

Métodos Estatísticos Multivariados

Teoria, interpretação e
aplicações em Python



VICTOR COSCRATO

Índice

Prefácio	3
I Parte I: Introdução	4
1 Introdução	5
1.1 Vetores aleatórios	6
1.2 Outras estruturas de dados multivariados	10
2 Álgebra matricial	13
3 A distribuição normal multivariada	20
3.1 Propriedades da distribuição normal multivariada	20
3.2 Visualizando a normal bivariada	21
3.3 Exercícios	22
II Parte II: Métodos multivariados	24
4 Análise de componentes principais	25
4.1 Variância como medida de informação	26
4.2 Centralização e escalonamento	28
4.3 Solução dos componentes principais	29
4.4 Interpretação dos componentes	32
4.5 Componentes principais amostrais	34
4.6 Escores	35
4.7 Redução de dimensão	37
4.8 Reconstrução e projeção ortogonal	39
4.9 Exemplo prático	40
4.10 Exercícios	44
4.11 Tópicos avançados	46
4.11.1 Soluções alternativas da ACP	46
4.11.2 ACP com kernel	48
4.11.3 Estabilidade da solução	49
4.11.4 Análise paralela	50
4.11.5 Outras reduções de dimensão	52

5	Análise de correspondência	54
5.1	O modelo de independência	56
5.2	Perfis e distâncias	57
5.3	Análise de correspondência simples	61
5.4	Coordenadas principais	63
5.5	Mapa de correspondência	64
5.5.1	Elementos suplementares	67
5.6	Exercícios	69
5.7	Análise de correspondência múltipla	70
5.7.1	Inércia na ACM	71
5.7.2	Solução e mapa de correspondência	72
5.7.3	Escolha das dimensões	74
5.7.4	Tabela de Burt	77
5.8	Exercícios	78
6	Análise fatorial	79
6.1	O modelo fatorial ortogonal	80
6.2	Estimação	83
6.2.1	Componentes principais	84
6.2.2	Fatoração do eixo principal	87
6.2.3	Máxima verossimilhança	88
6.3	Escolha do número de fatores	91
6.4	Rotação fatorial	92
6.4.1	Varimax	93
6.4.2	Outras rotações	96
6.5	Escores fatoriais	97
6.6	Exemplo prático	98
6.7	Exercícios	102
6.8	Tópicos avançados	104
6.8.1	Adequação dos dados	104
6.8.2	Análise fatorial confirmatória	106
6.8.3	Dados ordinais e correlações policóricas	106
6.8.4	A análise fatorial como modelo probabilístico	106
7	Análise de correlação canônica	108
7.1	Correlações canônicas	108
7.2	Interpretação	111
7.3	Inferência	113
7.4	Exemplo prático	114
7.5	Exercícios	118
	Leituras recomendadas	120

Prefácio

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

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

Os métodos cobertos são:

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

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

! Importante

Este material está (e continuará por um bom tempo) sendo revisto constantemente. Certamente existem erros e partes incompletas.

Parte I

Parte I: Introdução

1 Introdução

A análise multivariada é o campo da estatística dedicado a compreender conjuntos de dados com múltiplas variáveis correlacionadas. Os métodos multivariados permitem examinar simultaneamente as relações entre elas para extrair padrões e estruturas que de outra forma permaneceriam ocultos. Na prática, esses métodos atacam quatro problemas comuns: resumir conjuntos grandes reduzindo a sua dimensão sem perder variação relevante; descobrir grupos ou classificar novas observações em categorias já conhecidas; investigar a dependência entre variáveis para entender como elas se associam; e fazer previsões ou inferências quando há relações de modelagem estatística em jogo.

As técnicas que buscam esses objetivos costumam ser separadas em dois grupos. As **técnicas de dependência** partem de uma divisão prévia entre variáveis dependentes e independentes, e buscam prever ou explicar as primeiras a partir das segundas. Já as **técnicas de interdependência** tratam todas as variáveis igualmente, sem essa separação, e procuram descrever a estrutura do conjunto como um todo. A escolha da técnica também depende da natureza das variáveis: as **categóricas** (ou qualitativas) registram categorias e grupos, como gênero ou tipo de produto, enquanto as **métricas** (ou quantitativas) registram quantidades numéricas, como idade, altura, renda ou temperatura. Cada método foi construído tendo em vista certos tipos de variável, e é essa combinação entre objetivo e natureza dos dados que orienta a escolha.

A Tabela 1.1 reúne as técnicas apresentadas no livro segundo esses critérios, e a Figura 1.1 organiza a mesma informação como um roteiro de decisão, partindo das perguntas que costumamos fazer diante de um conjunto de dados. O diagrama é um guia simplificado e não exaustivo: ele serve para posicionar as técnicas discutidas aqui, e a escolha final deve sempre considerar o contexto do problema e as características detalhadas dos dados.

Tabela 1.1: Técnicas abordadas neste livro. Alguns dos capítulos ainda estão por vir.

