matriz de distâncias.

A **Distância de Mahalanobis** é uma medida de distância estatística que leva em conta a correlação entre as variáveis e é invariante à escala.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)}$$

Onde $\mathbf{S}^{-1}$ é a inversa da matriz de covariâncias amostral. Ela mede a distância entre os pontos em unidades de desvio padrão, ajustando a contribuição de cada variável pela estrutura de covariância dos dados. Já encontramos essa forma quadrática no expoente da distribuição Normal Multivariada (Capítulo 4).

6.3 Medidas para Variáveis Binárias

Quando os dados são binários (0 ou 1), a interpretação da distância muda. A distância Euclidiana quadrática, por exemplo, simplesmente conta o número de posições em que os dois vetores discordam.

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0, & \text{se } x_{ij} = x_{kj} \\ 1, & \text{se } x_{ij} \neq x_{kj} \end{cases}$$

O problema é que essa abordagem dá o mesmo peso para uma concordância de `1-1` e uma concordância de `0-0`. Em muitos contextos, a ausência conjunta de uma característica (concordância `0-0`) é menos informativa do que a presença conjunta (concordância `1-1`).

Para lidar com isso, podemos construir uma tabela de contingência para duas observações $\mathbf{x}_i$ e $\mathbf{x}_j$:

	Observação j = 1	Observação j = 0	Total
Obs i = 1	a	b	a+b
Obs i = 0	c	d	c+d
Total	a+c	b+d	p

Onde: - a: número de variáveis onde $x_{ik} = 1$ e $x_{jk} = 1$. - d: número de variáveis onde $x_{ik} = 0$ e $x_{jk} = 0$. - b e c: número de variáveis onde há discordância.

A distância Euclidiana quadrática corresponde a $b + c$.

Definição 6.4. O **Coeficiente de Jaccard** é uma medida de *similaridade* para dados binários que ignora as concordâncias `0-0`.

$$J(\mathbf{x}_i, \mathbf{x}_j) = \frac{a}{a + b + c}$$

A **Distância de Jaccard** é a sua contraparte de dissimilaridade, definida como $1 - J(\mathbf{x}_i, \mathbf{x}_j)$.

Definição 6.5. O **Coeficiente de Correspondência Simples** (*Simple Matching Coefficient*, SMC) considera tanto as presenças (`1-1`) quanto as ausências (`0-0`) como concordâncias. É útil quando a ausência de uma característica é tão informativa quanto a sua presença.

$$SMC = \frac{a + d}{a + b + c + d}$$

Definição 6.6. O **Coeficiente de Dice** (ou Sørensen-Dice) é outra medida de similaridade que, assim como Jaccard, ignora as concordâncias `0-0`. No entanto, ele dá um peso maior às concordâncias `1-1`.

$$Dice = \frac{2a}{2a + b + c}$$

Definição 6.7. O **Coeficiente de Russell-Rao** é uma medida mais simples que calcula a proporção de presenças conjuntas em relação ao total de variáveis.

$$RR = \frac{a}{a + b + c + d}$$

6.4 Matriz de Distâncias

Uma vez escolhida a medida de dissimilaridade, é comum pré-calcular todas as distâncias entre os pares de observações e organizá-las em uma **matriz de distâncias D**, de dimensão $n \times n$.

$$\mathbf{D} = \begin{pmatrix} 0 & d(\mathbf{x}_1, \mathbf{x}_2) & \cdots & d(\mathbf{x}_1, \mathbf{x}_n) \\ d(\mathbf{x}_2, \mathbf{x}_1) & 0 & \cdots & d(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(\mathbf{x}_n, \mathbf{x}_1) & d(\mathbf{x}_n, \mathbf{x}_2) & \cdots & 0 \end{pmatrix}$$

Esta matriz é simétrica, ou seja $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$, e possui zeros na diagonal principal. Ela serve como a entrada para muitos algoritmos de agrupamento, especialmente os hierárquicos.

Part II

Parte II: Métodos multivariados

7 Análise de Componentes Principais

A Análise de Componentes Principais (ACP ou *PCA* do acrônimo em inglês) é uma técnica estatística multivariada que transforma um conjunto de variáveis possivelmente correlacionadas em um novo conjunto de variáveis não correlacionadas, chamadas de **componentes principais**. O objetivo primário da ACP é a **redução de dimensionalidade**: representar a variabilidade presente nos dados originais com um número menor de variáveis, minimizando a perda de informação.

Cada componente principal é uma combinação linear das variáveis originais. O primeiro componente principal é construído para capturar a maior variabilidade possível nos dados. O segundo componente principal, ortogonal ao primeiro, captura a maior parte da variabilidade restante, e assim por diante. Ao final, o número de componentes principais é igual ao número de variáveis originais, mas a expectativa é que os primeiros componentes concentrem a maior parte da informação relevante.

Geometricamente, a ACP é uma projeção do espaço original de variáveis para um outro espaço com características mais interessantes: A variância dos dados é concentrada em direções especificas (os componentes principais) e não existem correlações entre os novos eixos.

Para construir a intuição geométrica, vamos começar com um exemplo simples, em duas dimensões.

Exemplo 7.1. Suponha que coletamos dados de duas variáveis, **Peso** (em kg) e **Altura** (em cm), de um grupo de 10 pessoas.

Tabela 7.1: Tabela de dados com Peso (kg) e Altura (cm).

	Peso	Altura
	65	170
	72	182
	58	165
	81	190
	75	178
	60	168
	68	175
	70	172
	78	185
	62	169
Média	68.90	175.40
Variância	59.88	66.71

Ao plotarmos esses dados (já centralizados), obtemos a nuvem de pontos abaixo. O sistema de eixos em preto (Peso, Altura) é a nossa perspectiva padrão.

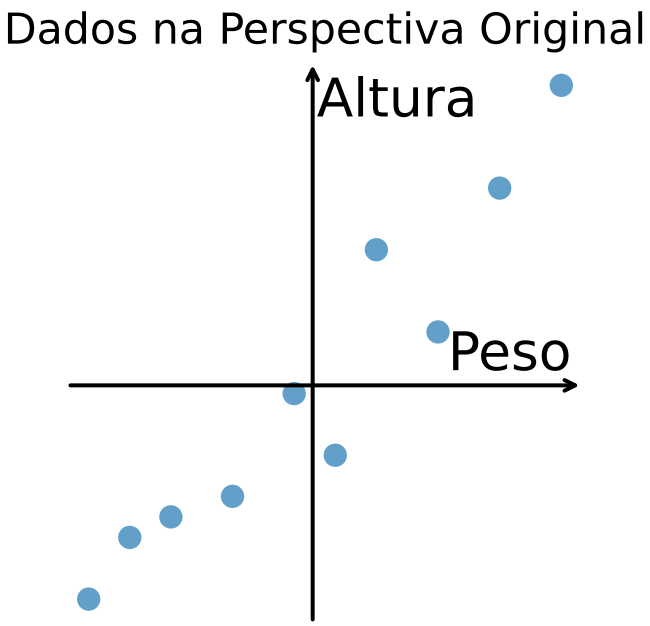

Figura 7.1: Diagrama de dispersão para os dados de Peso e Altura.

Observando o gráfico, notamos que a nuvem de pontos forma uma elipse inclinada, o que indica uma correlação entre Peso e Altura. Descrever os dados usando os eixos originais é perfeitamente válido, mas talvez não seja a forma mais eficiente. A maior parte da variabilidade ocorre ao longo de uma diagonal.

A ACP propõe uma rotação dos eixos para que eles se alinhem melhor com a estrutura dos dados. O resultado é um novo sistema de eixos, os **Componentes Principais** (CP_1 e CP_2), como mostrado abaixo.

O primeiro componente, CP_1, agora aponta na direção de maior "alongamento" da nuvem de pontos. O segundo, CP_2, é perpendicular ao primeiro e aponta na direção de maior variabilidade restante. Encontramos uma nova perspectiva que descreve a estrutura dos dados de forma mais natural e eficiente.

Em casos com mais de duas variáveis (p > 2), a lógica se estende: O **i-ésimo componente principal (**CP_i) aponta para a direção de maior variabilidade, sob a restrição de ser ortogonal (não correlacionado) a todos os componentes anteriores, $\text{Cov}[CP_j, CP_i] = 0 \; \forall \; j < i$.

Uma intuição geométrica para o problema é buscar o ângulo θ de rotação do eixo das variáveis tal que a variância dos componentes seja máxima. Essa rotação é simples de ser observada nesse exemplo bidimensional (veja Figura Figura 7.2).

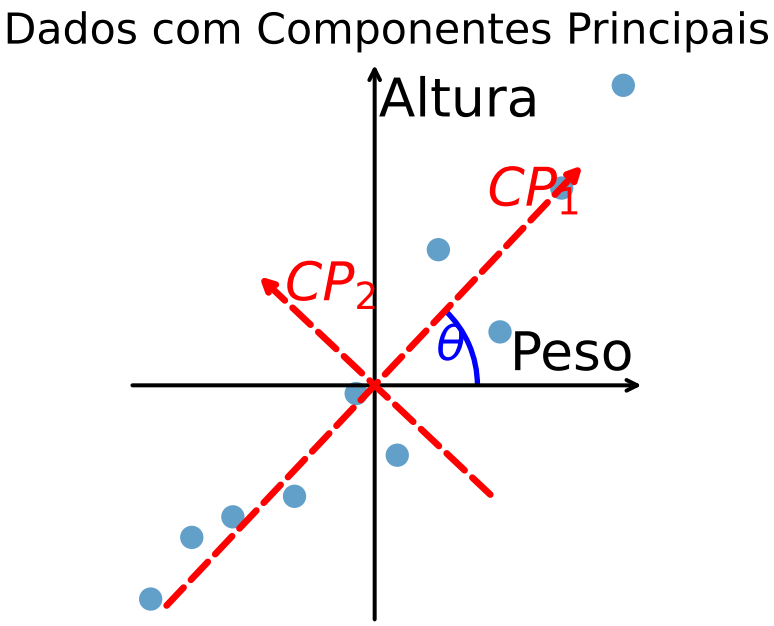

Figura 7.2: O sistema de eixos rotacionado (vermelho) se alinha com a máxima dispersão (variância) dos dados.

7.1 Variância como Medida de Informação

A essa altura, você deve estar se perguntando: por que a direção do "maior alongamento" é a mais importante? 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.

A ACP utiliza essa ideia para encontrar os eixos mais informativos. Ao rotacionar o sistema de coordenadas, ela não altera a variabilidade total dos dados, mas a redistribui de forma inteligente.

Definição 7.1. A **variância total** de um conjunto de dados com p variáveis é a soma das variâncias de cada variável individual. Matematicamente, se $\mathbf{x} = (X_1, \dots, X_p)'$ é o vetor de variáveis aleatórias com matriz de covariâncias $\boxtimes$, a variância total é definida como:

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

Onde σ_{jj} é a variância da j-ésima variável e $tr(\boxtimes)$ 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.

Exemplo 7.2. Voltando ao exemplo Exemplo 7.1, as variâncias das variáveis originais são:

- Variância do Peso: 59.88
- Variância da Altura: 66.71
- **Variância Total Original: 126.59**

Após a rotação, as variâncias ao longo dos novos eixos (os componentes principais) são:

- Variância de CP_1: **123.55**
- Variância de CP_2: **3.04**
- **Variância Total dos Componentes: 126.59**

Dois fatos cruciais se destacam:

1. **A variância total é conservada.** A soma das variâncias é a mesma nos dois sistemas de eixos. Nenhuma informação foi perdida; o ponto de vista foi apenas alterado.
2. **A variância foi eficientemente redistribuída.** O primeiro componente, CP_1, agora concentra **97.60%** da variância total. Isso significa que, se quiséssemos reduzir nossos dados de 2D para 1D, poderíamos manter apenas o CP_1 e ainda reter a maior parte da informação original. Essa é a essência da redução de dimensionalidade com ACP.

7.2 A Formalização Matemática

Com a intuição geométrica estabelecida, podemos formalizar a Análise de Componentes Principais. O objetivo é transformar um conjunto de variáveis correlacionadas $\mathbf{x} = (X_1, \ldots, X_p)'$ em um novo conjunto de variáveis não correlacionadas, os **componentes principais** $\mathbf{y} = (Y_1, \ldots, Y_p)'$. Cada componente é uma combinação linear das variáveis originais:

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p = \mathbf{e}_1'\mathbf{x} \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p = \mathbf{e}_2'\mathbf{x} \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p = \mathbf{e}_p'\mathbf{x} \end{aligned}$$

Em notação matricial, a transformação pode ser escrita de forma compacta:

$$\mathbf{Y} = \mathbf{E}'\mathbf{X} \tag{7.1}$$

Onde $\mathbf{Y}$ é o vetor $p \times 1$ de autovalores e $\mathbf{E}$ é a matriz $p \times p$ cujas colunas são os vetores de coeficientes $\mathbf{e}_k$.

Esses componentes são construídos para satisfazer duas condições fundamentais:

1. **Variâncias Ordenadas**: A variância do primeiro componente é a maior possível, a do segundo é a maior possível entre as direções não correlacionadas com o primeiro, e assim por diante. Ou seja, $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \geq \text{Var}(Y_p)$.
2. **Não Correlacionados**: Os componentes são ortogonais entre si, o que significa que $\text{Cov}[Y_i, Y_k] = 0$ para todo $i \neq k$.

7.2.1 O Problema de Maximização

O primeiro componente principal, $Y_1 = \mathbf{e}_1'\mathbf{x}$, é a combinação linear com variância máxima. A variância de Y_1 é dada por:

$$\text{Var}(Y_1) = \text{Var}(\mathbf{e}_1'\mathbf{x}) = \mathbf{e}_1'\text{Var}(\mathbf{x})\mathbf{e}_1 = \mathbf{e}_1'\boxtimes\mathbf{e}_1$$

Onde $\boxtimes$ é a matriz de covariâncias de $\mathbf{x}$. Para evitar que a variância seja aumentada simplesmente inflando os coeficientes em $\mathbf{e}_1$, impomos a restrição de que seu comprimento seja unitário, $\mathbf{e}_1'\mathbf{e}_1 = 1$. Formalmente, o problema de maximização para o primeiro componente principal se torna:

$$\begin{aligned} \max_{\mathbf{e}_1} \quad & \mathbf{e}_1'\boxtimes\mathbf{e}_1 \\ \text{sujeito a} \quad & \mathbf{e}_1'\mathbf{e}_1 = 1 \end{aligned}$$

Para maximizar a variância sujeita à restrição, utilizamos o método dos multiplicadores de Lagrange. A função a ser maximizada é:

$$L(\mathbf{e}_1, \lambda_1) = \mathbf{e}_1'\boxtimes\mathbf{e}_1 - \lambda_1(\mathbf{e}_1'\mathbf{e}_1 - 1)$$

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

$$\frac{\partial L}{\partial \mathbf{e}_1} = 2\boxtimes\mathbf{e}_1 - 2\lambda_1\mathbf{e}_1 = 0$$

O que nos leva à equação fundamental de autovalores e autovetores:

$$\boxtimes\mathbf{e}_1 = \lambda_1\mathbf{e}_1$$

Esta equação mostra que o vetor de coeficientes $\mathbf{e}_1$ deve ser um autovetor da matriz de covariâncias $\boxtimes$. Para encontrar a variância, pré-multiplicamos a equação por $\mathbf{e}_1'$:

$$\mathbf{e}_1'\boxtimes\mathbf{e}_1 = \lambda_1\mathbf{e}_1'\mathbf{e}_1$$

Como $\text{Var}(Y_1) = \mathbf{e}_1' \boxtimes \mathbf{e}_1$ e a restrição é $\mathbf{e}_1' \mathbf{e}_1 = 1$, 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 $\boxtimes$, e $\mathbf{e}_1$ é o **autovetor** correspondente.

> **Nota**
>
> A matriz de covariâncias $\boxtimes$ é, por construção, uma matriz simétrica e positiva semi-definida. Conforme discutido em Seção 5.3, o **Teorema Espectral** garante que os autovalores de tal matriz são reais e não-negativos, e que seus autovetores correspondentes a autovalores distintos são ortogonais. Esta propriedade é fundamental para a existência e unicidade dos componentes principais.

> **Nota**
>
> A demonstração acima, utilizando multiplicadores de Lagrange, é uma maneira moderna e elegante de conduzir a derivação do problema de máximização. Uma abordagem clássica restringe a norma de e através do quociente,
>
> $$\text{Var}(Y_1) = \max_{e_1} \frac{\mathbf{e}_1' \Sigma \mathbf{e}_1}{\mathbf{e}_1' \mathbf{e_1}}$$
>
> Este é um problema clássico na álgebra linear. Um teorema fundamental afirma que para qualquer matriz simétrica A, o máximo da forma quadrática $\mathbf{x}' A \mathbf{x}$, sujeito à restrição $\mathbf{x}'\mathbf{x} = 1$, é o **maior autovalor** de A. O vetor $\mathbf{x}$ que atinge esse máximo é o autovetor correspondente. Como a matriz de covariâncias $\boxtimes$ é simétrica, este teorema se aplica diretamente ao nosso problema.

7.2.2 Componentes Subsequentes

Uma vez encontrada a primeira direção de máxima variância, o segundo componente principal, $Y_2 = \mathbf{e}_2' \mathbf{x}$, busca capturar o máximo da variabilidade *restante*, sob a condição de ser não correlacionado com Y_1. A condição de componentes não correlacionados garante que a informação presente no segundo componente principal não é redundante com relação aquela já presente no primeiro. Formalmente, temos:

$$Cov(Y_1, Y_2) = Cov(\mathbf{e}_1' \mathbf{x}, \mathbf{e}_2' \mathbf{x}) = \mathbf{e}_1' \boxtimes \mathbf{e}_2$$

Como $\mathbf{e}_1$ é o primeiro autovetor, temos $\boxtimes \mathbf{e}_1 = \lambda_1 \mathbf{e}_1$, Assim:

$$\mathbf{e}_1'\boxtimes\mathbf{e}_2 = (\lambda_1\mathbf{e}_1)'\mathbf{e}_2 = \lambda_1\mathbf{e}_1'\mathbf{e}_2$$

Logo:

$$Cov(Y_1, Y_2) = 0 \iff \mathbf{e}_1'\mathbf{e}_2 = 0$$

Com essa condição bem definida, o problema para o segundo componente se torna:

$$\max_{\mathbf{e}_2} \quad \mathbf{e}_2'\boxtimes\mathbf{e}_2$$
$$\text{sujeito a} \quad \begin{cases}\mathbf{e}_2'\mathbf{e}_2 = 1\\ \mathbf{e}_1'\mathbf{e}_2 = 0\end{cases}$$

A função Lagrangiana agora inclui dois multiplicadores, λ_2 e ϕ:

$$L(\mathbf{e}_2, \lambda_2, \phi) = \mathbf{e}_2'\boxtimes\mathbf{e}_2 - \lambda_2(\mathbf{e}_2'\mathbf{e}_2 - 1) - \phi(\mathbf{e}_1'\mathbf{e}_2 - 0)$$

Derivando em relação a $\mathbf{e}_2$ e igualando a zero, temos:

$$\frac{\partial L}{\partial \mathbf{e}_2} = 2\boxtimes\mathbf{e}_2 - 2\lambda_2\mathbf{e}_2 - \phi\mathbf{e}_1 = \mathbf{0}$$

Pré-multiplicando por $\mathbf{e}_1'$:

$$2\mathbf{e}_1'\boxtimes\mathbf{e}_2 - 2\lambda_2\mathbf{e}_1'\mathbf{e}_2 - \phi\mathbf{e}_1'\mathbf{e}_1 = 0$$

Sabendo que:

- $\mathbf{e}_1'\boxtimes = \lambda_1\mathbf{e}_1'$
- $\mathbf{e}_1'\mathbf{e}_2 = 0$
- $\mathbf{e}_1'\mathbf{e}_1 = 1$

A equação se simplifica a $\phi = 0$. Substituindo $\phi = 0$ de volta na derivada, a equação se torna:

$$\boxtimes\mathbf{e}_2 = \lambda_2\mathbf{e}_2$$

Assim, $\mathbf{e}_2$ é o autovetor de $\boxtimes$ correspondente ao autovalor λ_2. Como λ_1 foi o maior autovalor, para maximizar a variância de Y_2, λ_2 deve ser o **segundo maior autovalor**. Este processo se generaliza para os componentes subsequentes.

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

É neste ponto que a conexão com a Seção 5.3 se torna explícita. A matriz $\mathbf{P}$ (Equação 7.1), cujas colunas são os autovetores da matriz de covariâncias $\boxtimes$, é exatamente a mesma matriz $\mathbf{P}$ da decomposição espectral $\boxtimes = \mathbf{P}\Lambda\mathbf{P}'$. Além disso, Λ é uma matriz diagonal contendo a variância de cada componente principal. Logo, podemos obter todos os componentes principais de maneira prática e simultânea através da decomposição espectral.

7.3 A Importância do Pré-processamento dos Dados

A Análise de Componentes Principais é, em sua essência, uma análise de variabilidade. A forma como medimos essa variabilidade impacta diretamente o resultado. Dois pré-processamentos são cruciais: a **centralização** e o **escalonamento**.

7.3.1 Centralização

Na Análise de Componentes Principais (ACP), a centralização dos dados — ou seja, a subtração da média de cada variável — é uma etapa fundamental não apenas para o cálculo da matriz de covariâncias, mas também para a projeção dos dados nos componentes principais.

Ao projetar os dados em um componente $\mathbf{e}_i$, é imprescindível que a projeção seja feita a partir dos dados centralizados, ou seja:

$$Y_i = \mathbf{e}_i^T(\mathbf{x} - \bar{\mathbf{x}})$$

Esse detalhe é essencial porque os autovetores da ACP são obtidos com base na matriz de covariâncias, a qual descreve a dispersão dos dados em torno da média, e não em torno da origem. Se aplicarmos a projeção diretamente sobre $\mathbf{x}$, sem subtrair a média, os componentes resultantes não representarão adequadamente as direções de maior variabilidade — e sim uma combinação da dispersão com a posição média dos dados.

Portanto, para que os componentes principais preservem a interpretação correta como combinações lineares que explicam a variância dos dados em torno do centro da nuvem de pontos, é indispensável que tanto o cálculo da matriz de covariâncias quanto a projeção dos dados utilizem os dados centralizados.

> **!** Importante
>
> No contexto de ACP, é comum e prático denotar por $\mathbf{x}$ o vetor de variáveis já centralizado. Utilizamos esse abuso de notação durante esse capítulo para simplificação do texto sem perda de generalidade.

7.3.2 Por que Escalonar? O Dilema da Covariância vs. Correlação

A Análise de Componentes Principais (ACP) é sensível à **escala** das variáveis. Se uma variável tiver uma variância numericamente muito maior que as outras — mesmo que apenas por causa da sua unidade de medida — ela poderá dominar os primeiros componentes principais.

Imagine incluir uma terceira variável no conjunto Altura/Peso: a **renda mensal**, medida em Reais. As variâncias poderiam ser aproximadamente:

- **Altura:** 80 cm^2
- **Peso:** 60 kg^2
- **Renda:** 4.000.000 $(RS)^2$

Nesse cenário, a variância da Renda é **milhares de vezes maior** que a das outras variáveis. Se aplicarmos a ACP diretamente na **matriz de covariâncias**, o primeiro componente principal será fortemente direcionado pela Renda, mesmo que sua correlação com as demais variáveis seja baixa. Isso ocorre porque a ACP estará apenas "seguindo" a direção da variável com maior variância — não necessariamente a mais informativa.

Para evitar esse viés, escalonamos as variáveis: cada uma é dividida por seu desvio padrão. Isso padroniza todas para variância igual a 1. Ao fazer isso, estamos na prática realizando a ACP sobre a **matriz de correlação** ($\mathbf{R}$) em vez da matriz de covariâncias ($\boxtimes$).

Vantagem do escalonamento:
Usar a matriz de correlação "democratiza" a análise. Todas as variáveis começam com a **mesma importância inicial** (variância 1), e a ACP passa a capturar a **estrutura de correlações**, ao invés de ser enviesada pelas **diferenças de escala**.

Quando usar a matriz de covariâncias?
Somente quando todas as variáveis estão na **mesma unidade de medida** e possuem uma interpretação comparável. Por exemplo, comparar a temperatura em Celsius em diferentes regiões pode fazer sentido sem escalonamento. Fora isso, a matriz de **correlação** é geralmente a escolha mais robusta e segura.

7.4 Componentes Principais Populacionais vs. Amostrais

Até este ponto, discutimos os componentes principais em um contexto populacional, onde a matriz de covariâncias ⊠ (ou correlação **R**) e seus autovalores λ_k e autovetores $\mathbf{e}_k$ são conhecidos. Na prática, quase sempre trabalhamos com uma amostra de dados. Nesse caso, não conhecemos os verdadeiros parâmetros populacionais e devemos estimá-los.

Os **componentes principais amostrais** são obtidos da mesma maneira, mas usando a matriz de covariâncias amostral **S** (ou a matriz de correlação amostral **R**). As quantidades resultantes são estimativas dos seus análogos populacionais:

- O k-ésimo autovalor amostral, $\hat{\lambda}_k$, é uma estimativa de λ_k.
- O k-ésimo autovetor amostral, $\hat{\mathbf{e}}_k$, é uma estimativa de $\mathbf{e}_k$.
- O k-ésimo componente principal amostral, $\hat{Y}_k = \hat{\mathbf{e}}_k' \mathbf{x}$, 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 (ˆ), mas é importante lembrar que, na aplicação prática, estamos sempre lidando com estimativas amostrais.

7.5 Escolhendo o Número de Componentes

A principal vantagem da ACP é a redução de dimensionalidade. Mas como decidimos quantos componentes ($q < p$) reter? A escolha de q envolve um trade-off entre a simplicidade (poucos componentes) e a fidelidade aos dados originais (muitos componentes). Não existe uma regra única, mas sim um conjunto de critérios que devem ser avaliados em conjunto.

7.5.1 Critério da Variância Explicada Acumulada

Este é o critério mais comum. 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.

7.5.2 Critério do Autovalor (Critério de Kaiser)

Proposto por Henry Kaiser, este critério 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.

7.5.3 Scree Plot (Gráfico de Cotovelo)

O Scree Plot, proposto por Raymond Cattell, é uma ferramenta visual que nos ajuda a identificar o número ideal de componentes. Ele é um gráfico de linha dos autovalores (variâncias dos componentes) em ordem decrescente.

Tipicamente, o gráfico mostra uma queda acentuada nos primeiros autovalores, seguida por um nivelamento gradual para os autovalores restantes. O ponto onde a curva "dobra" ou forma um "cotovelo" (*elbow*) é considerado 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. Os componentes após o cotovelo formam o "cascalho" (*scree*) na base de uma montanha e são considerados "ruído".

7.6 Interpretando os Componentes Principais

Uma vez que selecionamos o número de componentes a reter, o passo final é a **interpretação**. O que esses novos eixos, que são combinações de nossas variáveis originais, realmente significam?

Os coeficientes e_{kj} do autovetor $\mathbf{e}_k$ são chamados de **cargas** (*loadings*) e representam o peso da variável original X_j na formação do componente Y_k. Embora as cargas sejam importantes, a sua interpretação pode ser complicada, pois sua magnitude depende das unidades das variáveis originais.

Uma medida mais interpretável é a **correlação entre os componentes principais e as variáveis originais**, $Cor(Y_k, X_j)$. Ela nos diz o quão "alinhado" um componente está com cada variável original, numa escala padronizada de -1 a 1. A fórmula para essa correlação é:

$$Cor(Y_k, X_j) = \frac{e_{kj}\sqrt{\lambda_k}}{\sqrt{s_{jj}}}$$

Onde:

- e_{kj} é a carga da variável j no componente k.
- λ_k é o autovalor (variância) do componente k.

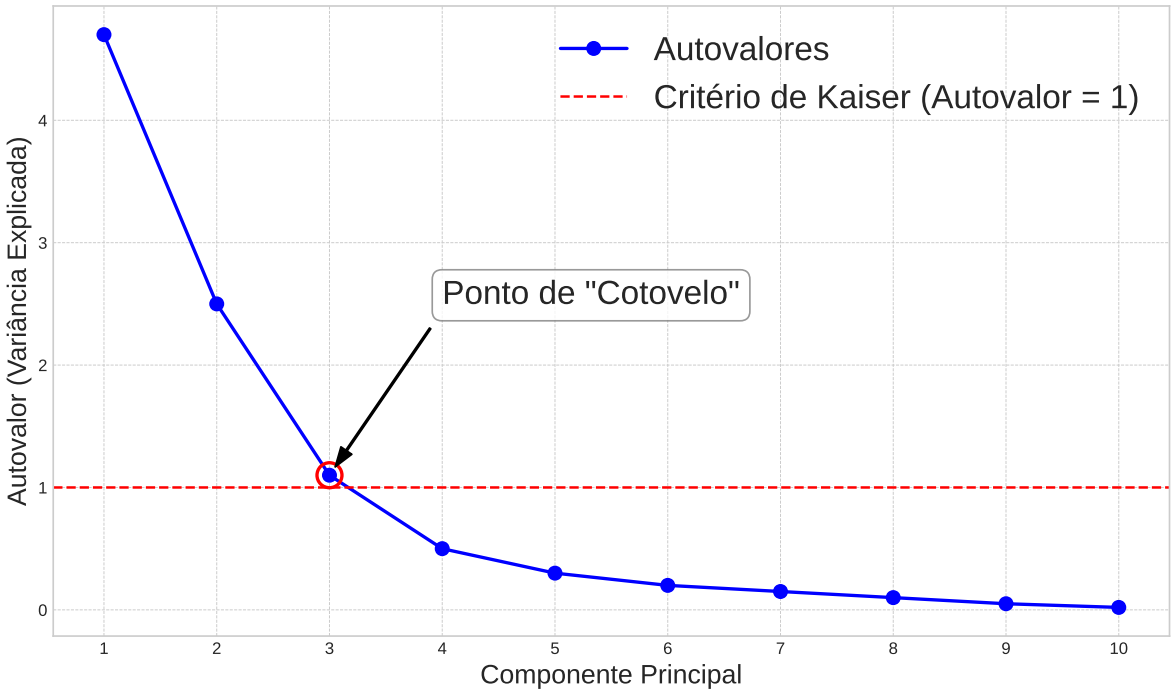

Figura 7.3: Exemplo de um Scree Plot. O 'cotovelo' em k=3 sugere a retenção de 3 componentes.

- s_{jj} é a variância da variável original j.

Quando a ACP é realizada sobre a **matriz de correlação** (ou seja, com dados padronizados), as variâncias s_{jj} são todas iguais a 1. Nesse caso, a fórmula simplifica para $Cor(Y_k, X_j) = e_{kj}\sqrt{\lambda_k}$. As correlações se tornam proporcionais às cargas, facilitando a interpretação.

Além disso, quando a ACP é realizada sobre a **matriz de correlações**, as variáveis são padronizadas. Nesse caso, uma opção comum e direta é avaliar os próprios *loadings* (os autovetores da matriz de correlação) para entender a contribuição de cada variável. Um *loading* alto (próximo de 1 ou -1) indica que a variável tem uma forte influência na construção daquele componente.

A etapa mais crucial da ACP é transformar os eixos matemáticos (os componentes) em descobertas práticos. A ferramenta visual mais adequada para essa tarefa é o **biplot**. O termo "biplot" significa "dois plots" (plot duplo), pois ele sobrepõe duas informações em um único gráfico:

1. **Os scores:** As coordenadas das observações no novo espaço dos componentes principais.
2. **Os loadings:** As contribuições das variáveis originais para a criação desses componentes.

O resultado é um mapa rico que mostra não apenas como as observações se agrupam, mas *por que* elas se agrupam daquela maneira. A interpretação de um biplot segue uma lógica visual. Vamos quebrar em partes:

1. **Eixos (Componentes Principais)**: O eixo horizontal é o CP1 e o vertical é o CP2. Eles são as "réguas" do nosso novo mapa e representam as direções de maior variabilidade nos dados. A porcentagem de variância que cada um explica é mostrada nos seus rótulos.
2. **Pontos (Observações)**: Cada ponto no gráfico é uma observação.
    - **Proximidade**: Pontos próximos uns dos outros representam observações com perfis semelhantes (conforme capturado pelos dois primeiros CPs).
    - **Agrupamentos**: Grupos de pontos (clusters) indicam subpopulações nos dados.
3. **Vetores (Variáveis Originais)**: Cada seta (vetor) representa uma das variáveis originais.
    - **Direção**: A direção da seta indica como a variável contribui para os dois componentes. Uma seta que aponta para a direita indica uma forte contribuição positiva para o CP1. Uma que aponta para cima, uma forte contribuição positiva para o CP2.
    - **Comprimento**: O comprimento da seta é proporcional a quão bem a variável é representada no espaço 2D do biplot. Setas mais longas significam que a variável tem uma forte influência nos componentes mostrados e é bem representada no gráfico. Setas curtas são menos importantes para os dois primeiros CPs ou sua variabilidade está melhor explicada em outros componentes (CP3, CP4, etc.).
    - **Relações entre Variáveis**: O ângulo entre os vetores nos informa sobre a correlação entre as variáveis originais.
        - **Ângulo pequeno (< 90°)**: As variáveis são positivamente correlacionadas.
        - **Ângulo de ~90°**: As variáveis não são correlacionadas.
        - **Ângulo obtuso (> 90°)**: As variáveis são negativamente correlacionadas.
4. **Relação entre Pontos e Vetores**: Para entender o perfil de um ponto (ou grupo de pontos), projete-o ortogonalmente sobre os vetores das variáveis. Se a projeção de um ponto cai na direção de um vetor, aquela observação tem um valor alto para aquela variável. Se cai na direção oposta, tem um valor baixo.

Com essas regras em mente, vamos analisar um biplot genérico.

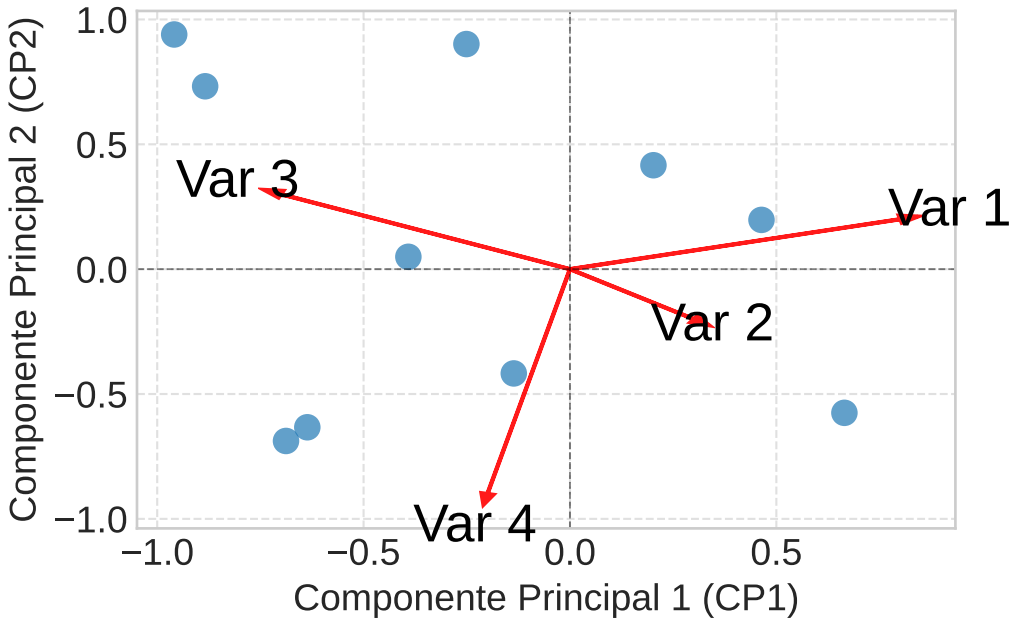

Figura 7.4: Exemplo de um biplot genérico para ilustrar a interpretação dos seus elementos. Os pontos representam as observações e as setas, as variáveis originais.

8 Análise Fatorial

A Análise Fatorial é uma técnica estatística utilizada para descrever a estrutura de covariância entre um conjunto de variáveis observadas. A hipótese central é que essa estrutura é gerada por um número menor de variáveis latentes não observáveis, denominadas **fatores comuns**.

Começamos com uma intuição. Suponha que temos as seguintes variáveis de gastos para diferentes famílias:

- X_1: Gasto em educação
- X_2: Gasto em cultura
- X_3: Gasto em alimentação

É razoável supor que essas variáveis sejam correlacionadas. Mais do que isso, pode existir um fator latente, como a **renda familiar** (F_1), que influencia todos esses gastos. A Análise Fatorial busca formalizar e quantificar essa relação.

8.1 O Modelo Fatorial Ortogonal

O modelo supõe que cada variável observada é linearmente dependente de um conjunto de fatores comuns, somado a um termo de variância individual, ou específico.

Definição 8.1. Seja $\mathbf{x}$ um vetor aleatório de p variáveis observadas com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boxtimes$. O modelo fatorial com m fatores comuns ($m < p$) postula que $\mathbf{x}$ é linearmente dependente de m fatores comuns $F_1, F_2, \dots, F_m$ e p termos de erro $\epsilon_1, \epsilon_2, \dots, \epsilon_p$. Em notação matricial, o modelo é:

$$\mathbf{x}_{(p\times 1)} - \boldsymbol{\mu}_{(p\times 1)} = \mathbf{L}_{(p\times m)}\mathbf{F}_{(m\times 1)} + \boldsymbol{\epsilon}_{(p\times 1)} \tag{8.1}$$

Onde:

- $\mathbf{L}$ é a matriz de **cargas fatoriais**: Uma matriz de pesos responsável por quantificar as relações entre as p variáveis e os m fatores.
- $\mathbf{F}$ é o vetor de **fatores comuns**, ou seja $\mathbf{F} = [F_1, F_2, \dots, F_m]'$.
- $\boldsymbol{\epsilon}$ é o vetor de **erros**, ou **variâncias específicas**, ou seja, $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \dots, \epsilon_p]'$

Para que o ajuste desse modelo seja factível, as seguintes suposições são feitas para o modelo ortogonal:

1. $E[\mathbf{F}] = \mathbf{0}$ e $Cov(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \mathbf{I}$.
2. $E[\boldsymbol{\epsilon}] = \mathbf{0}$ e $Cov(\boldsymbol{\epsilon}) = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \boldsymbol{\Psi}$, onde $\boldsymbol{\Psi}$ é uma matriz diagonal.
3. $Cov(\mathbf{F}, \boldsymbol{\epsilon}) = E[\mathbf{F}\boldsymbol{\epsilon}'] = \mathbf{0}$.

8.2 A Estrutura de Covariância Implícita

As suposições do modelo implicam uma estrutura específica para a matriz de covariâncias $\boldsymbol{\Sigma}$.

Teorema 8.1. *Sob as premissas do modelo fatorial ortogonal (Definição 8.1), a matriz de covariâncias $\boldsymbol{\Sigma}$ do vetor $\mathbf{x}$ é dada por:*

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi} \tag{8.2}$$

Comprovação. A partir do modelo fatorial em Equação 8.1, temos que $\mathbf{x} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}$. A matriz de covariâncias de $\mathbf{x}$ é, por definição, $\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']$. Substituindo a expressão do modelo, obtemos:

$$\begin{aligned}
\boldsymbol{\Sigma} &= E[(\mathbf{L}\mathbf{F} + \boldsymbol{\epsilon})(\mathbf{L}\mathbf{F} + \boldsymbol{\epsilon})'] \\
&= E[(\mathbf{L}\mathbf{F} + \boldsymbol{\epsilon})(\mathbf{F}'\mathbf{L}' + \boldsymbol{\epsilon}')] \\
&= E[\mathbf{L}\mathbf{F}\mathbf{F}'\mathbf{L}' + \mathbf{L}\mathbf{F}\boldsymbol{\epsilon}' + \boldsymbol{\epsilon}\mathbf{F}'\mathbf{L}' + \boldsymbol{\epsilon}\boldsymbol{\epsilon}'] \\
&= \mathbf{L}E[\mathbf{F}\mathbf{F}']\mathbf{L}' + \mathbf{L}E[\mathbf{F}\boldsymbol{\epsilon}'] + E[\boldsymbol{\epsilon}\mathbf{F}']\mathbf{L}' + E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']
\end{aligned}$$

Pelas suposições do modelo ortogonal (Definição 8.1):

1. $E[\mathbf{F}\mathbf{F}'] = \text{Cov}(\mathbf{F}) = \mathbf{I}$ (os fatores são não correlacionados e têm variância unitária).
2. $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$ (os erros são não correlacionados entre si).
3. $E[\mathbf{F}\boldsymbol{\epsilon}'] = \text{Cov}(\mathbf{F}, \boldsymbol{\epsilon}) = \mathbf{0}$ (os fatores e os erros são não correlacionados).

Substituindo essas esperanças na equação de $\boldsymbol{\Sigma}$, temos:

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{I}\mathbf{L}' + \mathbf{L}\mathbf{0} + \mathbf{0}\mathbf{L}' + \boldsymbol{\Psi} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$$

Isso completa a prova. □

Esta equação decompõe a variância de cada variável X_i em:

- **Comunalidade (h_i^2)**: A porção da variância de X_i explicada pelos m fatores comuns ($h_i^2 = \sum_{j=1}^{m} l_{ij}^2$).
- **Variância Específica (ψ_i)**: A porção da variância de X_i não explicada pelos fatores comuns ($Var(X_i) = \sigma_{ii} = h_i^2 + \psi_i$).

Exemplo 8.1. Suponha que a matriz de covariâncias de um vetor aleatório $\mathbf{x}$ com $p = 4$ variáveis seja:

$$\boxtimes = \begin{pmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 37 & 47 \\ 12 & 23 & 47 & 68 \end{pmatrix}$$

É possível mostrar que um modelo fatorial com $m = 2$ fatores comuns pode gerar essa estrutura de covariância. Uma solução possível para $\mathbf{L}$ e $\boxtimes$ é dada por:

$$\mathbf{L} = \begin{pmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{pmatrix}, \quad \boxtimes = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

O leitor pode verificar que $\boxtimes = \mathbf{LL}' + \boxtimes$. O modelo decompõe a variância de cada variável.

- Para X_1, a comunalidade é $h_1^2 = 4^2 + 1^2 = 17$, e sua variância total é $Var(X_1) = \sigma_{11} = h_1^2 + \psi_1 = 17 + 2 = 19$.
- Para X_2, a comunalidade é $h_2^2 = 7^2 + 2^2 = 53$, e sua variância total é $Var(X_2) = \sigma_{22} = h_2^2 + \psi_2 = 53 + 4 = 57$.

8.3 Problemas no Modelo Fatorial

- **Existência da Solução**: Nem sempre existe uma solução factível para o modelo fatorial com m fatores. A estimação dos parâmetros, especialmente com um número inadequado de fatores, pode levar a soluções impróprias, como uma variância específica negativa ($\hat{\psi}_i < 0$), conhecida como **caso de Heywood**. Isso viola a premissa de que ψ_i é uma variância e, portanto, deve ser não-negativa. Geralmente, uma solução imprópria indica que o modelo é inadequado para os dados.

Exemplo 8.2. Considere um modelo de um fator ($m = 1$) para $p = 3$ variáveis, com a seguinte matriz de correlação populacional:

$$\mathbf{R} = \begin{pmatrix} 1.0 & 0.4 & 0.9 \\ 0.4 & 1.0 & 0.7 \\ 0.9 & 0.7 & 1.0 \end{pmatrix}$$

O modelo fatorial para a matriz de correlação é $\mathbf{P} = \mathbf{LL'} + \boxtimes$. Para $m = 1$, as cargas são um vetor $\mathbf{L} = [l_{11}, l_{21}, l_{31}]'$. As covariâncias (correlações) são dadas por $\rho_{ij} = l_{i1}l_{j1}$. Temos o sistema:

- $\rho_{12} = l_{11}l_{21} = 0.4$
- $\rho_{13} = l_{11}l_{31} = 0.9$
- $\rho_{23} = l_{21}l_{31} = 0.7$

Multiplicando as três equações, obtemos $(l_{11}l_{21}l_{31})^2 = 0.4 \times 0.9 \times 0.7 = 0.252$. Isso nos permite resolver para as cargas:

- $l_{11}^2 = (l_{11}l_{21})(l_{11}l_{31})/(l_{21}l_{31}) = (0.4 \times 0.9)/0.7 \approx 0.514$
- $l_{21}^2 = (l_{11}l_{21})(l_{21}l_{31})/(l_{11}l_{31}) = (0.4 \times 0.7)/0.9 \approx 0.311$
- $l_{31}^2 = (l_{11}l_{31})(l_{21}l_{31})/(l_{11}l_{21}) = (0.9 \times 0.7)/0.4 = 1.575$

A comunalidade da terceira variável é $h_3^2 = l_{31}^2 = 1.575$. Como estamos modelando uma matriz de correlação, a variância total de cada variável é 1. A variância específica seria $\psi_3 = 1 - h_3^2 = 1 - 1.575 = -0.575$. Uma variância negativa é impossível, indicando que o modelo de um fator não é apropriado para descrever a estrutura de correlação dada.

- **Indeterminação da Solução (Rotação Fatorial)**: A solução para a matriz de cargas $\mathbf{L}$ não é única. Para qualquer matriz ortogonal $\mathbf{T}$ de dimensão $m \times m$ (ou seja, uma matriz tal que $\mathbf{TT'} = \mathbf{T'T} = \mathbf{I}$), podemos definir uma nova matriz de cargas $\mathbf{L}^* = \mathbf{LT}$ que resulta na mesma matriz de covariâncias.

Isso ocorre porque a parte da covariância explicada pelos fatores, $\mathbf{LL'}$, permanece inalterada:

$$\mathbf{L}^*(\mathbf{L}^*)' = (\mathbf{LT})(\mathbf{LT})' = \mathbf{LTT'L'} = \mathbf{L}(\mathbf{TT'})\mathbf{L'} = \mathbf{LIL'} = \mathbf{LL'}$$

Portanto, o modelo $\boxtimes = \mathbf{L}^*(\mathbf{L}^*)' + \boxtimes$ é equivalente ao modelo original. Essa propriedade é a base para a **rotação fatorial**, um procedimento que busca a solução $\mathbf{L}^*$ mais simples e interpretável, sem alterar o ajuste do modelo.

8.4 Adequabilidade do Modelo Fatorial

Antes de aplicar os métodos de estimação, pode-se avaliar se os dados são adequados para a Análise Fatorial. A lema fundamental da AF é que as variáveis observadas são correlacionadas e que essa correlação pode ser explicada por fatores latentes. Se as variáveis são ortogonais ou se a correlação entre elas é espúria, o modelo fatorial não é apropriado.

Dois dos principais diagnósticos para verificar a adequabilidade dos dados são o Teste de Esfericidade de Bartlett e a medida de adequação da amostra de Kaiser-Meyer-Olkin (KMO).

8.4.1 Teste de Esfericidade de Bartlett

O Teste de Esfericidade de Bartlett avalia a hipótese nula (H_0) de que a matriz de correlação populacional $\mathbf{P}$ é uma matriz identidade ($H_0 : \mathbf{P} = \mathbf{I}$). Se essa hipótese for verdadeira, as variáveis são não correlacionadas, e não há estrutura latente para ser extraída.

A estatística de teste é baseada no determinante da matriz de correlação amostral $\mathbf{R}$ e, sob H_0, segue aproximadamente uma distribuição Qui-quadrado. Para uma amostra de tamanho n e p variáveis, a estatística é:

$$\chi^2 = -\left[(n-1) - \frac{2p+5}{6}\right] \ln(|\mathbf{R}|)$$

Esta estatística tem, aproximadamente, uma distribuição χ^2 com $p(p-1)/2$ graus de liberdade. Um p-valor baixo (e.g., < 0.05) leva à rejeição de H_0, indicando que existe correlação suficiente entre as variáveis para justificar a aplicação da Análise Fatorial.

8.4.2 Medida de Adequação da Amostra (KMO)

Enquanto o teste de Bartlett avalia se a matriz de correlação como um todo se desvia significativamente da identidade, a medida de Kaiser-Meyer-Olkin (KMO) quantifica o quão adequados os dados são para a fatorização. O KMO compara a magnitude dos coeficientes de correlação observados com a magnitude dos coeficientes de correlação parcial.

A lógica é que, se as variáveis compartilham fatores comuns, as correlações parciais entre pares de variáveis (controlando pelas outras variáveis) devem ser pequenas. A estatística KMO é calculada como:

$$\text{KMO} = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

Onde r_{ij} é o coeficiente de correlação simples entre as variáveis X_i e X_j, e a_{ij} é o coeficiente de correlação parcial.

O valor do KMO varia de 0 a 1. Valores mais altos indicam que a Análise Fatorial é mais apropriada. Uma regra prática para a interpretação do KMO é:

- > **0.9**: Maravilhoso
- **0.8 - 0.9**: Meritório
- **0.7 - 0.8**: Razoável
- **0.6 - 0.7**: Medíocre
- **0.5 - 0.6**: Ruim
- < **0.5**: Inaceitável

Valores abaixo de 0.5 sugerem que a Análise Fatorial pode não ser uma boa ideia.

8.5 Métodos de Estimação

Assumindo uma amostra aleatória $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de uma população com matriz de covariâncias $\boxtimes$, o desafio é estimar $\mathbf{L}$ e $\boxtimes$ usando a matriz de covariâncias amostral $\mathbf{S}$ ou a matriz de correlação amostral $\mathbf{R}$.

Existem diversos métodos para estimar os parâmetros do modelo fatorial, cada um com suas próprias premissas e propriedades. Alguns dos mais conhecidos incluem:

- Método de Componentes Principais (MCP)
- Método da Máxima Verossimilhança (MMV)
- Método dos Fatores Principais (*Principal Axis Factoring*)
- Mínimos Quadrados Ponderados
- Mínimos Quadrados Generalizados

Neste capítulo, focaremos nos dois métodos mais amplamente utilizados na prática: o Método de Componentes Principais, por sua simplicidade computacional, e o Método da Máxima Verossimilhança, por sua fundamentação estatística robusta.

8.5.1 A Solução por Componentes Principais

O método de componentes principais (MCP) provê uma solução para $\mathbf{L}$ e $\boxtimes$ a partir da decomposição espectral da matriz de covariâncias amostral $\mathbf{S}$ (ou da matriz de correlações $\mathbf{R}$).

A ideia é que a matriz $\mathbf{S}$ pode ser decomposta em termos de seus pares de autovalor-autovetor $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$:

$$\mathbf{S} = \hat{\lambda}_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1' + \hat{\lambda}_2 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_2' + \cdots + \hat{\lambda}_p \hat{\mathbf{e}}_p \hat{\mathbf{e}}_p'$$

A estrutura do modelo fatorial é $\mathbf{S} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$. O MCP busca uma aproximação para $\mathbf{S}$ retendo apenas os m primeiros componentes, que explicam a maior parte da variabilidade total. A matriz $\hat{\mathbf{L}}\hat{\mathbf{L}}'$ é construída para igualar a contribuição desses componentes:

$$\hat{\mathbf{L}}\hat{\mathbf{L}}' = \hat{\lambda}_1\hat{\mathbf{e}}_1\hat{\mathbf{e}}_1' + \hat{\lambda}_2\hat{\mathbf{e}}_2\hat{\mathbf{e}}_2' + \cdots + \hat{\lambda}_m\hat{\mathbf{e}}_m\hat{\mathbf{e}}_m'$$

Uma solução explícita para $\hat{\mathbf{L}}$ que satisfaz essa equação é uma matriz $p \times m$ cujas colunas são os autovetores reescalados pelos respectivos autovalores. A matriz $\hat{\Psi}$ é então definida para garantir que as variâncias do modelo ($\text{diag}(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi})$) sejam iguais às variâncias amostrais ($\text{diag}(\mathbf{S})$).

Isso nos leva à seguinte definição formal.

Definição 8.2. Seja $\mathbf{S}$ a matriz de covariância amostral com pares de autovalor-autovetor $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$. A solução de componentes principais com m fatores é definida por:

- **Matriz de Cargas Estimada ($\hat{\mathbf{L}}$)**:

$$\hat{\mathbf{L}} = [\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 | \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 | \dots | \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m]$$

- **Matriz de Variâncias Específicas Estimada ($\hat{\Psi}$)**:

$$\hat{\Psi} = \text{diag}(\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}')$$

onde $\hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{l}_{ij}^2$.

Se a matriz de correlações $\mathbf{R}$ for utilizada, as cargas $\hat{\mathbf{L}}$ são calculadas a partir dos autovalores e autovetores de $\mathbf{R}$, e as variâncias específicas são $\hat{\psi}_i = 1 - \sum_{j=1}^{m} \hat{l}_{ij}^2$.

Por construção, este método força a diagonal da matriz de covariâncias do modelo, $\hat{\Psi} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$, a ser idêntica à diagonal de $\mathbf{S}$. O ajuste do modelo é então avaliado pela magnitude dos resíduos fora da diagonal. A matriz de resíduos é:

$$\mathbf{S} - \hat{\Psi} = \mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi})$$

Como $\hat{\Psi} = \text{diag}(\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}')$, os elementos da diagonal da matriz de resíduos são zero. Os resíduos fora da diagonal são os elementos de $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$. Pode-se demonstrar que a soma dos quadrados de todos os elementos da matriz $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$ (incluindo a diagonal) é:

$$\sum_{i=1}^{p}\sum_{j=1}^{p}(s_{ij} - \sum_{k=1}^{m}\hat{l}_{ik}\hat{l}_{jk})^2 = \sum_{k=m+1}^{p}\hat{\lambda}_k^2$$

Isso mostra que, para que o ajuste seja bom, a soma dos autovalores descartados $(\hat{\lambda}_{m+1}, \dots, \hat{\lambda}_p)$ deve ser pequena.

8.5.2 Método da Máxima Verossimilhança (MMV)

O método da máxima verossimilhança (MMV) é uma abordagem mais rigorosa para a estimação, baseada em suposições sobre a distribuição dos dados.

Suposições Adicionais:

1. O vetor de fatores comuns $\mathbf{F}$ e o vetor de erros $\boldsymbol{\epsilon}$ seguem uma distribuição normal multivariada:
   - $\mathbf{F} \sim N_m(\mathbf{0}, \mathbf{I})$
   - $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \boldsymbol{\Psi})$
2. $\mathbf{F}$ e $\boldsymbol{\epsilon}$ são independentes.

Sob essas condições, o vetor de variáveis observáveis $\mathbf{x}$ segue uma distribuição normal multivariada $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, onde $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$.

Dada uma amostra aleatória $\mathbf{x}_1, \dots, \mathbf{x}_n$, a função de log-verossimilhança (ignorando constantes) para os parâmetros $\mathbf{L}$ e $\boldsymbol{\Psi}$ é:

$$\log L(\mathbf{L}, \boldsymbol{\Psi}) = -\frac{n}{2} \ln |\boldsymbol{\Sigma}| - \frac{n}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})$$

onde $\mathbf{S}$ é a matriz de covariâncias amostral (versão ML, com divisor n). O objetivo é encontrar as estimativas $\hat{\mathbf{L}}$ e $\hat{\boldsymbol{\Psi}}$ que maximizam essa função, sujeito à restrição de que $\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}$ seja uma matriz diagonal para garantir a unicidade da solução.

A maximização é realizada por meio de algoritmos numéricos (como o de Newton-Raphson), pois não há uma solução analítica fechada. As estimativas resultantes, $\hat{\mathbf{L}}$ e $\hat{\boldsymbol{\Psi}}$, satisfazem um conjunto complexo de equações.

A principal vantagem do MMV é que ele permite um teste de hipóteses para a adequação do número de fatores m, comparando a matriz de covariâncias do modelo, $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}$, com a matriz amostral $\mathbf{S}$. Isso é fundamental na **Análise Fatorial Confirmatória (AFC)**

8.6 A Escolha do Número de Fatores (m)

A determinação do número de fatores, m, é uma das decisões mais importantes na Análise Fatorial. Um número muito baixo de fatores pode não capturar a estrutura de covariância subjacente, enquanto um número muito alto pode levar a um modelo superajustado e de difícil interpretação, violando o princípio da parcimônia.

A escolha de m geralmente envolve uma combinação de critérios estatísticos e julgamento prático. Vários dos métodos utilizados são análogos aos empregados na Análise de Componentes Principais (Capítulo 7). Os mais comuns são:

1. **Proporção da Variância Total Explicada**: Um critério comum é reter fatores suficientes para explicar uma proporção substancial (e.g., 70-90%) da variância total. No contexto do método de componentes principais para AF, a proporção da variância explicada pelo fator j é $\hat{\lambda}_j/\text{tr}(\mathbf{S})$.
2. **Critério de Kaiser (Autovalores > 1)**: Ao trabalhar com a matriz de correlação $\mathbf{R}$, o critério de Kaiser sugere reter apenas os fatores correspondentes a autovalores maiores que 1. A lógica é que um fator deve explicar pelo menos a variância de uma variável original.
3. **Gráfico de cotovelo (*Scree Plot*)**: Este é um gráfico dos autovalores ordenados ($\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$). Procura-se por um "cotovelo" no gráfico, um ponto onde a magnitude dos autovalores começa a diminuir drasticamente. O número de fatores a reter seria o número de pontos antes do início do platô.
4. **Teste de Hipóteses (para MMV)**: Quando o método da máxima verossimilhança é utilizado, é possível realizar um teste de razão de verossimilhanças para testar a hipótese nula de que m fatores são suficientes para descrever a estrutura de covariância.

Na prática, é recomendável utilizar uma combinação desses critérios. A interpretabilidade da solução fatorial resultante é, em última análise, o guia mais importante.

8.7 Rotação Fatorial

Como visto anteriormente, a solução para a matriz de cargas fatoriais $\hat{\mathbf{L}}$ não é única. Qualquer rotação ortogonal dos fatores resulta em uma nova matriz de cargas $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}$ que explica a estrutura de covariâncias dos dados exatamente da mesma forma, pois $\hat{\mathbf{L}}\hat{\mathbf{L}}' = \hat{\mathbf{L}}^*(\hat{\mathbf{L}}^*)'$.

Essa indeterminação, que a princípio parece um problema, é na verdade uma das ferramentas mais poderosas da Análise Fatorial. Ela nos permite girar a estrutura fatorial para uma posição que seja mais **simples e interpretável**, sem sacrificar o ajuste do modelo. O objetivo é alcançar o que o psicólogo Louis Thurstone chamou de **estrutura simples**.

A estrutura simples ideal teria as seguintes propriedades:

1. Cada variável deve ter pelo menos uma carga fatorial próxima de zero.
2. Cada fator deve ter várias cargas próximas de zero e algumas cargas altas.
3. Para cada par de fatores, deve haver variáveis com cargas altas em um fator, mas não no outro.

Em suma, busca-se uma matriz de cargas onde cada variável esteja fortemente associada a apenas um ou poucos fatores, e cada fator represente claramente um subconjunto de variáveis. Os métodos de rotação são algoritmos que buscam, de forma objetiva, uma matriz $\mathbf{T}$ que aproxime a matriz de cargas rotacionada $\hat{\mathbf{L}}^*$ a essa estrutura ideal.

As rotações dividem-se em duas categorias principais.

8.7.1 Rotações Ortogonais

Neste tipo de rotação, a matriz de transformação **T** é ortogonal, o que significa que os eixos dos fatores são girados, mas mantidos em um ângulo de 90 graus entre si. A consequência fundamental é que os fatores rotacionados **permanecem não correlacionados**.

Os métodos mais comuns de rotação ortogonal são:

- **Varimax**: É o método de rotação ortogonal mais popular. O objetivo do Varimax é simplificar as **colunas** da matriz de cargas fatoriais. Para cada fator, ele busca maximizar a variância das cargas ao quadrado, efetivamente empurrando as cargas para perto de 0 ou ±1. Isso facilita a identificação de quais variáveis estão associadas a cada fator. O critério Varimax maximiza a seguinte função:

$$V = \sum_{j=1}^{m} \left[\frac{1}{p} \sum_{i=1}^{p} (\hat{l}_{ij}^{*}/h_i)^4 - \left(\frac{1}{p} \sum_{i=1}^{p} (\hat{l}_{ij}^{*}/h_i)^2 \right)^2 \right]$$

  Onde $\hat{l}_{ij}^{*}$ são as cargas rotacionadas e h_i^2 são as comunalidades (que permanecem invariantes sob rotação).

- **Quartimax**: Este método foca em simplificar as **linhas** da matriz de cargas. Ele tenta fazer com que cada variável tenha carga alta em apenas um fator. O Quartimax foi o primeiro método analítico proposto, mas tende a criar um fator geral com cargas altas para muitas variáveis, o que pode dificultar a interpretação.

- **Equimax**: É um meio termo entre o Varimax e o Quartimax. Ele tenta simplificar tanto as linhas quanto as colunas da matriz de cargas simultaneamente.

8.7.2 Rotações Oblíquas

Em muitos campos, especialmente nas ciências sociais, é teoricamente razoável esperar que os fatores latentes sejam correlacionados. Por exemplo, os fatores "habilidade verbal" e "habilidade matemática" são distintos, mas é provável que sejam positivamente correlacionados.

As rotações oblíquas permitem que os fatores se tornem correlacionados. A matriz de transformação **T** não é mais ortogonal, e os eixos dos fatores podem ter ângulos diferentes de 90 graus. A vantagem é a capacidade de encontrar uma estrutura mais simples e teoricamente mais realista, ao custo de uma complexidade maior na interpretação, pois é preciso analisar tanto a matriz de cargas quanto a matriz de correlação entre os fatores.

Os métodos mais comuns incluem:

- **Promax**: É um método muito utilizado que funciona em duas etapas. Primeiro, ele realiza uma rotação ortogonal (geralmente Varimax). Em seguida, ele relaxa a restrição de ortogonalidade, permitindo que os fatores se correlacionem para buscar uma estrutura ainda mais simples (com mais cargas próximas de zero).
- **Oblimin Direto**: É um método mais geral que busca minimizar a covariância das cargas ao quadrado para pares de fatores. Ele possui um parâmetro (delta) que controla o grau de correlação permitido entre os fatores.

A escolha entre uma rotação ortogonal e oblíqua depende de considerações teóricas. Se não há uma razão forte para acreditar que os fatores são correlacionados, a rotação ortogonal (como a Varimax) é geralmente preferida por sua simplicidade. Se a correlação entre os fatores é esperada, uma rotação oblíqua pode fornecer uma representação mais fiel da realidade.

9 Análise de Agrupamentos

A Análise de Agrupamentos, ou Análise de *Clusters*, é uma técnica exploratória multivariada cujo objetivo é particionar um conjunto de observações em subgrupos (os *clusters*). A partição é feita de tal forma que as observações dentro de um mesmo grupo sejam semelhantes entre si, enquanto observações em grupos diferentes sejam o mais distintas possível.

Diferentemente de outras técnicas como a análise de regressão ou a análise discriminante, a análise de agrupamentos é um método de **aprendizagem não supervisionada**. Isso significa que não temos uma variável resposta ou rótulos pré-definidos para os grupos; o objetivo é descobrir a estrutura de agrupamentos inerente aos próprios dados.

O princípio fundamental é a maximização da homogeneidade intra-grupo e, ao mesmo tempo, a maximização da heterogeneidade entre grupos.

Diversas técnicas apresentadas nesse capítulo dependem da definição de uma medida para quantificar o quão semelhantes ou diferentes as observações são. Essa medida é formalizada como uma **medida de dissimilaridade** ou **distância**. Uma discussão detalhada sobre as diferentes métricas de distância pode ser encontrada na Capítulo 6.

Definição 9.1. Dado um conjunto de dados com n observações $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, um **agrupamento** (ou *clustering*) é uma partição de $\mathbf{X}$ em K subconjuntos disjuntos $C_1, C_2, \dots, C_K$, tal que:

1. $C_k \neq \emptyset$ para $k = 1, \dots, K$
2. $C_k \cap C_j = \emptyset$ para $k \neq j$
3. $\bigcup_{k=1}^{K} C_k = \mathbf{X}$

Para cada grupo C_k, definimos:

- **Tamanho:** $n_k = \#C_k$, o número de observações no grupo.
- **Centroide:** $\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i$, o centóide, ou vetor de médias, do grupo.

9.1 Decomposição da Variabilidade

Podemos formalizar o critério de "boa separação" dos grupos através de uma decomposição da variabilidade total dos dados, análoga à Análise de Variância (ANOVA).

Definição 9.2. Dado um conjunto de n observações e uma partição em K grupos $C_1, \dots, C_K$:

- **Soma de Quadrados Total (SQT):** Mede a dispersão total dos dados em torno da média geral $\bar{\mathbf{x}}$.
$$SQT = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})$$

- **Soma de Quadrados Intra-grupos (SQI):** Mede a dispersão dentro dos grupos. É a soma das dispersões de cada observação em relação ao centroide do seu próprio grupo, $\bar{\mathbf{x}}_k$. Também é conhecida como *Within-Cluster Sum of Squares* (WCSS).
$$SQI = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k)$$

- **Soma de Quadrados Entre-grupos (SQE):** Mede a dispersão entre os centroides dos grupos em relação à média geral.
$$SQE = \sum_{k=1}^{K} n_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$$
Onde n_k é o número de observações no grupo C_k.

O objetivo da análise de agrupamentos pode ser visto como encontrar a partição que **minimiza a SQI** (grupos coesos) e **maximiza a SQE** (grupos separados).

Teorema 9.1. *A soma de quadrados total pode ser decomposta como a soma da variabilidade dentro dos grupos e entre os grupos.*

$$SQT = SQI + SQE$$

Comprovação. A prova parte da decomposição do desvio de uma observação $\mathbf{x}_i \in C_k$ em relação à média geral $\bar{\mathbf{x}}$:

$$(\mathbf{x}_i - \bar{\mathbf{x}}) = (\mathbf{x}_i - \bar{\mathbf{x}}_k) + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$$

Elevando ao quadrado (no sentido de produto vetorial), temos:

$$(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}}) = [(\mathbf{x}_i - \bar{\mathbf{x}}_k) + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})]'[(\mathbf{x}_i - \bar{\mathbf{x}}_k) + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})]$$

$$= (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k) + (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}) + 2(\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$$

Agora, somamos sobre todas as observações $i = 1, \dots, n$. Para fazer isso, somamos primeiro dentro de cada grupo k e depois somamos os resultados sobre todos os grupos:

$$SQT = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})$$

$$= \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k) + \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}) + \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} 2(\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$$

Analisando cada termo:

1. O primeiro termo é, por definição, a Soma de Quadrados Intra-grupos (SQI).
2. No segundo termo, a expressão $(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$ é constante para todas as n_k observações no grupo C_k. Portanto, a soma interna resulta em $n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})$. Somar sobre k nos dá a Soma de Quadrados Entre-grupos (SQE).
3. Para o terceiro termo (o termo cruzado), podemos reescrevê-lo como:

$$2 \sum_{k=1}^{K} \left[\left(\sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \right)' (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}) \right]$$

   Pela definição do centroide $\bar{\mathbf{x}}_k$, a soma dos desvios em torno dele dentro de um grupo é zero: $\sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k) = \mathbf{0}$. Portanto, todo o terceiro termo é igual a zero.

Juntando os resultados, obtemos $SQT = SQI + SQE$. □

A prova deste teorema mostra que a variabilidade total é conservada e apenas particionada, de forma que minimizar a SQI é equivalente a maximizar a SQE.

9.2 Agrupamentos Hierárquicos

Os métodos de agrupamento hierárquico criam uma sequência de partições aninhadas, que pode ser representada visualmente por uma árvore chamada **dendrograma**. Existem duas abordagens principais:

1. **Aglomerativa (Bottom-Up):** Começa com cada observação em seu próprio grupo e, a cada passo, funde os dois grupos mais próximos até que reste apenas um único grupo contendo todas as observações.
2. **Divisiva (Top-Down):** Começa com todas as observações em um único grupo e, a cada passo, divide um grupo em dois até que cada observação esteja em seu próprio grupo.

> **Nota**
>
> Devido a dificuldades de implementação, agrupamentos divisivos raramente são utilizados na prática. Por esse motivo, os exemplos presentes nesta seção consideram apenas agrupamentos aglomerativos.

9.2.1 Métodos de Ligação (*Linkage*)

Enquanto as medidas de dissimilaridade retratam a distância entre observações, precisamos de uma regra que define a distância entre dois grupos. Para isso definimos alguns métodos de ligação populares a seguir.

- **Ligação Simples (*Single Linkage*):** A distância entre dois grupos é a distância mínima entre quaisquer dois pontos dos grupos. Tende a produzir grupos "alongados" e é sensível a ruído.

$$d(A, B) = \min_{\mathbf{a} \in A, \mathbf{b} \in B} d(\mathbf{a}, \mathbf{b})$$

- **Ligação Completa (*Complete Linkage*):** A distância é o máximo da distância entre quaisquer dois pontos. Produz grupos mais compactos e esféricos.

$$d(A, B) = \max_{\mathbf{a} \in A, \mathbf{b} \in B} d(\mathbf{a}, \mathbf{b})$$

- **Ligação Média (*Average Linkage*):** A distância é a média de todas as distâncias entre os pares de pontos dos dois grupos. É um meio-termo entre a simples e a completa.

$$d(A, B) = \frac{1}{n_A n_B} \sum_{\mathbf{a} \in A} \sum_{\mathbf{b} \in B} d(\mathbf{a}, \mathbf{b})$$

- **Método de Ward:** Este método se baseia em um critério de minimização da variância. A cada passo do algoritmo aglomerativo, ele funde o par de grupos que leva ao **menor aumento possível na Soma de Quadrados Intra-grupos (SQI)**. O objetivo é encontrar, a cada passo, a fusão mais "econômica" em termos de perda de coesão interna.

  Suponha que estejamos considerando fundir dois grupos, C_i e C_j. O aumento na SQI, que denotamos por $\Delta(C_i, C_j)$, é a diferença entre a SQI do novo grupo fundido (C_{ij}) e a soma das SQIs dos grupos individuais antes da fusão.

$$\Delta(C_i, C_j) = \text{SQI}(C_{ij}) - (\text{SQI}(C_i) + \text{SQI}(C_j))$$

A SQI para o novo grupo $C_{ij} = C_i \cup C_j$ é $\sum_{\mathbf{x} \in C_{ij}} (\mathbf{x} - \bar{\mathbf{x}}_{ij})'(\mathbf{x} - \bar{\mathbf{x}}_{ij})$, onde $\mathbf{x}_{ij}$ é o centroide do novo grupo. Podemos reescrever essa soma como:

$$\sum_{\mathbf{x} \in C_i} (\mathbf{x} - \bar{\mathbf{x}}_{ij})'(\mathbf{x} - \bar{\mathbf{x}}_{ij}) + \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \bar{\mathbf{x}}_{ij})'(\mathbf{x} - \bar{\mathbf{x}}_{ij})$$

Usando a mesma lógica da decomposição da variância, podemos mostrar que $\sum_{\mathbf{x} \in C_i} (\mathbf{x} - \bar{\mathbf{x}}_{ij})'(\mathbf{x} - \bar{\mathbf{x}}_{ij}) = \text{SQI}(C_i) + n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ij})'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ij})$.

Aplicando o resultado para ambos os termos, a SQI do novo grupo é:

$$\text{SQI}(C_{ij}) = \text{SQI}(C_i) + \text{SQI}(C_j) + n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ij})'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ij}) + n_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{ij})'(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{ij})$$

Portanto, o aumento na SQI é:

$$\Delta(C_i, C_j) = n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ij})'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{ij}) + n_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{ij})'(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_{ij})$$

Substituindo $\bar{\mathbf{x}}_{ij} = \frac{n_i \bar{\mathbf{x}}_i + n_j \bar{\mathbf{x}}_j}{n_i + n_j}$ e simplificando a álgebra, chegamos a:

$$\Delta(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$$

A fórmula final nos dá uma maneira eficiente de calcular o critério de Ward. A cada passo, o algoritmo calcula $\Delta(C_i, C_j)$ para todos os pares de grupos e realiza a fusão para o par que tiver o menor valor. Note que a fórmula depende da distância euclidiana entre centroides, por esse motivo, o método de Ward tende a produzir grupos de tamanho semelhante e formato esférico.

9.2.2 O Dendrograma

O dendrograma é a principal ferramenta de visualização para o agrupamento hierárquico. Ele mostra como, passo a passo, as observações são fundidas em grupos. O eixo Y representa a distância ou dissimilaridade em que as fusões ocorrem. Quanto mais alta a "ponte" que une dois grupos, mais diferentes eles são.

O gráfico abaixo, gerado a partir de um exemplo simples de 4 observações, mostra a anatomia de um dendrograma.

O exemplo abaixo apresenta os dendogramas de maneira prática, além de exemplificar também como a função de ligação escolhida impacta no agrupamento formado.

Exemplo 9.1. Vamos analisar o comportamento dos métodos de ligação com a matriz de 5x5 abaixo.

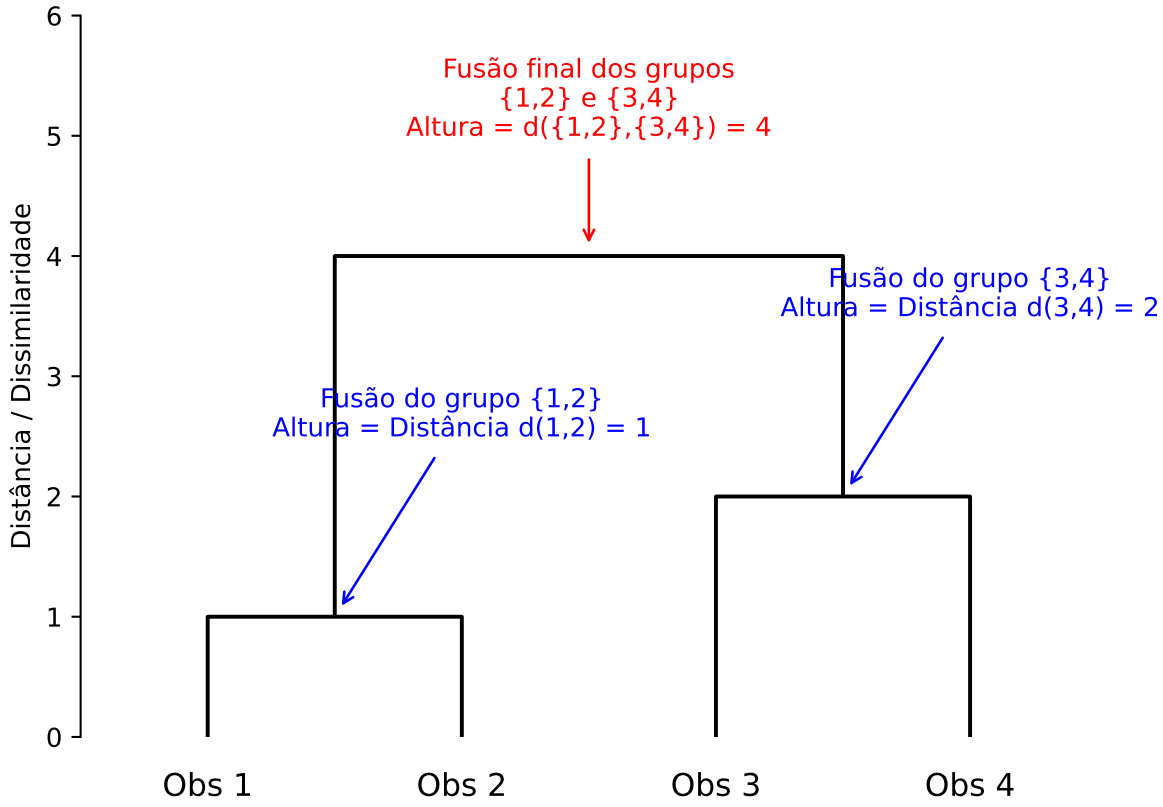

Figura 9.1: Anatomia de um dendrograma simples.

$$\mathbf{D} = \left(\begin{array}{c|ccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \hline \mathbf{1} & 0 & & & & \\ \mathbf{2} & 1 & 0 & & & \\ \mathbf{3} & 8 & 2 & 0 & & \\ \mathbf{4} & 6 & 4 & 11 & 0 & \\ \mathbf{5} & 7 & 9 & 8 & 7 & 0 \end{array}\right)$$

1. Usando a Ligação Simples (*Single Linkage*)

- **Passo 1:** A menor distância na matriz é `d(1,2) = 1`. Fundimos `{1,2}` na altura 1.
- **Passo 2:** A distância do novo grupo `{1,2}` para `{3}` é `min(d(1,3), d(2,3)) = min(8, 2) = 2`. Todas as outras distâncias entre grupos ou pontos restantes são maiores que 2. Assim, fundimos `{1,2}` com `{3}` na altura 2.
- **Passo 3:** A distância de `{1,2,3}` para `{4}` é `min(d(1,4), d(2,4), d(3,4)) = min(6, 4, 11) = 4`. Esta é a próxima menor distância, então fundimos `{1,2,3}` com `{4}` na altura 4.
- **Passo 4:** A distância de `{1,2,3,4}` para `{5}` é `min(d(1,5), d(2,5), d(3,5), d(4,5)) = min(7, 9, 8, 7) = 7`. Fundimos o último ponto na altura 7.
- **Resultado:** O método cria uma longa cadeia: `((({1,2},3),4),5)`.

2. Usando a Ligação Completa (*Complete Linkage*)

- **Passo 1:** A fusão inicial é a mesma: `{1,2}` (altura 1).
- **Passo 2:** A distância de `{1,2}` para `{4}` é `max(d(1,4), d(2,4)) = max(6, 4) = 6`. Já a distância para `{3}` é `max(d(1,3), d(2,3)) = max(8, 2) = 8`. A menor distância entre os grupos existentes é 6, então fundimos `{1,2}` com `{4}`.
- **Passo 3:** Temos os grupos `{1,2,4}`, `{3}` e `{5}`. A próxima menor distância entre os grupos restantes é `d(3,5) = 8`. Fundimos `{3,5}`.
- **Passo 4:** A distância entre `{1,2,4}` e `{3,5}` é `max(d(1,3), d(1,5), d(2,3), d(2,5), d(4,3), d(4,5)) = max(8, 7, 2, 9, 11, 7) = 11`. A fusão final ocorre na altura 11.
- **Resultado:** O método cria dois grupos distintos, `({1,2,4}, {3,5})`, antes da fusão final.

Os resultados são estruturalmente diferentes, como mostram os dendrogramas.

```
import numpy as np
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.spatial.distance import squareform
import matplotlib.pyplot as plt

# Matriz de distância 5x5 do exemplo
D = np.array([
    [0, 1, 8, 6, 7],
    [1, 0, 2, 4, 9],
    [8, 2, 0, 11, 8],
```

```
    [6, 4, 11, 0, 7],
    [7, 9, 8, 7, 0]
])

condensed_D = squareform(D)
labels = ['1', '2', '3', '4', '5']

fig, axes = plt.subplots(1, 2, figsize=(7, 5))
fig.suptitle('Dendrogramas Puros (Sem Definição de grupos)')

# Ligação Simples
linked_single = linkage(condensed_D, 'single')
dendrogram(linked_single, orientation='top', labels=labels, ax=axes[0], color_threshold=0, above_
axes[0].set_title('Ligação Simples')
axes[0].set_xlabel('Observação')
axes[0].set_ylabel('Distância')

# Ligação Completa
linked_complete = linkage(condensed_D, 'complete')
dendrogram(linked_complete, orientation='top', labels=labels, ax=axes[1], color_threshold=0, abov
axes[1].set_title('Ligação Completa')
axes[1].set_xlabel('Observação')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

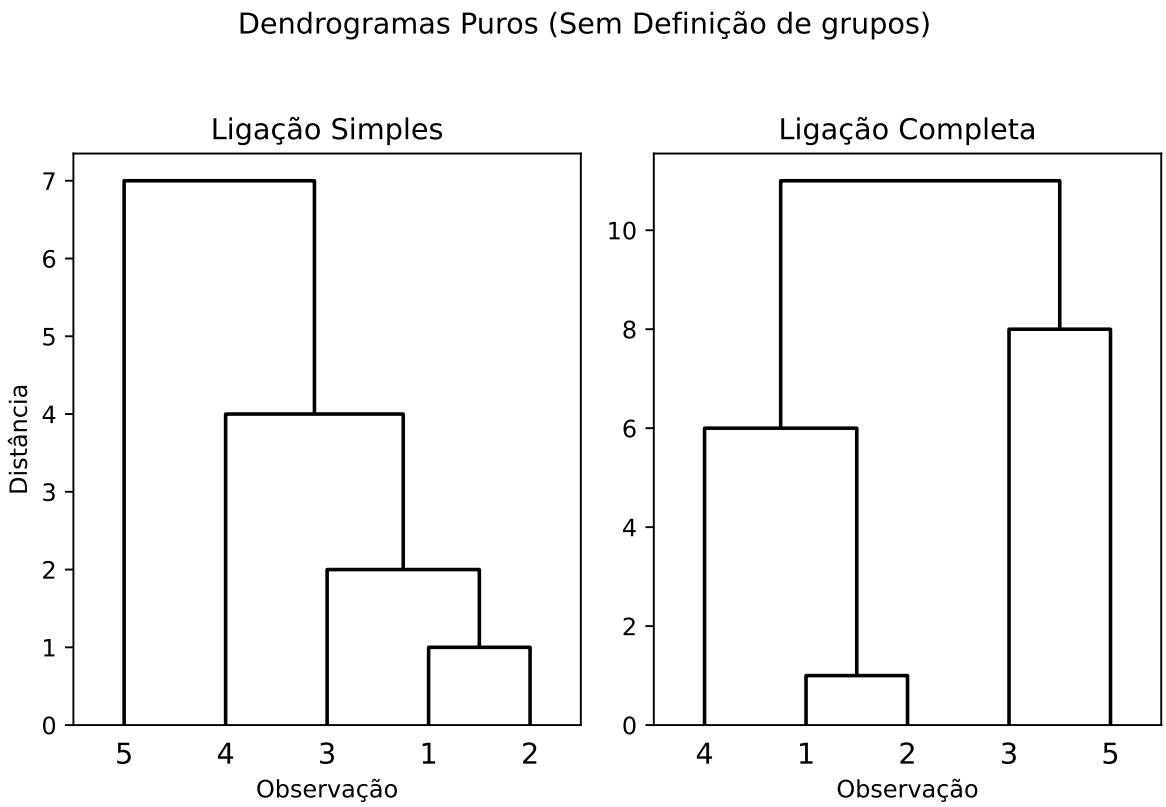

Figura 9.2: Comparação de dendrogramas para a matriz 5x5. As cores foram removidas para mostrar a estrutura pura.

O dendrograma não apenas mostra a hierarquia das fusões, mas também é a principal ferramenta para decidir o número final de grupos. A estratégia consiste em "cortar" a árvore em uma determinada altura. Todas as ramificações que estão abaixo da linha de corte constituem os grupos.

A regra geral é procurar por um corte que cruze as conexões mais longas. Uma conexão longa representa uma fusão que ocorreu a uma distância (ou aumento de SQI, no caso de Ward) muito maior do que as fusões anteriores. Isso sugere que estamos unindo grupos que são naturalmente muito diferentes entre si. Portanto, cortar a árvore logo acima dessa grande "distância de fusão" é uma escolha sensata.

A Figura 9.3 demonstra como a escolha de diferentes alturas de corte leva a diferentes números de grupos.

```
import numpy as np
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.spatial.distance import squareform
```

```
import matplotlib.pyplot as plt

# Matriz de distância 5x5 do exemplo
D = np.array([
    [0, 1, 8, 6, 7],
    [1, 0, 2, 4, 9],
    [8, 2, 0, 11, 8],
    [6, 4, 11, 0, 7],
    [7, 9, 8, 7, 0]
])

condensed_D = squareform(D)
labels = ['1', '2', '3', '4', '5']
linked_complete = linkage(condensed_D, 'complete')

fig, axes = plt.subplots(1, 3, figsize=(7, 5))
fig.suptitle('Visualização dos Cortes no Dendrograma (Ligação Completa)')

# --- Corte para K=2 ---
cut_height_k2 = 9
dendrogram(
    linked_complete,
    orientation='top',
    labels=labels,
    ax=axes[0],
    color_threshold=0,
    above_threshold_color='k'
)
axes[0].axhline(y=cut_height_k2, color='r', linestyle='--')
axes[0].set_title('Corte para K=2 grupos')
axes[0].set_xlabel('Observação')
axes[0].set_ylabel('Distância')

# --- Corte para K=3 ---
cut_height_k3 = 7
dendrogram(
    linked_complete,
    orientation='top',
    labels=labels,
    ax=axes[1],
    color_threshold=0,
    above_threshold_color='k'
```

```
)
axes[1].axhline(y=cut_height_k3, color='r', linestyle='--')
axes[1].set_title('Corte para K=3 grupos')
axes[1].set_xlabel('Observação')

# --- Corte para K=4 ---
cut_height_k4 = 3
dendrogram(
    linked_complete,
    orientation='top',
    labels=labels,
    ax=axes[2],
    color_threshold=0,
    above_threshold_color='k'
)
axes[2].axhline(y=cut_height_k4, color='r', linestyle='--')
axes[2].set_title('Corte para K=4 grupos')
axes[2].set_xlabel('Observação')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

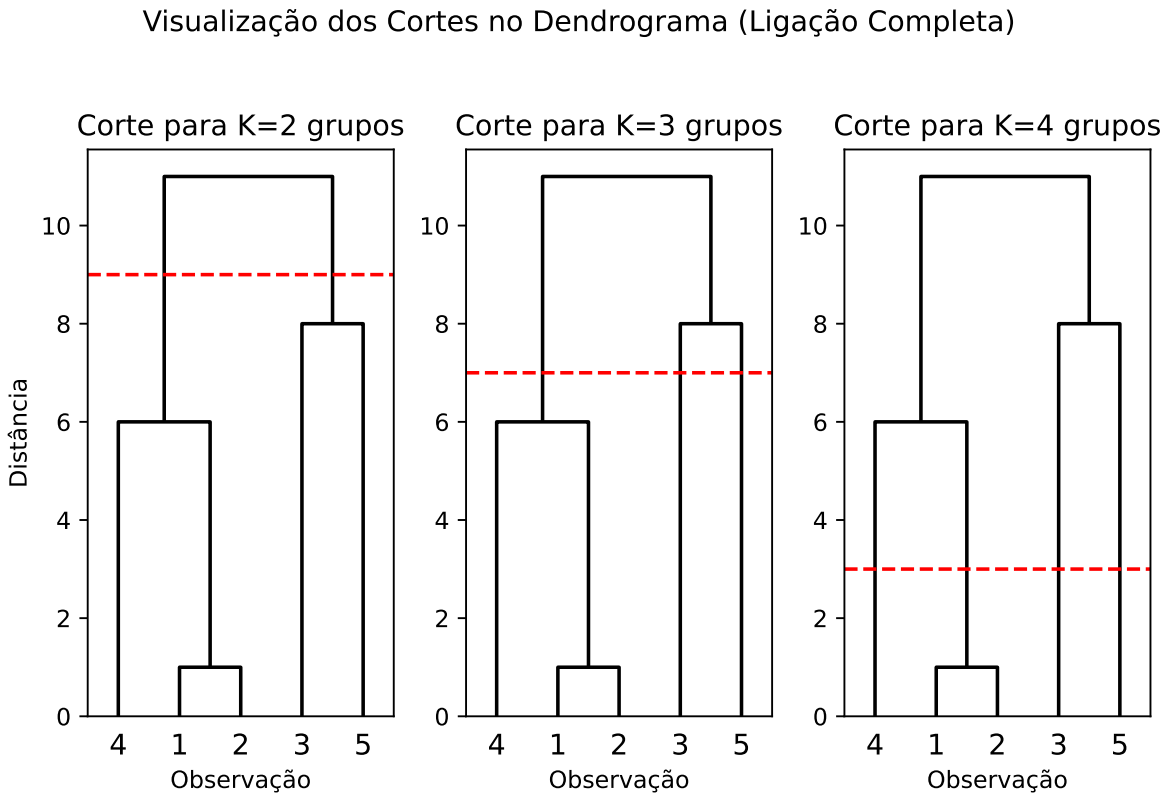

Figura 9.3: Ilustração do corte no dendrograma de Ligação Completa para obter K=2 (esquerda), K=3 (centro) e K=4 (direita) grupos.

A interpretação dos resultados para cada corte é a seguinte:

- **K=2:** Ao cortar o dendrograma na altura 9, obtemos dois grupos. O primeiro é `{1, 2, 4}` e o segundo é `{3, 5}`. Esta partição representa a estrutura de mais alto nível nos dados, separando os dois grupos mais distintos.
- **K=3:** Abaixando a linha de corte para a altura 7, o grupo `{3, 5}` (que era formado na altura 8) é quebrado. O resultado são três grupos: `{1, 2, 4}`, `{3}` e `{5}`.
- **K=4:** Com um corte ainda mais baixo, na altura 3, quebramos o grupo `{1, 2, 4}` (formado na altura 6). Os grupos resultantes são `{1, 2}`, `{4}`, `{3}` e `{5}`.

Observe que o primeiro grupo `{1, 2}` se forma em uma altura de uma unidade de distância. Já a segunda ligação, `{1, 2}` com `{4}` ocorre na altura `d({1, 2}, {4})=6`. Isso indica que já existe um salto na distância da ligação logo no segundo passo. Por isso, uma escolha sensível é $K = 4$.

> **O Eixo Vertical no Dendrograma**
>
> Para os métodos de ligação simples, completa e média o eixo vertical do dendrograma represente diretamente a distância da fusão. Já para o **Método de Ward**, a altura da fusão representa o **aumento na Soma de Quadrados Intra-grupos (SQI)** resultante da união dos dois grupos. Esse valor não é uma distância, mas sim uma medida de perda de homogeneidade.

9.2.3 Limitações do Agrupamento Hierárquico

Apesar de sua simplicidade e elegância, os métodos hierárquicos possuem limitações importantes. Uma delas é sua complexidade computacional, que cresce rapidamente para dados com muitas observações devido a grande quantidade de comparações.

Além disso, a decisão de fusão é final e não pode ser desfeita. Se um grupo inicial for inadequada, o erro se propagará por toda a hierarquia. Ou seja, o agrupamento é muito sensível à estrutura inicial. Finalmente, todos os métodos de ligação carregam suas vantagens e problemas.

Essas limitações motivam o uso de métodos não hierárquicos, como o K-médias, que abordaremos a seguir.

9.3 Agrupamento Não-Hierárquico

Diferente dos métodos hierárquicos, os métodos particionais, como o K-Médias, dividem os dados em um número K de grupos pré-especificado. O K-Médias é um dos algoritmos de agrupamento mais populares e eficientes.

O objetivo do K-Médias é particionar as n observações em K grupos de modo a minimizar a Soma de Quadrados Intra-grupos (SQI), também chamada de inércia.

$$\text{SQI} = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k)$$

Onde $\bar{\mathbf{x}}_k$ é o centroide (média) do grupo C_k.

9.3.1 O Algoritmo K-Médias (*K-Means*)

O algoritmo K-médias é um processo iterativo que busca minimizar a SQI. Seus passos podem ser definidos matematicamente da seguinte forma:

1. **Inicialização:** Escolha K centroides iniciais $\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}$. Esta escolha pode ser feita selecionando K observações aleatórias do conjunto de dados.

2. **Atribuição:** Em cada iteração t, atribua cada observação $\mathbf{x}_i$ ao grupo cujo centroide é o mais próximo. Matematicamente, para cada observação i, encontramos o índice do grupo c_i que minimiza a distância Euclidiana quadrada:

$$c_i^{(t)} = \arg\min_k (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t-1)})'(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t-1)})$$

   Isso particiona os dados nos conjuntos $C_1^{(t)}, \ldots, C_K^{(t)}$.

3. **Atualização:** Recalcule o centroide de cada grupo como a média de todas as observações atribuídas a ele na iteração atual:

$$\boldsymbol{\mu}_k^{(t)} = \frac{1}{\#C_k^{(t)}} \sum_{\mathbf{x}_i \in C_k^{(t)}} \mathbf{x}_i$$

4. **Repetição:** Repita os passos 2 e 3 até que as atribuições dos grupos não mudem mais, ou seja, $c_i^{(t)} = c_i^{(t-1)}$ para todas as observações i.

O grande problema do método K-médias é a escolha dos centroides iniciais. O algoritmo tem a garantia de convergir, mas dependendo das condições iniciais pode chegar a um mínimo local, e não necessariamente o mínimo global da SQI. Uma opção comum é executar o algoritmo várias vezes com diferentes inicializações e escolher o resultado com a menor SQI.

A necessidade de pré-especificar o número de grupos, K, também pode ser uma desvantagem. Para contornar o problema, é comum executar o algoritmo para uma gama de valores de K e calcular a Soma de Quadrados Intra-grupos (SQI) em cada caso. A ideia é escolher K tal que a SQI seja baixa, mas sendo parcimonioso com o número de grupos. Vale lembrar que a SQI é inversamente proporcional a K – se $K = 1$, a SQI é máxima e se $K = n$ a SQI é zero. Na prática, aumentamos K progressivamente observando os decréscimo na SQI, seguimos aumentando K enquanto esse decréscimo for grande.

A seguir, definimos um procedimento mais robusto, que soluciona os problemas mencionados combinando diferentes métodos de agrupamento.

9.3.2 Procedimento sugerido para análise de agrupamentos

Levando em consideração as vantagens e desvantagens dos agrupamentos hierárquicos e de K-médias, podemos definir um procedimento que simplifica as escolhas durante um problema prático de análise de agrupamentos.

- Inicie com um agrupamento hierárquico. O método de Ward costuma ser uma boa primeira opção, no entanto, experimente também outros métodos de ligação conforme necessário.

- Observe o dendograma para escolher o corte que seja mais plausível. Esse corte determina um número de grupos ótimo K^*.
- Calcule o centroide para cada um dos grupos obtidos via agrupamento hierárquico $\mu_1, \cdots, \mu_K$.
- Utilize os centoides obtidos como centroides iniciais para o método K-médias com K^* grupos.

Exemplificamos esse procedimento com dados reais no Capítulo 12.

9.4 Agrupamento Baseado em Modelos

Uma abordagem mais avançada e flexível é o agrupamento baseado em modelos. A ideia central é assumir que os dados são gerados a partir de uma **mistura de distribuições de probabilidade**, onde cada componente da mistura corresponde a um grupo.

O modelo mais comum é o **Modelo de Mistura Gaussiana (*Gaussian Mixture Model*, GMM)**. Ele assume que cada grupo segue uma distribuição normal multivariada. A densidade de probabilidade de todo o conjunto de dados é uma soma ponderada de K densidades Gaussianas:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boxtimes_k)$$

Onde, para cada grupo k:

- π_k: é o peso da mistura. Por ser uma probabilidade, deve satisfazer $\pi_k \geq 0$ e $\sum_{k=1}^{K} \pi_k = 1$.
- $\boldsymbol{\mu}_k$: é o vetor de médias (o centroide do grupo).
- $\boxtimes_k$: é a matriz de covariâncias (descreve a forma e orientação do grupo).

9.4.1 O Algoritmo de Expectation-Maximization (EM)

Os parâmetros do GMM $(\pi_k, \boldsymbol{\mu}_k, \boxtimes_k)$ são estimados usando o algoritmo de **Expectation-Maximization (EM)**, um processo iterativo que alterna entre dois passos.

1. **Inicialização:** Inicialize os parâmetros do modelo $\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boxtimes_k^{(0)}$ para cada grupo $k = 1, \ldots, K$. Isso pode ser feito de forma aleatória ou usando o resultado de um algoritmo mais simples, como o K-médias.
2. **Passo-E (Expectation):** Em cada iteração t, calculamos a "responsabilidade" de cada grupo k por cada observação $\mathbf{x}_i$. A responsabilidade, denotada por $p_{ik}^{(t)}$, é a probabilidade posterior

de que a observação $\mathbf{x}_i$ tenha sido gerada pela componente k, dados os parâmetros da iteração anterior. Usando a regra de Bayes, a responsabilidade é calculada como:

$$p_{ik}^{(t)} = \frac{\pi_k^{(t-1)} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^{K} \pi_j^{(t-1)} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\Sigma}_j^{(t-1)})}$$

Essencialmente, para cada ponto, calculamos a probabilidade de ele pertencer a cada um dos K grupos.

3. **Passo-M (Maximization):** Nesta etapa, usamos as responsabilidades $p_{ik}^{(t)}$ para reestimar os parâmetros do modelo, maximizando a verossimilhança esperada. As atualizações para a iteração t são:

   - **Pesos da mistura:** O peso $\pi_k^{(t)}$ é a proporção média de responsabilidade do grupo k sobre todas as observações.
   $$\pi_k^{(t)} = \frac{N_k^{(t)}}{n}, \quad \text{onde } N_k^{(t)} = \sum_{i=1}^{n} p_{ik}^{(t)}$$

   - **Médias:** A média $\boldsymbol{\mu}_k^{(t)}$ é uma média ponderada de todas as observações, onde os pesos são as responsabilidades.
   $$\boldsymbol{\mu}_k^{(t)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} p_{ik}^{(t)} \mathbf{x}_i$$

   - **Covariâncias:** A matriz de covariâncias $\boldsymbol{\Sigma}_k^{(t)}$ é uma média ponderada das covariâncias, centrada na nova média.
   $$\boldsymbol{\Sigma}_k^{(t)} = \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} p_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)})'$$

4. **Repetição:** Os passos E e M são repetidos até que a log-verossimilhança do modelo, $\ln p(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$, convirja para um valor estável.

Exemplificamos a aplicação do modelo de mistura Gaussiana para análise de agrupamentos na Capítulo 13.

9.4.2 Vantagens do Agrupamento Baseado em Modelos

- **Agrupamento "Suave":** O GMM fornece uma probabilidade de pertencimento a cada grupo para cada observação, em vez de uma atribuição "dura" (tudo ou nada) como o K-Médias.

- **Flexibilidade de Formato:** Ao permitir que cada grupo tenha sua própria matriz de covariâncias $\boxtimes_k$, os GMMs podem identificar grupos com diferentes formas (elípticas) e orientações, enquanto o K-Médias assume implicitamente que os grupos são esféricos.
- **Seleção de Modelo Criteriosa:** A natureza estatística do método permite o uso de critérios de informação, como o **Critério de Informação Bayesiano (BIC)** ou o **Critério de Informação de Akaike (AIC)**, para selecionar o número ideal de grupos K e a estrutura da matriz de covariâncias de forma mais objetiva. O modelo com o menor valor de BIC é geralmente preferido.

9.5 Índices de validação de agrupamentos

Uma das perguntas mais importantes na análise de grupo é "qual o número ideal de grupos?". Embora o procedimento que combina o método hierárquico com o K-médias ajude a guiar essa escolha, existem diversas métricas quantitativas que avaliam a qualidade de uma partição de grupos, independentemente do método utilizado para obtê-la. Esses índices nos permitem comparar os resultados para diferentes valores de K e escolher o que for estatisticamente mais robusto.

9.5.1 Método da Silhueta

A análise de silhueta é uma das técnicas mais populares. Ela mede o quão bem cada observação se encaixa em seu grupo em comparação com outros grupos. O coeficiente de silhueta para uma única observação i é:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Onde: - $a(i)$: A distância média de i para todas as outras observações no *mesmo* grupo (coesão). - $b(i)$: A distância média de i para todas as observações no grupo *vizinho mais próximo* (separação).

O valor de $s(i)$ varia de -1 a 1. Para um dado K, calculamos o coeficiente de silhueta médio para todas as observações. O valor de K que **maximiza** a pontuação média da silhueta é considerado o ideal.

9.5.2 Índice de Calinski-Harabasz

Também conhecido como Critério da Razão de Variâncias, este índice avalia a qualidade do agrupamento pela razão entre a dispersão entre os grupos e a dispersão intra-grupo. A fórmula é:

$$CH(K) = \frac{SQE/(K-1)}{SQI/(n-K)}$$

Onde *SQE* é a Soma de Quadrados Entre-grupos, *SQI* é a Soma de Quadrados Intra-grupos, n é o número total de observações e K é o número de grupos.

Intuitivamente, um bom agrupamento tem uma alta dispersão entre os grupos (SQE alta) e uma baixa dispersão dentro dos grupos (SQI baixa). Portanto, procuramos o valor de K que **maximiza** o índice de Calinski-Harabasz.

9.5.3 Índice de Davies-Bouldin

Este índice mede a "similaridade" média de cada grupo com seu grupo mais semelhante. A similaridade entre dois grupos C_i e C_j é definida como:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Onde s_i é a dispersão média dentro do grupo i (por exemplo, a distância média de cada ponto ao centroide) e d_{ij} é a distância entre os centroides dos grupos i e j.

Para cada grupo i, encontramos o grupo j que maximiza R_{ij}. O índice de Davies-Bouldin é a média desses valores máximos sobre todos os grupos:

$$DB(K) = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i}(R_{ij})$$

Um valor baixo de $DB(K)$ indica que os grupos estão bem separados e compactos. Portanto, procuramos o valor de K que **minimiza** o índice de Davies-Bouldin.

Em resumo, a escolha de K deve ser guiada por uma combinação dessas ferramentas, a análise do dendrograma e, mais importante, o contexto do problema. Se os grupos resultantes não forem interpretáveis ou úteis, o valor de K deve ser reconsiderado.

9.6 Interpretações em análises de agrupamentos

Uma vez que os grupos são formados, o passo final e mais importante é a **interpretação**. Um agrupamento matematicamente sólido é inútil se não pudermos extrair dele insights práticos. A interpretação envolve a caracterização e a validação dos grupos.

9.6.1 Perfil dos grupos

O objetivo aqui é responder à pergunta: "Quem são os membros de cada grupo?". Para isso, descrevemos cada grupo em termos das variáveis usadas na análise (e também de outras variáveis descritivas, se disponíveis). O método mais comum é calcular estatísticas descritivas para cada variável, segmentadas por grupo.

- **Variáveis Contínuas:** Calcule a média e o desvio padrão de cada variável para cada grupo. Em seguida, compare a média de um grupo com a média geral da amostra. Por exemplo: "O grupo 2 tem uma renda média 30% acima da média geral, enquanto o grupo 1 tem uma renda 20% abaixo".
- **Variáveis Categóricas:** Calcule a distribuição de frequência (ou porcentagens) de cada categoria para cada grupo. Por exemplo: "O grupo 3 é composto por 80% de clientes do sexo feminino, enquanto na amostra total essa proporção é de 55%".

A criação de uma tabela de perfis, com os grupos nas linhas e as estatísticas das variáveis nas colunas, é uma ferramenta extremamente eficaz para resumir e comunicar os resultados.

Depois de entender o perfil de cada grupo, é uma boa prática dar a eles nomes descritivas. Nomes como "Jovens Urbanos Conectados", "Famílias Rurais Conservadoras" ou "Clientes de Alto Risco" são muito mais fáceis de comunicar e lembrar do que "Grupo 1", "Grupo 2", etc. O nome deve capturar a essência do que torna aquele grupo único.

A interpretação é um processo iterativo. Às vezes, os perfis dos grupos podem sugerir que o número de grupos escolhido não foi o ideal, levando o analista a revisitar as etapas anteriores.

9.6.2 Visualização do agrupamento

A visualização é pode auxiliar na validação da separação dos grupos.

- **Dados de Baixa Dimensão (2D ou 3D):** Um simples gráfico de dispersão, com os pontos identificados de acordo com seu grupo designado, é suficiente.
- **Dados de Alta Dimensão:** Quando temos mais de três variáveis, precisamos primeiro reduzir a dimensionalidade para poder visualizar. A **Análise de Componentes Principais (ACP)** é a técnica mais comum para isso. Podemos plotar as observações em um gráfico de dispersão usando os dois primeiros componentes principais como eixos e identificar os pontos por grupo. Isso nos dá uma visão da separação dos grupos no espaço que captura a maior parte da variabilidade dos dados.

Part III

Exemplos

10 Exemplo manual: ACP

Neste exemplo, vamos detalhar passo a passo a aplicação da Análise de Componentes Principais (ACP) em um pequeno conjunto de dados. O objetivo é demonstrar manualmente todos os cálculos, desde a preparação dos dados até a interpretação dos resultados, seguindo a metodologia apresentada no Capítulo 3.

10.1 O Cenário

Vamos expandir o exemplo da intuição geométrica, que usava **Peso** e **Altura**. Adicionaremos uma terceira variável, **Renda** (em milhares de R$), para um grupo de 5 indivíduos. A ideia é que Peso e Altura sejam correlacionados, mas a Renda não tenha uma correlação forte com eles.

Nosso conjunto de dados inicial é:

Indivíduo	Peso (kg)	Altura (cm)	Renda (R$ 1000)
1	65	170	5.5
2	72	182	4.0
3	58	165	7.0
4	81	190	3.5
5	75	178	5.0

10.2 Passo 1: Preparação dos Dados

Conforme discutido no Capítulo 3, a ACP é sensível à escala das variáveis. Portanto, o primeiro passo é **padronizar** os dados. Isso envolve duas etapas: centralizar (subtrair a média) e escalonar (dividir pelo desvio padrão).

10.2.1 1.1. Calcular a Média e o Desvio Padrão

Primeiro, calculamos a média e o desvio padrão para cada variável.

$$\bar{x}_{peso} = \frac{65 + 72 + 58 + 81 + 75}{5} = 70.2\,\text{kg}$$

$$\bar{x}_{altura} = \frac{170 + 182 + 165 + 190 + 178}{5} = 177.0\,\text{cm}$$

$$\bar{x}_{renda} = \frac{5.5 + 4.0 + 7.0 + 3.5 + 5.0}{5} = 5.0\,(R\$1000)$$

Agora, os desvios padrão (usando a fórmula com denominador $n - 1$):

$$s_{peso} = \sqrt{\frac{(65 - 70.2)^2 + \ldots + (75 - 70.2)^2}{4}} = 8.64\,\text{kg}$$

$$s_{altura} = \sqrt{\frac{(170 - 177)^2 + \ldots + (178 - 177)^2}{4}} = 9.67\,\text{cm}$$

$$s_{renda} = \sqrt{\frac{(5.5 - 5.0)^2 + \ldots + (5.0 - 5.0)^2}{4}} = 1.35\,(R\$1000)$$

10.2.2 1.2. Padronizar os Dados

Com as médias e desvios padrão, podemos padronizar cada observação x_{ij} usando a fórmula $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$.

Por exemplo, para o Indivíduo 1:

$$z_{1,peso} = \frac{65 - 70.2}{8.64} = -0.60$$

$$z_{1,altura} = \frac{170 - 177}{9.67} = -0.72$$

$$z_{1,renda} = \frac{5.5 - 5.0}{1.35} = 0.37$$

Aplicando isso a todos os dados, obtemos a matriz de dados padronizados $\mathbf{Z}$:

$$\mathbf{Z} = \begin{pmatrix} -0.60 & -0.72 & 0.37 \\ 0.21 & 0.52 & -0.74 \\ -1.41 & -1.24 & 1.48 \\ 1.25 & 1.34 & -1.11 \\ 0.56 & 0.10 & 0.00 \end{pmatrix}$$

10.3 Passo 2: Calcular a Matriz de Correlação

Como estamos trabalhando com dados padronizados, a ACP será realizada sobre a **matriz de correlação R**. A matriz de correlação pode ser calculada como:

$$\mathbf{R} = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z}$$

Calculando $\mathbf{Z}'\mathbf{Z}$:

$$\mathbf{Z}'\mathbf{Z} = \begin{pmatrix} 4.00 & 3.85 & -0.81 \\ 3.85 & 4.00 & -1.18 \\ -0.81 & -1.18 & 4.00 \end{pmatrix}$$

Dividindo por $n - 1 = 4$, obtemos a matriz de correlação $\mathbf{R}$:

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.96 & -0.20 \\ 0.96 & 1.00 & -0.29 \\ -0.20 & -0.29 & 1.00 \end{pmatrix}$$

Como esperado, a correlação entre Peso e Altura (0.96) é muito alta, enquanto a Renda tem uma correlação fraca e negativa com as outras duas variáveis.

10.4 Passo 3: Decomposição Espectral da Matriz de Correlação

O próximo passo é encontrar os autovalores (λ) e autovetores ($\mathbf{e}$) da matriz de correlação $\mathbf{R}$. Eles são a solução da equação $\mathbf{Re} = \lambda\mathbf{e}$, que é equivalente a resolver $(\mathbf{R} - \lambda\mathbf{I})\mathbf{e} = \mathbf{0}$.

Isso requer encontrar as raízes do polinômio característico $det(\mathbf{R} - \lambda\mathbf{I}) = 0$.

$$det\begin{pmatrix} 1.00 - \lambda & 0.96 & -0.20 \\ 0.96 & 1.00 - \lambda & -0.29 \\ -0.20 & -0.29 & 1.00 - \lambda \end{pmatrix} = 0$$

Resolver este determinante cúbico manualmente é trabalhoso. Usando uma calculadora ou software, encontramos os seguintes autovalores:

$$\lambda_1 = 1.98 \quad \lambda_2 = 1.00 \quad \lambda_3 = 0.02$$

10.4.1 Interpretação dos Autovalores

A variância total no sistema é a soma dos autovalores (que é igual ao traço da matriz $\mathbf{R}$, ou seja, 3). - **Variância Total** = $1.98 + 1.00 + 0.02 = 3.00$

A proporção da variância explicada por cada componente é: - **CP1**: $\frac{1.98}{3.00} = 66.0\%$ - **CP2**: $\frac{1.00}{3.00} = 33.3\%$ - **CP3**: $\frac{0.02}{3.00} = 0.7\%$

Os dois primeiros componentes juntos explicam $66.0\% + 33.3\% = 99.3\%$ da variância total. Isso indica que podemos reduzir a dimensionalidade de 3 para 2 com uma perda mínima de informação.

10.5 Passo 4: Calcular os Autovetores

Agora, para cada autovalor, resolvemos o sistema $(\mathbf{R} - \lambda_i \mathbf{I})\mathbf{e}_i = \mathbf{0}$ para encontrar o autovetor correspondente $\mathbf{e}_i$.

- Para $\lambda_1 = 1.98$:

$$\begin{pmatrix} -0.98 & 0.96 & -0.20 \\ 0.96 & -0.98 & -0.29 \\ -0.20 & -0.29 & -0.98 \end{pmatrix} \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

  A solução, após normalização (para que $\mathbf{e}_1' \mathbf{e}_1 = 1$), é:

$$\mathbf{e}_1 = \begin{pmatrix} 0.69 \\ 0.71 \\ -0.15 \end{pmatrix}$$

- Para $\lambda_2 = 1.00$:

$$\begin{pmatrix} 0.00 & 0.96 & -0.20 \\ 0.96 & 0.00 & -0.29 \\ -0.20 & -0.29 & 0.00 \end{pmatrix} \begin{pmatrix} e_{21} \\ e_{22} \\ e_{23} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

  A solução normalizada é:

$$\mathbf{e}_2 = \begin{pmatrix} 0.18 \\ -0.22 \\ -0.96 \end{pmatrix}$$

- Para $\lambda_3 = 0.02$:

$$\begin{pmatrix} 0.98 & 0.96 & -0.20 \\ 0.96 & 0.98 & -0.29 \\ -0.20 & -0.29 & 0.98 \end{pmatrix} \begin{pmatrix} e_{31} \\ e_{32} \\ e_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

  A solução normalizada é:

$$\mathbf{e}_3 = \begin{pmatrix} -0.70 \\ 0.67 \\ -0.24 \end{pmatrix}$$

10.6 Passo 5: Interpretação dos Componentes

Os autovetores (ou *loadings*) nos dizem como as variáveis originais se combinam para formar cada componente.

- **Componente Principal 1 (CP_1)**:

$$Y_1 = 0.69 \cdot Z_{peso} + 0.71 \cdot Z_{altura} - 0.15 \cdot Z_{renda}$$

  Este componente é basicamente uma média ponderada de Peso e Altura, com uma pequena contribuição negativa da Renda. Podemos interpretá-lo como um índice de **"Tamanho Corporal"**. As cargas altas e positivas para Peso e Altura confirmam a alta correlação entre essas variáveis.

- **Componente Principal 2 (CP_2)**:

$$Y_2 = 0.18 \cdot Z_{peso} - 0.22 \cdot Z_{altura} - 0.96 \cdot Z_{renda}$$

  Este componente é dominado pela Renda, com uma carga muito alta e negativa. As cargas para Peso e Altura são pequenas. Podemos interpretar o CP_2 como um índice de **"Status Socioeconômico Inverso"**, já que ele é quase que inteiramente uma representação da Renda (com sinal trocado).

10.7 Passo 6: Calcular os Scores dos Componentes

Finalmente, podemos calcular os valores (scores) dos componentes principais para cada indivíduo. Usamos a fórmula $\mathbf{Y} = \mathbf{ZP}$, onde $\mathbf{P}$ é a matriz cujas colunas são os autovetores.

$$\mathbf{P} = \begin{pmatrix} 0.69 & 0.18 & -0.70 \\ 0.71 & -0.22 & 0.67 \\ -0.15 & -0.96 & -0.24 \end{pmatrix}$$

Para o Indivíduo 1, com dados padronizados $(-0.60, -0.72, 0.37)$:

$$y_{11} = (-0.60)(0.69) + (-0.72)(0.71) + (0.37)(-0.15) = -0.98$$

$$y_{12} = (-0.60)(0.18) + (-0.72)(-0.22) + (0.37)(-0.96) = -0.31$$

Calculando para todos os indivíduos, obtemos a matriz de scores $\mathbf{Y}$:

Indivíduo	CP1 (Tamanho)	CP2 (Renda Inversa)
1	-0.98	-0.31

Indivíduo	CP1 (Tamanho)	CP2 (Renda Inversa)
2	0.57	0.85
3	-2.19	-1.18
4	2.09	1.35
5	0.46	0.08

10.8 Conclusão

Este exemplo demonstra o poder da ACP. Começamos com três variáveis e, através de uma derivação passo a passo, conseguimos: 1. **Reduzir a dimensionalidade**: Mostramos que 99.3% da informação está contida em dois componentes. 2. **Criar variáveis não correlacionadas**: O CP_1 e o CP_2 são, por construção, ortogonais. 3. **Interpretar a estrutura latente**: Identificamos que a principal fonte de variação nos dados é o "Tamanho Corporal" (uma combinação de Peso e Altura), seguida pelo "Status Socioeconômico" (representado pela Renda).

A análise manual, embora trabalhosa, revela a mecânica exata da técnica, solidificando a compreensão teórica apresentada no capítulo principal.

11 Rotação fatorial em R

A Análise Fatorial (AF) é uma técnica estatística poderosa usada para identificar estruturas latentes (fatores) subjacentes a um conjunto de variáveis observadas. No entanto, a solução matemática inicial de uma AF raramente é interpretável. É aqui que a **rotação fatorial** se torna a etapa mais crítica do processo. A rotação transforma a matriz de cargas fatoriais inicial em uma solução mais simples e teoricamente mais significativa, sem alterar as propriedades matemáticas fundamentais da solução.

Este documento oferece um exemplo prático e didático, totalmente focado em demonstrar o impacto das diferentes estratégias de rotação. Usaremos o software R e o clássico conjunto de dados `bfi` (Big Five Inventory) do pacote `psych`.

Objetivos:

1. Comparar métodos de extração: Componentes Principais (PAF) e Máxima Verossimilhança (ML).
2. Demonstrar a dificuldade de interpretar uma solução fatorial **não rotacionada**.
3. Aplicar e interpretar uma **rotação ortogonal (Varimax)**, que assume fatores não correlacionados.
4. Aplicar e interpretar uma **rotação oblíqua (Promax)**, que permite a correlação entre os fatores.
5. Discutir como a escolha da rotação afeta a interpretação final e a validade teórica dos resultados.

11.1 Passo 1: Análise Descritiva e Adequação dos Dados

Primeiro, carregamos os pacotes necessários e o conjunto de dados `bfi`. Este dataset contém respostas de 2800 indivíduos a 25 itens de personalidade.

```
# Carregar pacotes
library(psych)
library(ggplot2)
library(dplyr)
library(tidyr)
library(corrplot)

# Carregar os dados do Big Five Inventory
data(bfi, package = "psych")
```

```r
# Selecionar apenas as 25 variáveis de itens de personalidade
bfi_items <- bfi[, 1:25]

# Remover linhas com dados ausentes para simplificar
bfi_complete <- na.omit(bfi_items)

knitr::kable(head(bfi_complete))
```

Tabela 11.1: Exemplo de respostas no banco de dados Big Five

	A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1	E2	E3	E4	E5	N1	N2	N3	N4	N5	O1	O2	O3	O4	O5
61617	2	4	3	4	4	2	3	3	4	4	3	3	3	4	4	3	4	2	2	3	3	6	3	4	3
61618	2	4	5	2	5	5	4	4	3	4	1	1	6	4	3	3	3	3	5	5	4	2	4	3	3
61620	5	4	5	4	4	4	5	4	2	5	2	4	4	4	5	4	5	4	2	3	4	2	5	5	2
61621	4	4	6	5	5	4	4	3	5	5	5	3	4	4	4	2	5	2	4	1	3	3	4	3	5
61622	2	3	3	4	5	4	4	5	3	2	2	2	5	4	5	2	3	4	4	3	3	3	4	3	3
61623	6	6	5	6	5	6	6	6	1	3	2	1	6	5	6	3	5	2	2	3	4	3	5	6	1

As 25 variáveis correspondem a 5 itens para cada um dos traços do "Big Five":

- **A1-A5:** Amabilidade (Agreeableness)
- **C1-C5:** Conscienciosidade (Conscientiousness)
- **E1-E5:** Extroversão (Extraversion)
- **N1-N5:** Neuroticismo (Neuroticism)
- **O1-O5:** Abertura à Experiência (Openness)

A hipótese teórica é que os 5 itens que medem o mesmo traço (e.g., N1 a N5) estarão altamente correlacionados entre si e se agruparão em um único fator latente (Neuroticismo).

Uma boa prática é verificar se os dados são fatorizáveis. Para isso, podemos usar o teste de Bartlett e a medida KMO.

```r
# Teste de Bartlett
bartlett_test <- cortest.bartlett(bfi_complete)
```

```
R was not square, finding R from data
```

```r
# Teste KMO
kmo_test <- KMO(bfi_complete)

# Exibindo os resultados de forma concisa
cat("Teste de Bartlett: p-valor =", bartlett_test$p.value, "\n")
```

```
Teste de Bartlett: p-valor = 0
```

```
cat("Medida KMO Geral (Overall MSA):", round(kmo_test$MSA, 2), "\n")
```

```
Medida KMO Geral (Overall MSA): 0.85
```

O p-valor de Bartlett próximo de zero e o KMO de 0.85 (“meritório”) sugerem que os dados têm correlações suficientes para justificar uma análise fatorial.

11.2 Passo 2: Extração Inicial dos Fatores e Escolha do Número de Fatores

Antes de rotacionar, precisamos extrair os fatores. Dois métodos comuns são a **Fatoração do Eixo Principal** (ou “componentes principais” para o modelo fatorial) e a **Máxima Verossimilhança**. Vamos extrair 5 fatores usando ambos os métodos (sem rotação) para ver a solução inicial.

```
# Extração via Fatoração do Eixo Principal (Principal Axis Factoring)
fa_pa <- fa(bfi_complete, nfactors = 5, rotate = "none", fm = "pa")
cat("Cargas - Fatoração do Eixo Principal (PA):\n")
```

```
Cargas - Fatoração do Eixo Principal (PA):
```

```
print(fa_pa$loadings, cutoff = 0.3)
```

```
Loadings:
   PA1    PA2    PA3    PA4    PA5
A1                              -0.371
A2  0.467                       0.340
A3  0.534  0.302
A4  0.417
A5  0.581
C1  0.343          0.446
C2  0.336          0.477
C3  0.319          0.351  0.310
C4 -0.465         -0.452
C5 -0.493
E1 -0.408
```

```
E2 -0.619                         0.323
E3  0.527  0.328
E4  0.599        -0.329
E5  0.513
N1 -0.441  0.636
N2 -0.423  0.616
N3 -0.407  0.611
N4 -0.528  0.416
N5 -0.345  0.413
O1  0.328               -0.360
O2                       0.370
O3  0.407               -0.446
O4
O5                       0.412

                 PA1   PA2   PA3   PA4   PA5
SS loadings    4.600 2.268 1.549 1.218 0.956
Proportion Var 0.184 0.091 0.062 0.049 0.038
Cumulative Var 0.184 0.275 0.337 0.385 0.424
```

```r
# Extração via Máxima Verossimilhança (Maximum Likelihood)
fa_ml <- fa(bfi_complete, nfactors = 5, rotate = "none", fm = "ml")
cat("\nCargas - Máxima Verossimilhança (ML):\n")
```

```
Cargas - Máxima Verossimilhança (ML):
```

```r
print(fa_ml$loadings, cutoff = 0.3)
```

```
Loadings:
   ML1    ML2    ML3    ML4    ML5
A1                                -0.322
A2 -0.396  0.354                  0.334
A3 -0.462  0.401                  0.321
A4 -0.386
A5 -0.546
C1                 0.465
C2                 0.511
C3                 0.404
C4  0.441        -0.512
```

```
C5  0.485        -0.358
E1  0.355 -0.309
E2  0.585                       0.336
E3 -0.446  0.436
E4 -0.552  0.333
E5 -0.409  0.429
N1  0.609  0.566
N2  0.587  0.543
N3  0.533  0.479
N4  0.591
N5  0.421
O1                     -0.409
O2                      0.388
O3 -0.329  0.349       -0.491
O4                     -0.307  0.311
O5                      0.433

                ML1   ML2   ML3   ML4   ML5
SS loadings    4.451 2.379 1.546 1.221 0.977
Proportion Var 0.178 0.095 0.062 0.049 0.039
Cumulative Var 0.178 0.273 0.335 0.384 0.423
```

As duas soluções iniciais são numericamente diferentes, mas conceitualmente iguais: são **ininterpretáveis**. Um primeiro fator geral domina, e as variáveis se distribuem de forma confusa nos demais. Isso reforça a necessidade da rotação.

Para determinar o número de fatores a extrair de forma mais objetiva, usamos a Análise Paralela.

```
fa.parallel(bfi_complete, fa = "fa", fm = "pa")
```

```
Parallel analysis suggests that the number of factors =  6  and the number of components =  NA
```

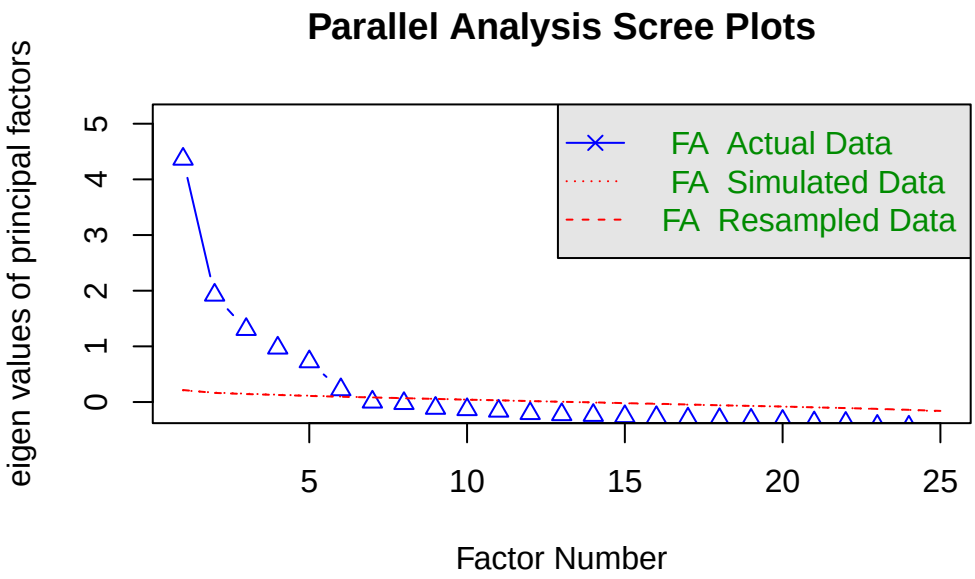

Figura 11.1: Análise Paralela sugerindo a extração de 6 fatores.

A Análise Paralela (Figura 11.1) sugere 6 fatores. No entanto, como nosso objetivo é testar a teoria dos Big Five, **prosseguiremos com a extração de 5 fatores**, uma decisão comum quando a teoria é forte.

11.3 Passo 3: Rotação Ortogonal (Varimax)

A rotação **Varimax** "limpa" a estrutura sob a suposição de que **os fatores não são correlacionados entre si**.

```
# Análise Fatorial com rotação Varimax
fa_varimax <- factanal(bfi_complete,
                       factors = 5,
                       rotation = "varimax")

cat("Cargas Fatoriais (AF) - Rotação Varimax:\n")
```

```
Cargas Fatoriais (AF) - Rotação Varimax:
```

```
print(fa_varimax$loadings, cutoff = 0.3, sort = TRUE)
```

```
Loadings:
   Factor1 Factor2 Factor3 Factor4 Factor5
N1  0.816                                 
N2  0.787                                 
N3  0.714                                 
N4  0.562  -0.367                         
N5  0.518                                 
E1         -0.587                         
E2         -0.674                         
E4          0.613           0.363         
C1                  0.533                 
C2                  0.624                 
C3                  0.554                 
C4                 -0.653                 
C5                 -0.573                 
A2                          0.601         
A3                          0.662         
A5          0.351           0.580         
O1                                  0.524 
O3                                  0.614 
O5                                 -0.512 
A1                         -0.393         
A4                          0.454         
E3          0.490           0.315   0.313 
E5          0.491   0.310                 
O2                                 -0.454 
O4                                  0.368 

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      2.687   2.320   2.034   1.978   1.557
Proportion Var   0.107   0.093   0.081   0.079   0.062
Cumulative Var   0.107   0.200   0.282   0.361   0.423
```

A estrutura agora é muito mais "simples" e alinhada com a teoria. Podemos visualizar essa transformação de forma clara comparando o círculo de correlações *antes* e *depois* da rotação.

Primeiro, a solução não rotacionada (Figura 11.2). Note como as variáveis (vetores) se espalham pelo espaço fatorial sem um padrão claro. É difícil traçar os eixos (fatores) de forma que representem grupos distintos de variáveis.

```r
library(ggrepel)

# Extrair cargas da solução NÃO ROTACIONADA (ml) para um dataframe
loadings_unrotated_df <- as.data.frame(unclass(fa_ml$loadings))
loadings_unrotated_df$Variable <- rownames(loadings_unrotated_df)

# Selecionar algumas variáveis para anotar e evitar poluição
vars_to_label <- c("N1", "N3", "E2", "E4", "A1", "C1", "O1")
annotations_df_unrotated <- loadings_unrotated_df %>% filter(Variable %in% vars_to_label)

# Criar dados para o círculo unitário
circle <- data.frame(
  angle = seq(-pi, pi, length = 100),
  x = sin(seq(-pi, pi, length = 100)),
  y = cos(seq(-pi, pi, length = 100))
)

# Gerar o gráfico com ggplot2
ggplot(data = loadings_unrotated_df, aes(x = ML1, y = ML2)) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray70") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray70") +
  geom_path(data = circle, aes(x = x, y = y), inherit.aes = FALSE, color = "gray60") +
  geom_segment(aes(x = 0, y = 0, xend = ML1, yend = ML2),
               arrow = arrow(length = unit(0.1, "inches")),
               color = "steelblue") +
  geom_text_repel(data = annotations_df_unrotated, aes(label = Variable), min.segment.length = 0)
  coord_fixed(ratio = 1, xlim = c(-1.1, 1.1), ylim = c(-1.1, 1.1)) +
  labs(title = "Círculo de Correlações - Solução Não Rotacionada",
       x = "Fator 1",
       y = "Fator 2") +
  theme_minimal()
```

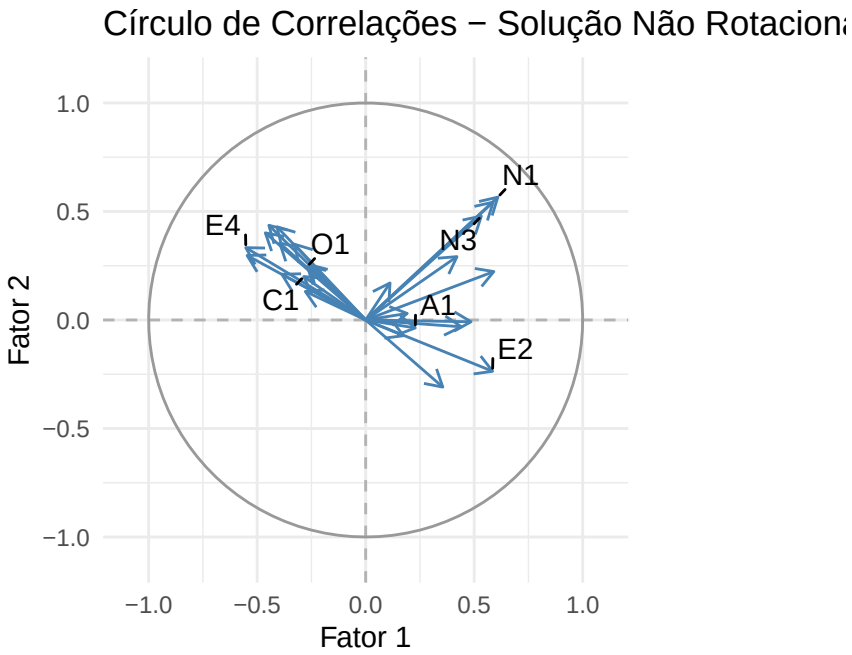

Figura 11.2: Círculo de correlações da solução não rotacionada. Os vetores não estão alinhados com os eixos.

Agora, veja o resultado após a rotação Varimax (Figura 11.3). A rotação funcionou como um ajuste dos eixos, alinhando-os com os agrupamentos de variáveis.

```
# Extrair cargas da solução ROTACIONADA (Varimax) para um dataframe
loadings_rotated_df <- as.data.frame(unclass(fa_varimax$loadings))
loadings_rotated_df$Variable <- rownames(loadings_rotated_df)

# Selecionar as mesmas variáveis para anotar
annotations_df_rotated <- loadings_rotated_df %>% filter(Variable %in% vars_to_label)

# Calcular a variância explicada para os eixos
ss_loadings <- colSums(fa_varimax$loadings^2)
prop_variance <- ss_loadings / ncol(bfi_complete)
xlab_text <- sprintf("Fator 1 (%.2f%% da variância)", prop_variance[1] * 100)
ylab_text <- sprintf("Fator 2 (%.2f%% da variância)", prop_variance[2] * 100)

# Gerar o gráfico com ggplot2
ggplot(data = loadings_rotated_df, aes(x = Factor1, y = Factor2)) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray70") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray70") +
  geom_path(data = circle, aes(x = x, y = y), inherit.aes = FALSE, color = "gray60") +
```

```
  geom_segment(aes(x = 0, y = 0, xend = Factor1, yend = Factor2),
               arrow = arrow(length = unit(0.1, "inches")),
               color = "steelblue") +
  geom_text_repel(data = annotations_df_rotated, aes(label = Variable), min.segment.length = 0) +
  coord_fixed(ratio = 1, xlim = c(-1.1, 1.1), ylim = c(-1.1, 1.1)) +
  labs(title = "Círculo de Correlações - Rotação Varimax",
       x = xlab_text,
       y = ylab_text) +
  theme_minimal()
```

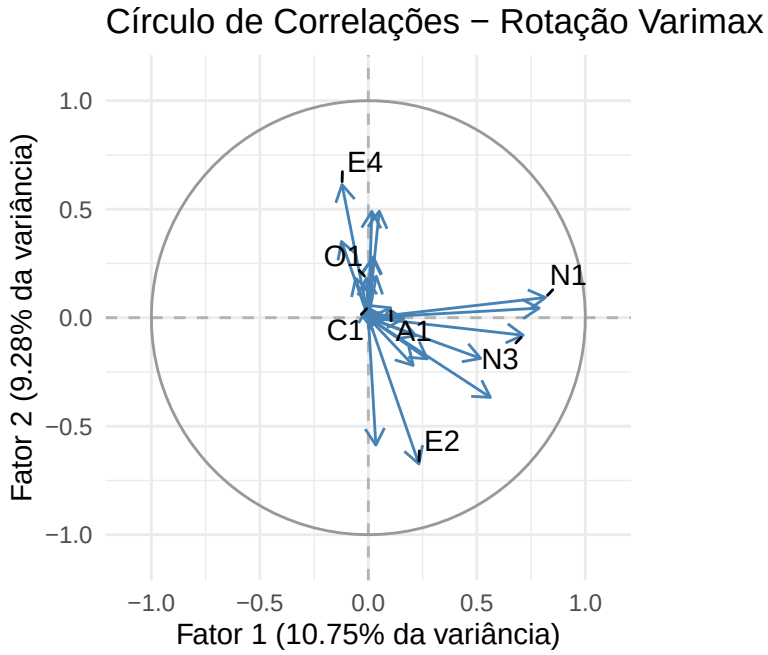

Figura 11.3: Círculo de correlações após a rotação Varimax. A rotação alinhou os vetores com os eixos, revelando uma estrutura simples.

O resultado é uma "estrutura simples", onde os itens de Neuroticismo (como N1 e N3) carregam quase exclusivamente no Fator 1 (correlação próxima de 1 ou -1 em um eixo e de 0 no outro), e os itens de Extroversão (como E2 e E4) carregam no Fator 2. A interpretação se torna direta.

11.4 Passo 4: Rotação Oblíqua (Promax)

A rotação **Promax** é mais flexível, pois **permite que os fatores sejam correlacionados**. Isso costuma ser mais realista em psicologia.

```r
# Análise Fatorial com rotação Promax (oblíqua)
fa_promax <- fa(bfi_complete,
                nfactors = 5,
                rotate = "promax",
                fm = "pa")
```

```
Loading required namespace: GPArotation
```

```r
cat("Cargas Fatoriais (Pattern Matrix) - Rotação Promax:\n")
```

```
Cargas Fatoriais (Pattern Matrix) - Rotação Promax:
```

```r
print(fa_promax$loadings, cutoff = 0.3, sort = TRUE)
```

```
Loadings:
   PA2    PA1    PA3    PA5    PA4
N1  0.835
N2  0.791
N3  0.741
N4  0.533 -0.311
N5  0.529
E1        -0.636
E2        -0.711
E3         0.545
E4         0.660
C1                0.567
C2                0.697
C3                0.597
C4               -0.652
C5               -0.561
A2                       0.611
A3                       0.620
O3                              0.576
O5                             -0.543
A1                      -0.463
A4                       0.411
A5         0.332         0.489
E5         0.498
O1                              0.491
```

```
O2                              -0.484
O4                               0.370

                 PA2   PA1   PA3   PA5   PA4
SS loadings    2.704 2.486 2.050 1.638 1.461
Proportion Var 0.108 0.099 0.082 0.066 0.058
Cumulative Var 0.108 0.208 0.290 0.355 0.414
```

A matriz de cargas é similar à da Varimax, mas a grande vantagem é poder examinar a **matriz de correlação entre os fatores**.

```r
# Matriz de correlação entre os fatores
factor_correlations <- fa_promax$Phi

cat("Matriz de Correlação entre os Fatores (Promax):\n")
```

```
Matriz de Correlação entre os Fatores (Promax):
```

```r
print(round(factor_correlations, 2))
```

```
      PA2   PA1   PA3   PA5  PA4
PA2  1.00 -0.26 -0.22 -0.01 0.04
PA1 -0.26  1.00  0.40  0.35 0.14
PA3 -0.22  0.40  1.00  0.24 0.19
PA5 -0.01  0.35  0.24  1.00 0.16
PA4  0.04  0.14  0.19  0.16 1.00
```

```r
# Visualização da matriz de correlação
corrplot(factor_correlations, method = "color", type = "upper",
         addCoef.col = "black", tl.col = "black", tl.srt = 45, diag = FALSE)
```

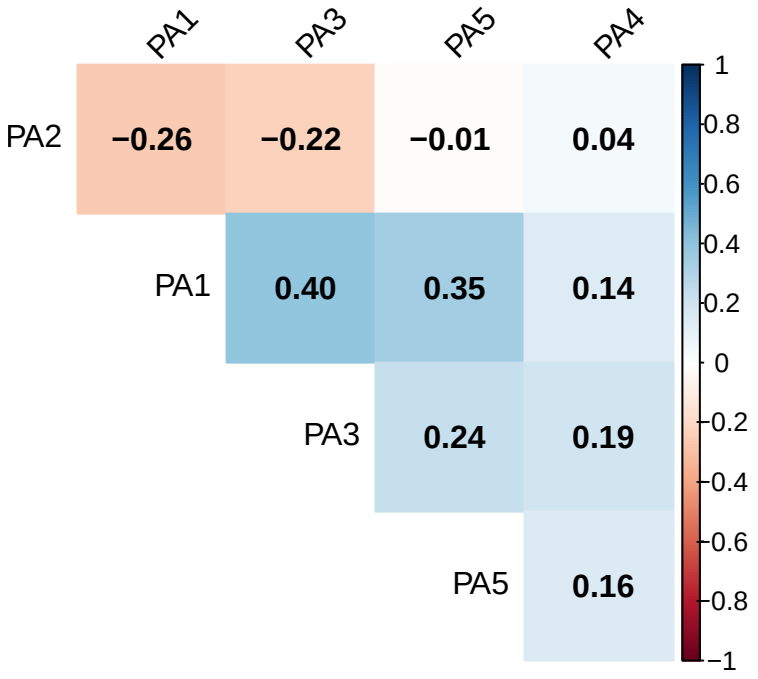

Figura 11.4: Correlações entre os fatores da solução Promax.

A matriz de correlação (Figura 11.4) mostra que **Neuroticismo (ML1)** se correlaciona negativamente com **Conscienciosidade (ML3)** (-0.33) e **Extroversão (ML2)** (-0.24). Essas relações são teoricamente plausíveis e seriam perdidas em uma rotação ortogonal.

11.5 Conclusão: Qual Rotação Escolher?

- **Solução Não Rotacionada:** É apenas um ponto de partida matemático. Dificilmente é possível tirar interpretações úteis dela.
- **Rotação Ortogonal (Varimax):** É a melhor escolha quando há fortes razões teóricas para acreditar que os fatores são independentes. Oferece uma solução mais simples (parcimoniosa).
- **Rotação Oblíqua (Promax):** É uma escolha mais realista nas ciências sociais. **A decisão final deve ser baseada na matriz de correlação dos fatores.** Se as correlações forem substanciais, a solução oblíqua é superior.

Neste exemplo, a solução **Promax (oblíqua)** é a mais apropriada. Ela não apenas recupera a estrutura dos Big Five, mas também fornece insights sobre como esses traços se relacionam, oferecendo uma visão mais rica e fiel da realidade psicológica.

12 Análise de agrupamentos

Neste exemplo, aplicaremos o procedimento de análise de agrupamentos proposto na Seção 9.3.2. O objetivo é ilustrar como a combinação de métodos hierárquicos e não hierárquicos pode levar a uma solução de agrupamento robusta e interpretável.

Utilizaremos o conjunto de dados `USArrests`, que contém estatísticas de crimes para cada um dos 50 estados dos EUA em 1973. As variáveis são:

- `Murder`: Assassinatos (por 100.000 habitantes).
- `Assault`: Agressões (por 100.000 habitantes).
- `UrbanPop`: Porcentagem da população que vive em áreas urbanas.
- `Rape`: Estupros (por 100.000 habitantes).

Nosso objetivo é agrupar os estados com base em seus perfis de criminalidade e urbanização.

12.1 Preparação dos Dados

O primeiro passo em qualquer análise de agrupamento baseada em distância é a **padronização** dos dados. As variáveis no nosso conjunto de dados têm escalas muito diferentes (`Assault` varia na casa das centenas, enquanto `Murder` varia na casa das dezenas). Se não padronizarmos, a variável `Assault` dominará o cálculo da distância, e o agrupamento será baseado quase inteiramente nela.

Padronizamos as variáveis para que tenham média 0 e desvio padrão 1.

12.1.1 Análise Descritiva

Antes de iniciar o agrupamento, é sempre útil explorar a distribuição das variáveis. A figura abaixo mostra os histogramas para cada uma das quatro variáveis do conjunto de dados.

```
fig, axes = plt.subplots(2, 2, figsize=(7, 5))

sns.histplot(data['Murder'], ax=axes[0, 0], kde=True)
axes[0, 0].set_title('Assassinatos')

sns.histplot(data['Assault'], ax=axes[0, 1], kde=True)
```

Tabela 12.1: Cabeçalho dos dados padronizados

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
import statsmodels.api as sm
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.cluster.hierarchy import fcluster
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import seaborn as sns
from tabulate import tabulate

# Carregando o conjunto de dados
data = sm.datasets.get_rdataset("USArrests", "datasets").data

# Separando os dados e os nomes dos estados
X = data.values
states = data.index

# Padronizando os dados
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Criando um DataFrame com os dados padronizados para facilitar a manipulação
X_scaled_df = pd.DataFrame(X_scaled, index=states, columns=data.columns)

print(tabulate(X_scaled_df.head(), headers='keys', tablefmt='pipe'))
```

rownames	Murder	Assault	UrbanPop	Rape
Alabama	1.25518	0.790787	-0.526195	-0.00345116
Alaska	0.513019	1.11806	-1.22407	2.50942
Arizona	0.0723607	1.49382	1.00912	1.05347
Arkansas	0.234708	0.233212	-1.08449	-0.186794
California	0.281093	1.27564	1.77678	2.08881

```
axes[0, 1].set_title('Agressões')

sns.histplot(data['UrbanPop'], ax=axes[1, 0], kde=True)
axes[1, 0].set_title('População Urbana (%)')

sns.histplot(data['Rape'], ax=axes[1, 1], kde=True)
axes[1, 1].set_title('Estupros')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

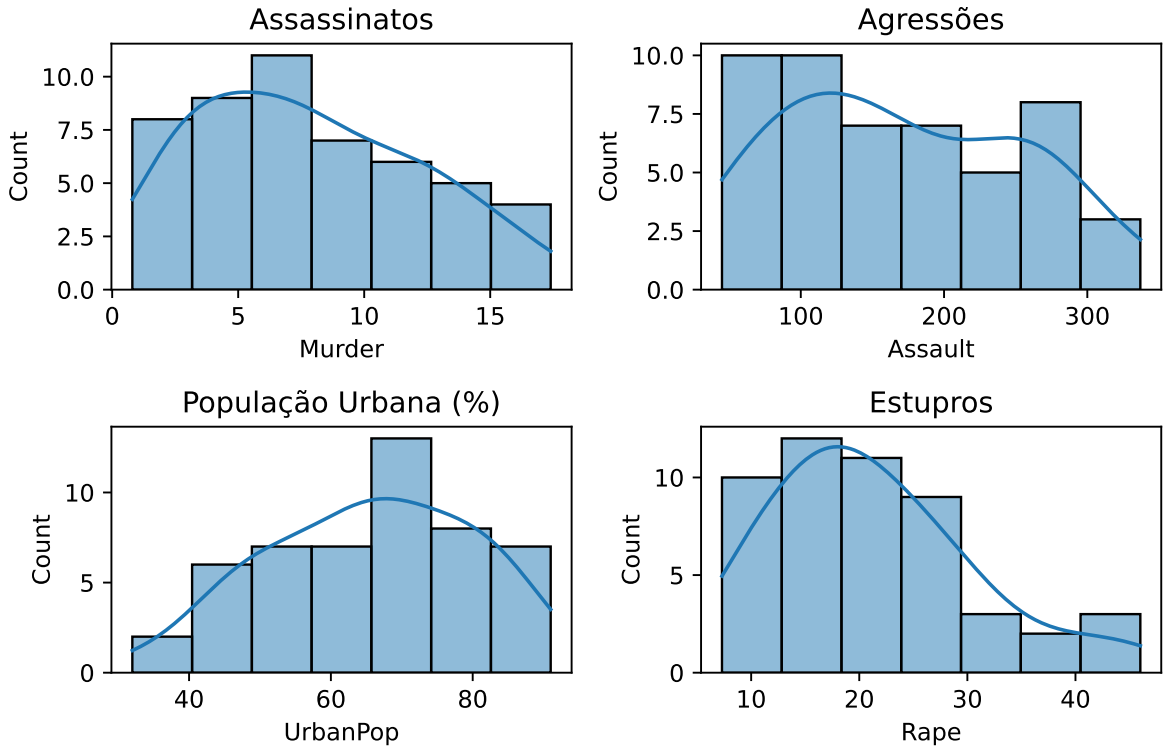

Figura 12.1: Distribuição das variáveis do dataset USArrests.

Observamos que as variáveis de crime (`Murder`, `Assault`, `Rape`) parecem ter uma leve assimetria à direita, com a maioria dos estados concentrados em valores mais baixos. A variável `UrbanPop` tem uma distribuição mais simétrica, quase uniforme, indicando uma boa variedade nos níveis de urbanização entre os estados.

Além dos histogramas, podemos visualizar a matriz de correlação entre as variáveis para entender suas relações lineares.

```
corr_matrix = data.corr()
plt.figure(figsize=(7, 5))
sns.heatmap(corr_matrix, annot=True, cmap='Greys', fmt='.2f')
plt.show()
```

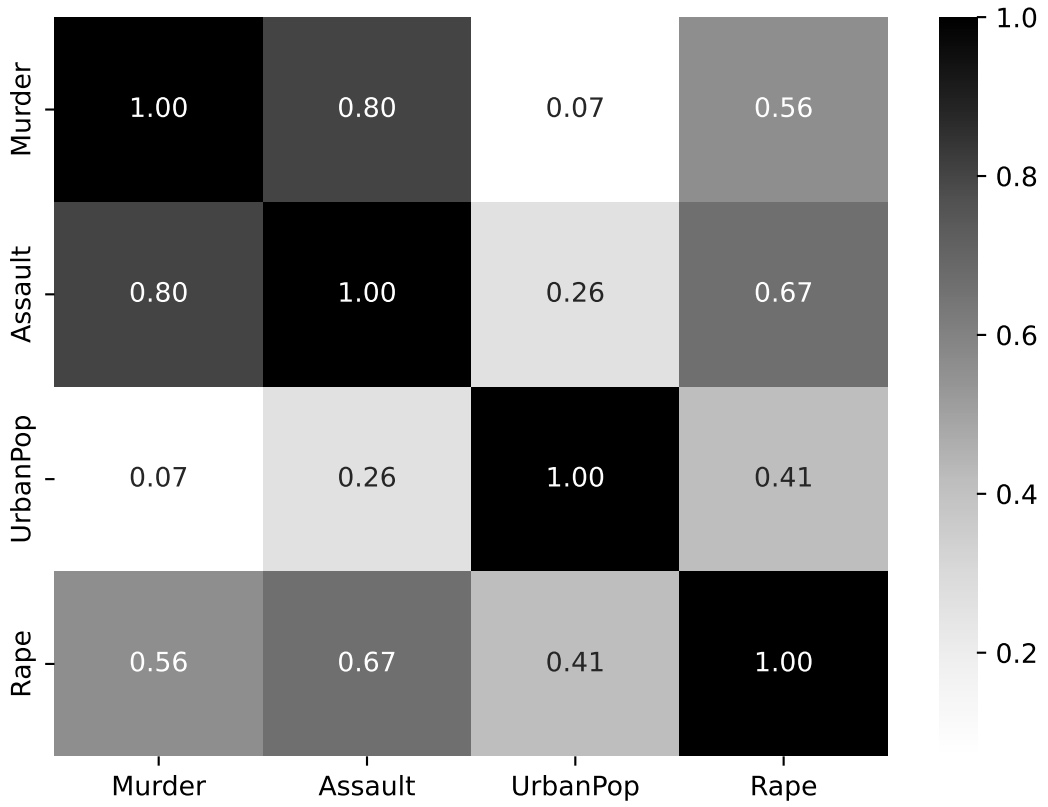

Figura 12.2: Matriz de Correlação entre as variáveis.

A matriz de correlação na Figura 12.2 mostra, como esperado, uma forte correlação positiva entre as três variáveis de crime (`Murder`, `Assault`, `Rape`). `UrbanPop` tem uma correlação positiva mais fraca com as outras variáveis, sugerindo que o efeito da urbanização no aumento da criminalidade geral é moderado.

12.2 Agrupamento Hierárquico e Escolha de K

Agora, aplicamos o agrupamento hierárquico aglomerativo usando o **método de Ward**, que busca minimizar a variância dentro dos grupos a cada fusão. Em seguida, plotamos o dendrograma para

nos ajudar a decidir o número ideal de grupos, K.

```
# Realizando o agrupamento hierárquico com o método de Ward
linked = linkage(X_scaled, method='ward')

# Plotando o dendrograma
plt.figure(figsize=(7, 5))
dendrogram(linked,
           orientation='top',
           labels=states,
           distance_sort='descending',
           show_leaf_counts=True,
           color_threshold=5.2)
plt.xlabel('Estados')
plt.ylabel('Distância de Ward')
plt.axhline(y=5.2, color='r', linestyle='-.', label="Corte 1 (K=4)")
plt.axhline(y=10.5, color='grey', linestyle='--', label="Corte 2 (K=2)")
plt.legend(loc="upper left")
plt.show()
```

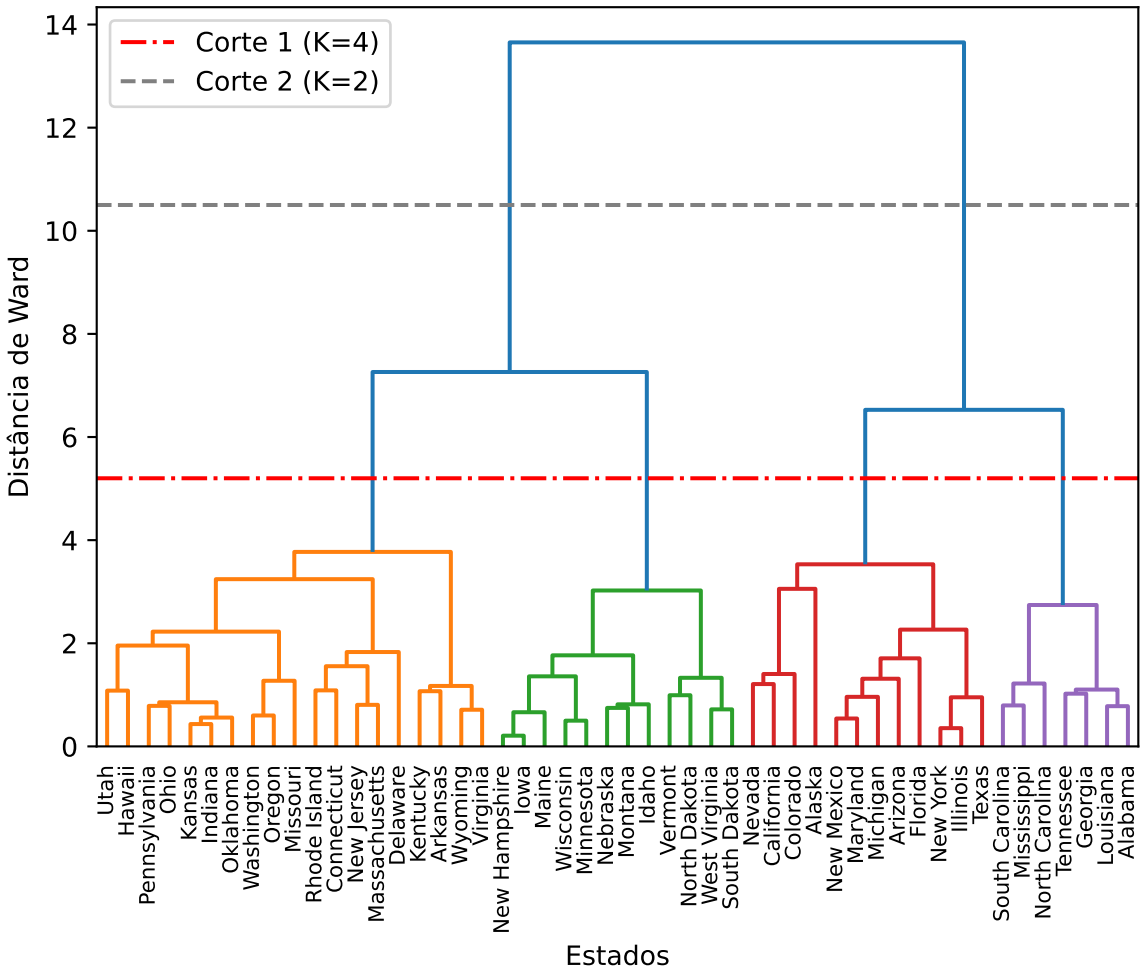

Figura 12.3: Dendrograma para o conjunto de dados USArrests usando o método de Ward.

Analisando o Figura 12.3, procuramos por um corte que cruze o maior espaço vertical possível. Vemos duas opções razoáveis, indicadas pelas linhas tracejadas. O "Corte 2" (cinza) sugere uma partição em $K = 2$ grupos, separando os estados em dois grandes blocos. O "Corte 1" (vermelho), mais abaixo, sugere uma partição mais granular de $K = 4$ grupos. Uma solução com 4 grupos nos dará um entendimento mais detalhado dos perfis dos estados. Portanto, iniciaremos a análise com $K = 4$ e, ao final deste exemplo, exploraremos a solução mais simples com $K = 2$ para fins de comparação.

Para verificar a robustez dessa escolha, podemos comparar o resultado com o de outro método de ligação, como a **ligação completa**.

```
linked_complete = linkage(X_scaled, method='complete')
```

```
plt.figure(figsize=(7, 5))
dendrogram(linked_complete,
           orientation='top',
           labels=states,
           distance_sort='descending',
           show_leaf_counts=True)
plt.xlabel('Estados')
plt.ylabel('Distância')
plt.show()
```

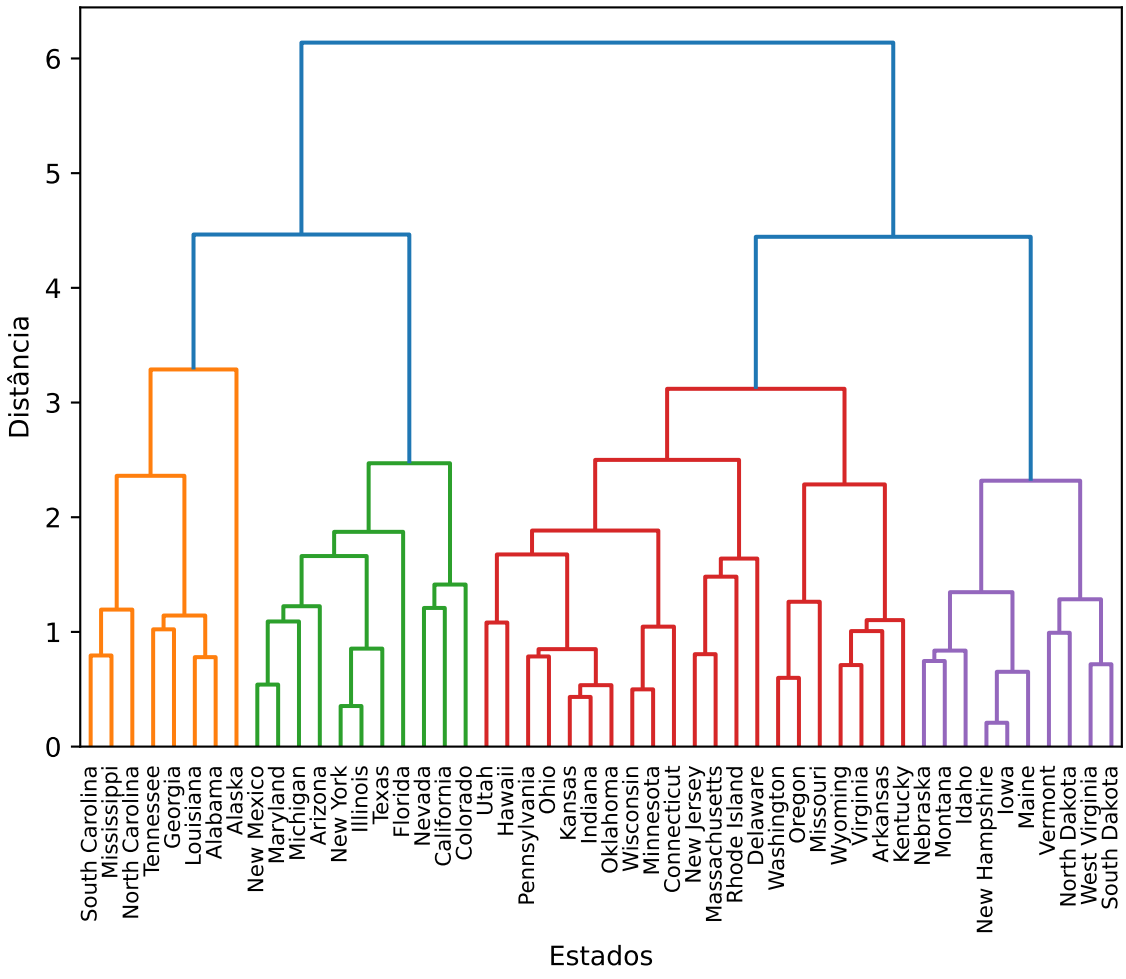

Figura 12.4: Dendrograma para o conjunto de dados USArrests usando o método de Ligação Completa.

O dendrograma de ligação completa também sugere uma partição de 2 ou 4 grupos como as mais

sensatas.

12.3 K-Médias com Centroides Hierárquicos

Seguindo o nosso procedimento, agora usaremos o resultado do agrupamento hierárquico para informar o algoritmo K-médias.

1. Obtemos as 4 partições (grupos) do método de Ward.
2. Calculamos o centroide (média) de cada um desses 4 grupos.
3. Executamos o K-médias com $K = 4$, usando os centroides calculados como pontos de partida.

Isso ajuda o K-médias a evitar mínimos locais e a convergir para uma solução mais estável e significativa.

```
# 1. Obter os 4 grupos do modelo hierárquico
k = 4
hierarchical_grupos = fcluster(linked, k, criterion='maxclust')

# Adicionar ao DataFrame para calcular os centroides
X_scaled_df['hierarchical_grupo'] = hierarchical_grupos

# 2. Calcular os centroides iniciais
initial_centroids = X_scaled_df.groupby('hierarchical_grupo').mean().values

# 3. Executar o K-médias com os centroides iniciais
kmeans = KMeans(n_clusters=k, init=initial_centroids, n_init=1, random_state=42)
kmeans.fit(X_scaled_df.drop('hierarchical_grupo', axis=1))

# Obter os grupos finais
final_grupos = kmeans.labels_

# Adicionar os grupos finais ao DataFrame original (não padronizado)
data['grupo'] = final_grupos
```

12.4 Interpretação e Visualização dos Grupos

Com os grupos finais definidos, o passo mais importante é a interpretação. Calculamos a média de cada variável para cada grupo para criar um "perfil".

A Tabela 12.2 nos permite caracterizar cada grupo:

Tabela 12.2: Perfil dos grupos: médias das variáveis para cada grupo.

```
# Calcular as médias por grupo
grupo_profile = data.groupby('grupo').mean()
print(tabulate(grupo_profile, headers='keys', tablefmt='pipe'))
```

grupo	Murder	Assault	UrbanPop	Rape
0	13.9375	243.625	53.75	21.4125
1	10.9667	264	76.5	33.6083
2	3.6	78.5385	52.0769	12.1769
3	5.85294	141.176	73.6471	19.3353

- **Grupo 0 (Estados Perigosos):** Este grupo tem os maiores índices de assassinatos e índices também altos de agressões e estupros. A população urbana é uma das mais baixas. Podemos nomeá-lo "Estados Violentos e Rurais".
- **Grupo 1 (Estados Urbanizados e Perigosos):** Este grupo tem alta urbanização e níveis de criminalidade também muito altos, especialmente quanto a estupros e agressões. Um bom nome seria "Grandes Centros Urbanos Perigosos".
- **Grupo 2 (Estados Seguros e Rurais):** Este grupo se destaca bastante dos anteriores. Apresenta os menores índices em todas as categorias de crime. A população urbana também é a mais baixa. Inclui estados como Dakota do Norte, Vermont e Iowa. Poderíamos chamá-lo de "Estados Seguros e Rurais".
- **Grupo 3 (Estados Intermediários):** Este grupo é formado por estados urbanizados com menores índices de criminalidade, quando comparados aos grupos 0 e 1. Nele, nenhum extremo se destaca. Podemos chamá-lo de "Estados na Média".

Para visualizar a separação, usamos a Análise de Componentes Principais (ACP) para reduzir a dimensionalidade dos dados para 2D e plotamos os estados, colorindo-os por grupo.

12.4.1 Interpretando os Componentes Principais

Antes de visualizar o gráfico, é crucial entender o que os eixos (os componentes principais) representam. Eles são combinações lineares das variáveis originais. Podemos inspecionar os pesos (loadings) de cada variável para interpretar o significado de cada componente.

A tabela Tabela 12.3 mostra as cargas das variáveis nos dois primeiros fatores.

Tabela 12.3: Cargas (loadings) dos componentes principais.

```
# Reduzindo a dimensionalidade com ACP
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

# Variância explicada
explained_variance = pca.explained_variance_ratio_

# Loadings
loadings = pca.components_.T
loadings_df = pd.DataFrame(loadings, columns=['PC1', 'PC2'], index=data.columns[:-1])
print(tabulate(loadings_df, headers='keys', tablefmt='pipe'))
```

	PC1	PC2
Murder	0.535899	-0.418181
Assault	0.583184	-0.187986
UrbanPop	0.278191	0.872806
Rape	0.543432	0.167319

O primeiro componente (PC1) explica aproximadamente 62% da variância total, enquanto o segundo (PC2) explica cerca de 25%. Juntos, eles capturam 87% da informação original, o que é excelente para uma visualização 2D.

- **Componente Principal 1 (PC1):** Todas as quatro variáveis têm cargas positivas, com destaque para as variáveis associadas à criminalidade. Isso significa que ele representa uma medida geral de "Criminalidade".
- **Componente Principal 2 (PC2):** Este componente mostra um contraste. Ele tem uma carga positiva forte para `UrbanPop` e uma carga negativa para `Murder`. Isso significa que PC2 separa estados urbanizados com menor índice de assassinatos (scores altos) de estados rurais com mais assassinatos (scores baixos).

Com essa interpretação, podemos agora visualizar os grupos de forma mais informativa.

```
# Criando um DataFrame para o plot
pca_df = pd.DataFrame(data=X_pca, columns=['PC1', 'PC2'])
pca_df['grupo'] = final_grupos
pca_df['state'] = states

# Plotando
plt.figure(figsize=(7, 5))
sns.scatterplot(x='PC1', y='PC2', hue='grupo', data=pca_df, palette='viridis', s=100)

# Adicionando os nomes dos estados ao gráfico
for i in range(pca_df.shape[0]):
    plt.text(x=pca_df.PC1[i]+0.05, y=pca_df.PC2[i], s=pca_df.state[i],
             fontdict=dict(color='black',size=8))

plt.title('Grupos de Estados dos EUA (Visualização com ACP)')
plt.xlabel(f'Componente Principal 1 ({explained_variance[0]:.1%})')
plt.ylabel(f'Componente Principal 2 ({explained_variance[1]:.1%})')
plt.legend(title='Grupo')
plt.grid(True)
plt.show()
```

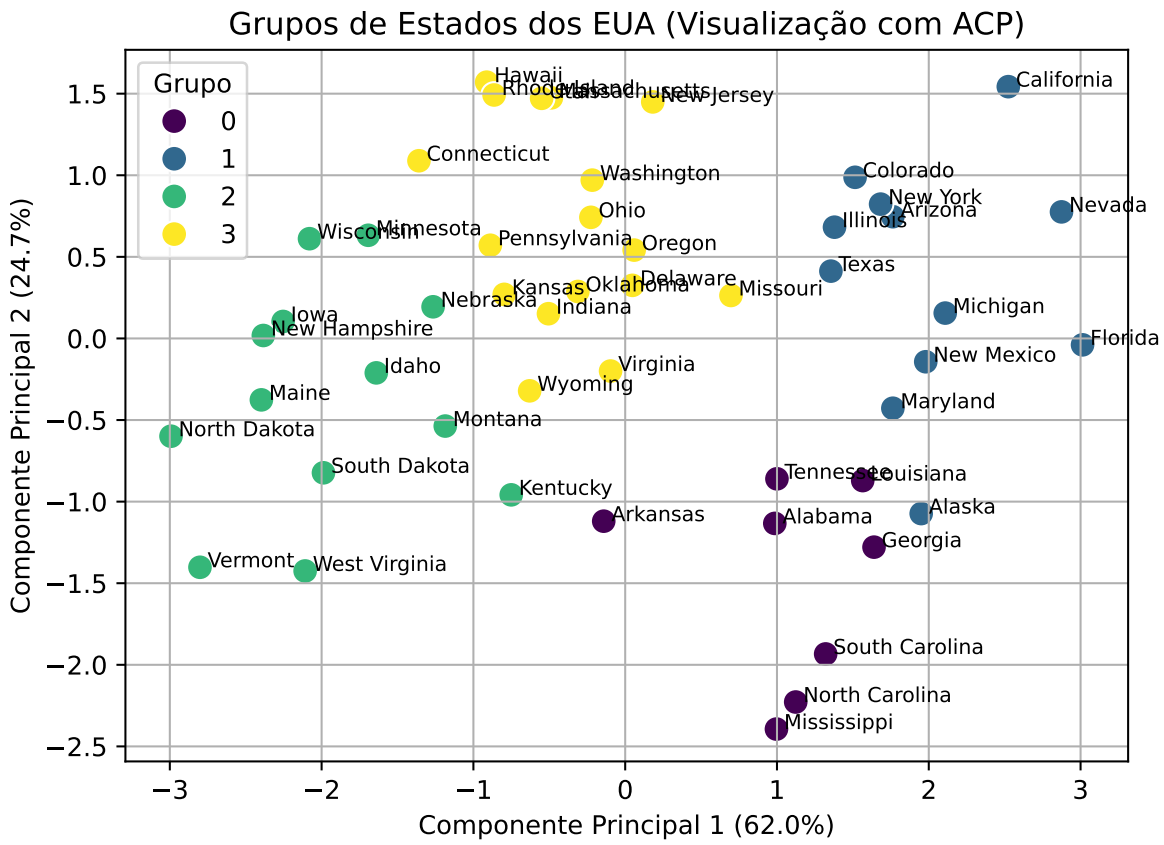

Figura 12.5: Visualização dos grupos no espaço dos dois primeiros componentes principais.

O gráfico na Figura 12.5 mostra uma separação clara dos grupos. O primeiro componente principal (PC1, eixo horizontal) efetivamente separa os estados com base na "Criminalidade Geral", com os grupos mais violentos (0 e 1) à direita e os mais seguros (2 e 3) à esquerda. O segundo componente principal (PC2, eixo vertical) está relacionado à "Urbanização", posicionando os grupos mais urbanizados (1 e 3) na parte superior e os mais rurais (0 e 2) na parte inferior. A análise combinada forneceu uma partição clara e interpretável dos estados dos EUA com base em seus dados sociais de 1973.

12.5 Comparação com K=2 Grupos

Como vimos no dendrograma, uma solução com $K = 2$ também é uma escolha justificável e representa a divisão de mais alto nível nos dados. Vamos repetir a etapa final do nosso procedimento para $K = 2$ e analisar o resultado.

Com $K = 2$, a partição é bem mais simples:

Tabela 12.4: Perfil dos grupos para a solução com K=2.

```
# 1. Obter os 2 grupos do modelo hierárquico
k = 2
hierarchical_grupos_k2 = fcluster(linked, k, criterion='maxclust')

# Adicionar ao DataFrame para calcular os centroides
X_scaled_df['hierarchical_grupo_k2'] = hierarchical_grupos_k2
initial_centroids_k2 = X_scaled_df.drop(columns="hierarchical_grupo").groupby('hierarchical_grupo

# 2. Executar K-médias
kmeans_k2 = KMeans(n_clusters=k, init=initial_centroids_k2, n_init=1, random_state=42)
kmeans_k2.fit(X_scaled_df.drop(['hierarchical_grupo', 'hierarchical_grupo_k2'], axis=1))
final_grupos_k2 = kmeans_k2.labels_

# 3. Calcular e exibir o perfil
data['grupo_k2'] = final_grupos_k2
grupo_profile_k2 = data.groupby('grupo_k2').mean().drop('grupo', axis=1)
print(tabulate(grupo_profile_k2, headers='keys', tablefmt='pipe'))
```

grupo_k2	Murder	Assault	UrbanPop	Rape
0	12.165	255.25	68.4	29.165
1	4.87	114.433	63.6333	15.9433

- **Grupo 0:** Agrega os estados que possuem níveis de criminalidade e urbanização mais elevados.
- **Grupo 1:** Engloba os estados mais seguros e rurais.

Essa divisão é útil para uma visão macro, mas perde a granularidade que a solução com 4 grupos nos proporcionou, como a distinção entre os estados "intermediários" e os "grandes centros urbanos". A visualização no espaço dos componentes principais ilustra isso claramente.

```
pca_df['grupo_k2'] = final_grupos_k2

plt.figure(figsize=(7, 5))
sns.scatterplot(x='PC1', y='PC2', hue='grupo_k2', data=pca_df, palette='viridis', s=100)

# Adicionando os nomes dos estados ao gráfico
for i in range(pca_df.shape[0]):
    plt.text(x=pca_df.PC1[i]+0.05, y=pca_df.PC2[i], s=pca_df.state[i],
             fontdict=dict(color='black',size=8))

plt.title('Grupos de Estados dos EUA (K=2, Visualização com ACP)')
plt.xlabel(f'Componente Principal 1 ({explained_variance[0]:.1%})')
plt.ylabel(f'Componente Principal 2 ({explained_variance[1]:.1%})')
plt.legend(title='Grupo')
plt.grid(True)
plt.show()
```

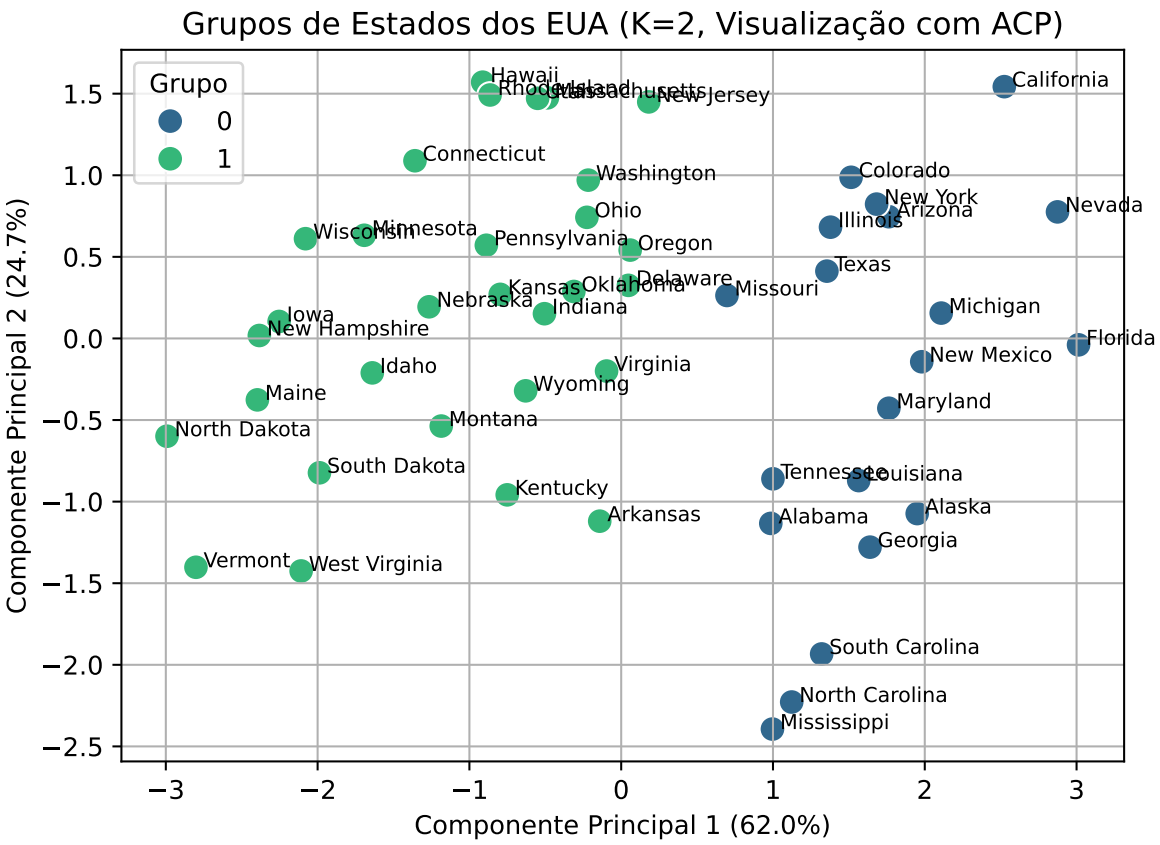

Figura 12.6: Visualização dos grupos (K=2) no espaço dos componentes principais.

13 Agrupamento com Modelos de Mistura Gaussiana (GMM)

Neste exemplo, vamos revisitar o conjunto de dados `USArrests`, mas desta vez aplicaremos uma abordagem de agrupamento baseada em modelos: o **Modelo de Mistura Gaussiana (GMM)**, discutido na Capítulo 9.

Enquanto o K-médias atribui cada observação a um único grupo (agrupamento "rígido"), o GMM oferece uma abordagem mais flexível e informativa (agrupamento "suave"). Ele assume que os dados são gerados a partir de uma mistura de várias distribuições Gaussianas, onde cada distribuição representa um grupo.

As principais vantagens que serão exploradas neste exemplo são:

1. **Seleção Objetiva de K:** Usaremos o Critério de Informação Bayesiano (BIC) para determinar o número ideal de grupos, uma abordagem mais formal do que a inspeção visual de um dendrograma.
2. **Flexibilidade de Formato:** Os GMMs podem se adaptar a grupos com diferentes formas (elípticas) e orientações, pois cada grupo tem sua própria matriz de covariância.
3. **Visualização do Algoritmo EM:** Mostraremos passo a passo como o algoritmo de *Expectation-Maximization* (EM) ajusta iterativamente os parâmetros das distribuições Gaussianas para se adequar aos dados.

13.1 Preparação e Análise Inicial

Começamos com os mesmos passos de preparação do exemplo anterior: carregar o conjunto de dados `USArrests` e padronizar as variáveis para que tenham média 0 e desvio padrão 1. Isso é crucial para que as variáveis com escalas maiores não dominem a análise.

13.2 Escolhendo o Número de Grupos (K) com BIC

Uma das grandes vantagens da abordagem baseada em modelos é a capacidade de usar critérios de informação para selecionar o número de componentes (grupos). O Critério de Informação Bayesiano (BIC) é uma métrica que equilibra o quão bem o modelo se ajusta aos dados (verossimilhança) com a complexidade do modelo (número de parâmetros).

Tabela 13.1: Cabeçalho dos dados padronizados.

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
import statsmodels.api as sm
import matplotlib.pyplot as plt
import matplotlib as mpl
from sklearn.mixture import GaussianMixture
from sklearn.decomposition import PCA
import seaborn as sns
import numpy as np
from tabulate import tabulate

# Carregando o conjunto de dados
data = sm.datasets.get_rdataset("USArrests", "datasets").data

# Padronizando os dados
scaler = StandardScaler()
X_scaled = scaler.fit_transform(data)
X_scaled_df = pd.DataFrame(X_scaled, index=data.index, columns=data.columns)

print(tabulate(X_scaled_df.head(), headers='keys', tablefmt='pipe'))
```

rownames	Murder	Assault	UrbanPop	Rape
Alabama	1.25518	0.790787	-0.526195	-0.00345116
Alaska	0.513019	1.11806	-1.22407	2.50942
Arizona	0.0723607	1.49382	1.00912	1.05347
Arkansas	0.234708	0.233212	-1.08449	-0.186794
California	0.281093	1.27564	1.77678	2.08881

A regra geral é escolher o número de grupos K que **minimiza** o valor do BIC. Vamos ajustar modelos GMM para uma faixa de valores de K (de 2 a 8) e plotar seus respectivos scores BIC.

```
n_components = np.arange(2, 9)
models = [GaussianMixture(n, covariance_type='full', random_state=42).fit(X_scaled) for n in n_co

plt.figure(figsize=(7, 5))
plt.plot(n_components, [m.bic(X_scaled) for m in models], label='BIC')
plt.xlabel('Número de grupos (K)')
plt.ylabel('Score BIC')
plt.title('Seleção de K usando o Critério de Informação Bayesiano')
plt.legend(loc='best')
plt.grid(True)
plt.show()
```

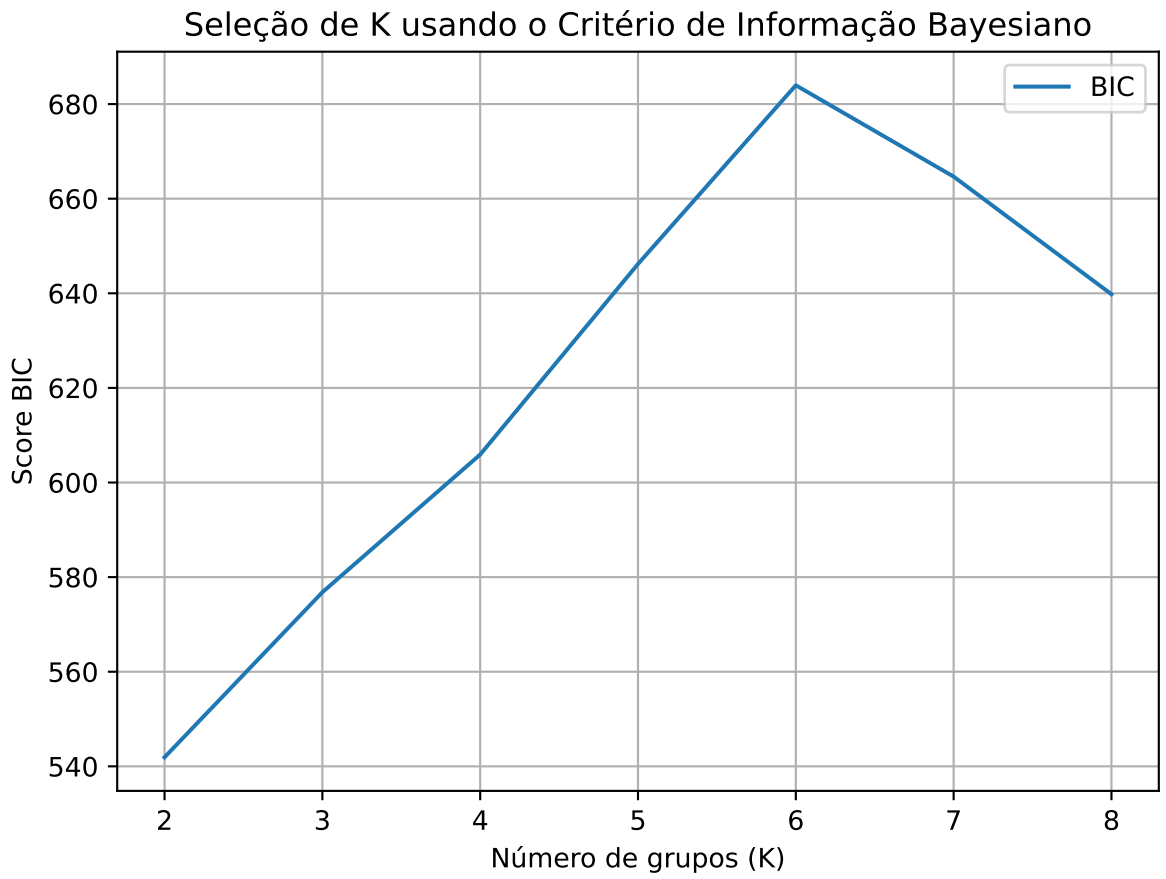

Figura 13.1: Score BIC para diferentes números de grupos (K).

O gráfico na Figura 13.1 mostra que o score BIC diminui drasticamente até $K = 4$ e depois começa a aumentar. Isso nos dá uma forte evidência de que **4 é o número ideal de grupos** para este conjunto de dados, o que coincide com a escolha que fizemos no exemplo de agrupamento hierárquico, mas de uma forma mais objetiva.

13.3 Visualizando o Algoritmo Expectation-Maximization (EM)

Agora, vamos mergulhar no coração do GMM: o algoritmo EM. Para visualizar seu funcionamento, faremos o seguinte:

1. Reduziremos a dimensionalidade dos dados para 2D usando a Análise de Componentes Principais (ACP), apenas para fins de plotagem.
2. Inicializaremos um modelo GMM com $K = 4$.
3. Executaremos o algoritmo EM passo a passo e plotaremos o resultado em cada etapa, mostrando como as elipses (que representam as distribuições Gaussianas) se movem e se ajustam para "encontrar" os grupos nos dados.

Primeiro, vamos definir uma função auxiliar para desenhar as elipses que representam as distribuições de probabilidade de cada componente Gaussiano.

```
def draw_ellipse(position, covariance, ax=None, **kwargs):
    """Desenha uma elipse representando a covariância de um componente GMM."""
    ax = ax or plt.gca()

    # Decomposição da matriz de covariância
    if covariance.shape == (2, 2):
        U, s, Vt = np.linalg.svd(covariance)
        angle = np.degrees(np.arctan2(U[1, 0], U[0, 0]))
        width, height = 2 * np.sqrt(s)
    else:
        angle = 0
        width, height = 2 * np.sqrt(covariance)

    # Desenha a elipse
    for nsig in range(1, 4):
        ax.add_patch(mpl.patches.Ellipse(position, nsig * width, nsig * height,
                                         angle=angle, **kwargs))

def plot_gmm(gmm, X, label=True, ax=None):
    """Função para plotar os resultados do GMM."""
    ax = ax or plt.gca()
    labels = gmm.fit(X).predict(X)
```

```
    if label:
        ax.scatter(X[:, 0], X[:, 1], c=labels, s=40, cmap='viridis', zorder=2)
    else:
        ax.scatter(X[:, 0], X[:, 1], s=40, zorder=2)

    w_factor = 0.2 / gmm.weights_.max()
    for pos, covar, w in zip(gmm.means_, gmm.covariances_, gmm.weights_):
        draw_ellipse(pos, covar, ax=ax, alpha=w * w_factor, fill=True, facecolor='gray')

# Reduzindo a dimensionalidade com ACP para visualização
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

Agora, vamos observar o algoritmo EM em ação.

```
# Modelos GMM para cada etapa da visualização
gmm_iter0 = GaussianMixture(n_components=4, covariance_type='full', max_iter=0, n_init=1, init_pa
gmm_iter1 = GaussianMixture(n_components=4, covariance_type='full', max_iter=1, n_init=1, init_pa
gmm_iter5 = GaussianMixture(n_components=4, covariance_type='full', max_iter=5, n_init=1, init_pa
gmm_final = GaussianMixture(n_components=4, covariance_type='full', max_iter=100, n_init=1, init_

# Inicializa os modelos
gmm_iter0.fit(X_pca)
gmm_iter1.fit(X_pca)
gmm_iter5.fit(X_pca)
gmm_final.fit(X_pca)

fig, axes = plt.subplots(2, 2, figsize=(7, 7), sharex=True, sharey=True)

# Plot Iteração 0: Inicialização
axes[0, 0].scatter(X_pca[:, 0], X_pca[:, 1], s=15, cmap='viridis')
for pos, covar, w in zip(gmm_iter0.means_, gmm_iter0.covariances_, gmm_iter0.weights_):
    draw_ellipse(pos, covar, ax=axes[0, 0], alpha=0.2, fill=True, facecolor='gray')
axes[0, 0].set_title('Iteração 0: Inicialização (K-médias)')

# Plot Iteração 1
labels1 = gmm_iter1.predict(X_pca)
axes[0, 1].scatter(X_pca[:, 0], X_pca[:, 1], c=labels1, s=15, cmap='viridis')
for pos, covar, w in zip(gmm_iter1.means_, gmm_iter1.covariances_, gmm_iter1.weights_):
    draw_ellipse(pos, covar, ax=axes[0, 1], alpha=0.2, fill=True, facecolor='gray')
axes[0, 1].set_title('Iteração 1: Após 1 passo EM')
```

```python
# Plot Iteração 5
labels5 = gmm_iter5.predict(X_pca)
axes[1, 0].scatter(X_pca[:, 0], X_pca[:, 1], c=labels5, s=15, cmap='viridis')
for pos, covar, w in zip(gmm_iter5.means_, gmm_iter5.covariances_, gmm_iter5.weights_):
    draw_ellipse(pos, covar, ax=axes[1, 0], alpha=0.2, fill=True, facecolor='gray')
axes[1, 0].set_title('Iteração 5')

# Plot Convergência
labels_final = gmm_final.predict(X_pca)
axes[1, 1].scatter(X_pca[:, 0], X_pca[:, 1], c=labels_final, s=15, cmap='viridis')
for pos, covar, w in zip(gmm_final.means_, gmm_final.covariances_, gmm_final.weights_):
    draw_ellipse(pos, covar, ax=axes[1, 1], alpha=0.2, fill=True, facecolor='gray')
axes[1, 1].set_title('Convergência Final')

for ax in axes.flat:
    ax.set_xticks([])
    ax.set_yticks([])

plt.tight_layout()
plt.show()
```

```
/home/victor/Documents/EstatisticaMultivariada/.venv/lib/python3.13/site-packages/sklearn/mixture
  warnings.warn(
/home/victor/Documents/EstatisticaMultivariada/.venv/lib/python3.13/site-packages/sklearn/mixture
  warnings.warn(
/tmp/ipykernel_232124/2253255802.py:16: UserWarning: No data for colormapping provided via 'c'. Para
  axes[0, 0].scatter(X_pca[:, 0], X_pca[:, 1], s=15, cmap='viridis')
```

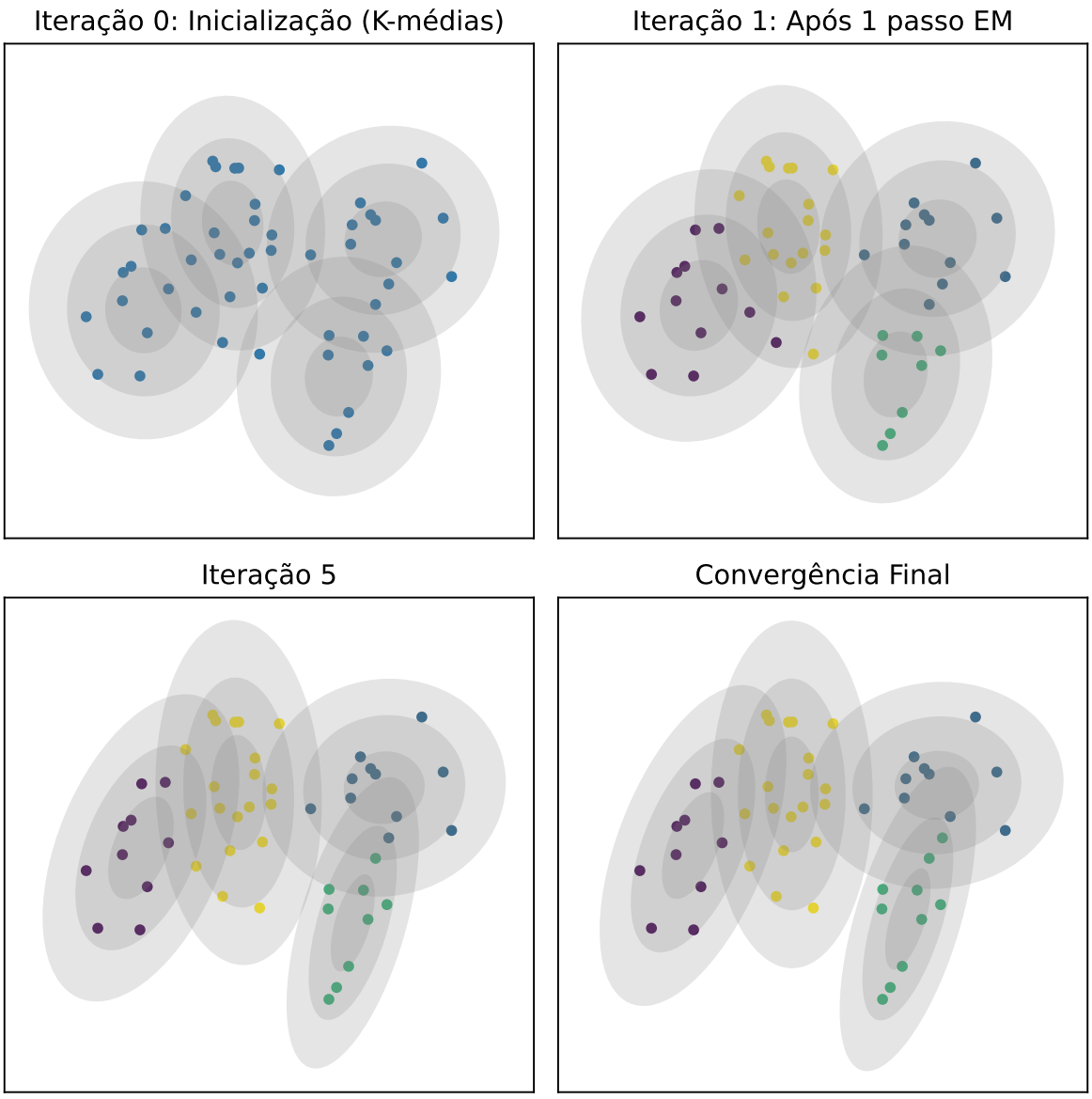

Figura 13.2: Visualização das iterações do algoritmo EM.

A Figura 13.2 ilustra o processo de ajuste do GMM:

1. **Iteração 0 (Inicialização):** O algoritmo é inicializado usando o resultado do K-médias. As elipses são posicionadas nos centroides do K-médias e são esféricas, pois a covariância inicial é simples.
2. **Iteração 1 (Após 1 passo EM):**

Tabela 13.2: Perfil dos grupos GMM: médias das variáveis originais para cada grupo.

```
# Usando o modelo final para obter os grupos
final_model = GaussianMixture(n_components=4, covariance_type='full', random_state=42).fit(X_scal
data['gmm_grupo'] = final_model.predict(X_scaled)

# Calcular as médias por grupo
gmm_profile = data.groupby('gmm_grupo').mean()
print(tabulate(gmm_profile, headers='keys', tablefmt='pipe'))
```

gmm_grupo	Murder	Assault	UrbanPop	Rape
0	3.94	82.8	54.3333	12.7467
1	14.6714	251.286	54.2857	21.6857
2	10.9667	264	76.5	33.6083
3	6	148.062	72.75	19.7063

- **Passo-E (*Expectation*):** O modelo calcula a probabilidade de cada ponto pertencer a cada uma das 4 distribuições Gaussianas (as "responsabilidades").
- **Passo-M (*Maximization*):** Usando essas probabilidades como pesos, o modelo atualiza os parâmetros de cada Gaussiana: a média (centro da elipse), a covariância (formato e orientação da elipse) e o peso da mistura (importância geral do grupo). Vemos que as elipses já se moveram e mudaram de forma para se adaptar melhor aos dados.

3. **Iteração 5:** Após mais algumas iterações, as elipses continuam a se mover e a se deformar, aproximando-se cada vez mais da estrutura subjacente dos dados. As atribuições de grupo (cores) tornam-se mais estáveis.
4. **Convergência Final:** O algoritmo para quando as mudanças nos parâmetros e na verossimilhança do modelo se tornam insignificantes. As elipses agora representam a forma, o tamanho e a orientação dos 4 grupos de dados identificados.

13.4 Interpretação dos Grupos

Agora que o modelo convergiu, podemos analisar os perfis dos grupos resultantes, assim como fizemos no exemplo do K-médias.

A Tabela 13.2 nos permite caracterizar os grupos. A numeração dos grupos pode ser diferente da do K-médias, mas os perfis são conceitualmente semelhantes:

- **Grupo 0 (Estados Seguros e Rurais):** Níveis muito baixos em todas as categorias de crime e a menor média de população urbana. Corresponde ao grupo 2 do exemplo de K-médias.

- **Grupo 1 (Estados Perigosos e Rurais):** Índices de criminalidade muito altos, especialmente `Murder` e `Assault`, combinados com uma população urbana relativamente baixa. Corresponde ao grupo 0 do K-médias.
- **Grupo 2 (Estados Urbanizados e Seguros):** Alta população urbana, mas com os níveis de crime mais moderados entre os grupos urbanizados. Corresponde ao grupo 3 ("Estados na Média") do K-médias.
- **Grupo 3 (Grandes Centros Urbanos Perigosos):** A maior população urbana e níveis de `Assault` e `Rape` muito elevados. Corresponde ao grupo 1 do K-médias.

Finalmente, vamos visualizar o resultado final no espaço dos componentes principais, com os nomes dos estados para uma interpretação mais clara.

```
# Criando um DataFrame para o plot
pca_df = pd.DataFrame(data=X_pca, columns=['PC1', 'PC2'])
pca_df['gmm_grupo'] = data['gmm_grupo'].values
pca_df['state'] = data.index

# Plotando
plt.figure(figsize=(7, 5))
sns.scatterplot(x='PC1', y='PC2', hue='gmm_grupo', data=pca_df, palette='viridis', s=100)

# Adicionando os nomes dos estados ao gráfico
for i in range(pca_df.shape[0]):
    plt.text(x=pca_df.PC1[i]+0.05, y=pca_df.PC2[i], s=pca_df.state[i],
             fontdict=dict(color='black',size=8))

# Interpretando os eixos da ACP
explained_variance = pca.explained_variance_ratio_
plt.title('Grupos GMM de Estados dos EUA (Visualização com ACP)')
plt.xlabel(f'Componente Principal 1 (Criminalidade Geral) ({explained_variance[0]:.1%})')
plt.ylabel(f'Componente Principal 2 (Urbanização) ({explained_variance[1]:.1%})')
plt.legend(title='Grupo')
plt.grid(True)
plt.show()
```

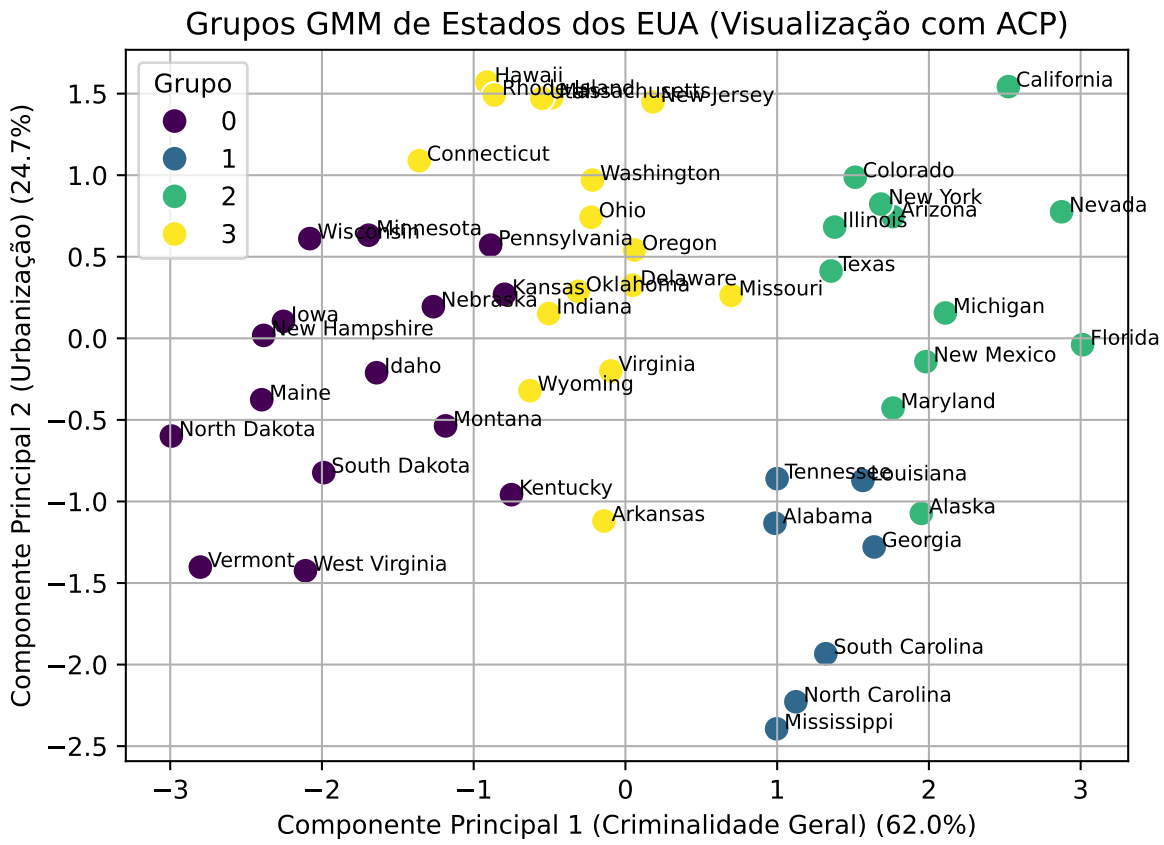

Figura 13.3: Visualização dos grupos GMM no espaço dos dois primeiros componentes principais.

A Figura 13.3 mostra a partição final. O resultado é muito semelhante ao obtido com o procedimento hierárquico + K-médias, o que reforça a validade da estrutura de 4 grupos encontrada nos dados. No entanto, a abordagem GMM nos forneceu um caminho mais formal para a seleção de K (via BIC) e uma compreensão mais profunda de como os grupos são modelados como distribuições de probabilidade, capazes de capturar formas e orientações mais complexas do que as esferas implícitas no K-médias.