Técnica	Objetivo(s)	Tipo de Análise
Análise de Componentes Principais (ACP)	Relação entre variáveis métricas ou Redução de dimensionalidade	Interdependência
Análise de Correspondência Simples (ACS) e Múltipla (ACM)	Relação entre variáveis categóricas	Interdependência
Análise Fatorial (AF)	Fatorização conjunta de variáveis métricas	Interdependência

Técnica	Objetivo(s)	Tipo de Análise
Análise de Agrupamento	Formação de grupos homogêneos	Interdependência
Análise Discriminante	Classificação de observações	Dependência
Correlação Canônica (ACC)	Relação entre conjuntos de variáveis métricas	Dependência

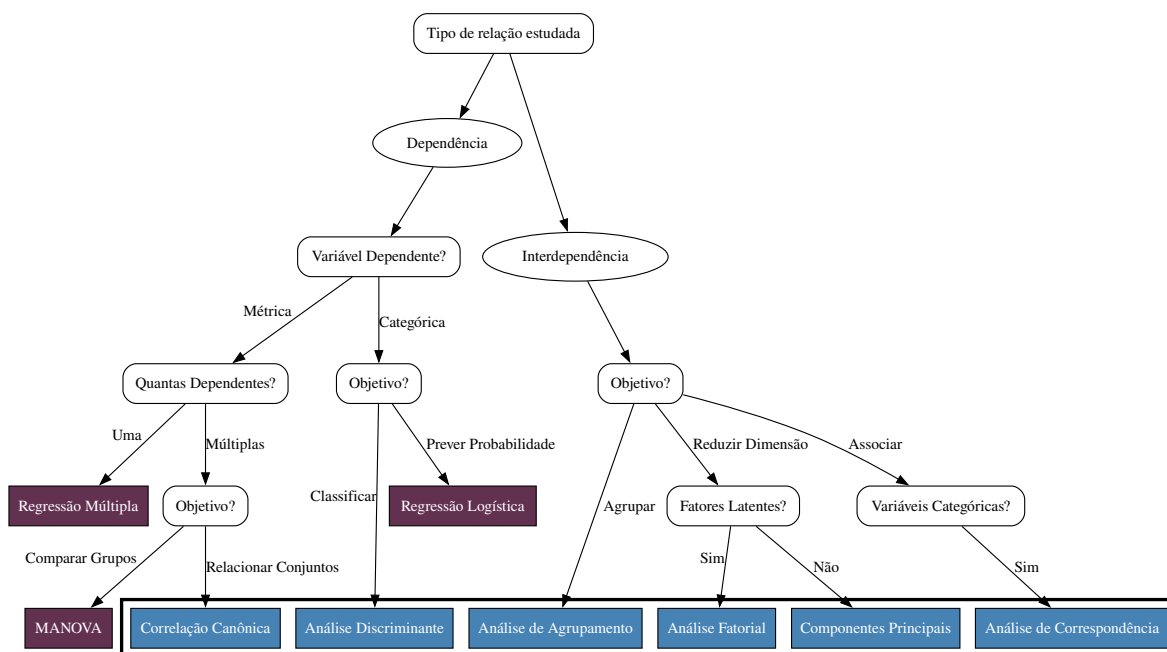


Figura 1.1: Diagrama de decisão para escolha de técnica de Análise Multivariada. Os nós azuis indicam as técnicas abordadas neste livro.

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.

1.1 Vetores aleatórios

Em estudos estatísticos multivariados, temos uma população de interesse (como os estudantes de uma universidade) e um conjunto de p características observadas (como notas em diferentes disciplinas e horas de estudo). Como essas características variam entre os elementos da população, cada uma delas pode ser tratada como uma variável aleatória, com sua própria distribuição de probabilidade. Para facilitar, podemos organizar essas variáveis em um **vetor aleatório**.

Definição 1.1 (Vetor aleatório). Um **vetor aleatório** $\mathbf{x}$ é um vetor-coluna cujos componentes são p variáveis aleatórias, $X_1, X_2, \dots, X_p$.

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

A esse vetor associamos uma distribuição multivariada que governa o comportamento da população. Suas propriedades e estrutura de variabilidade formam a base para todas as técnicas do livro.

Cuidado com a Notação

- Uma letra minúscula em negrito (e.g., $\mathbf{x}$) denota **vetores** em geral, sejam eles aleatórios, observados ou de constantes; o contexto distingue os casos.
- 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).
- O sobrescrito T denota **transposição**: $\mathbf{A}^T$ é a transposta de $\mathbf{A}$. Essa operação troca linhas por colunas; em particular, o vetor-coluna $\mathbf{x}$ da definição acima se torna o vetor-linha $\mathbf{x}^T = (X_1, X_2, \dots, X_p)$.

Assim como as variáveis aleatórias, os vetores aleatórios são caracterizados por parâmetros como a média e a variância, com o acréscimo de que agora também há parâmetros associados às correlações entre as variáveis em estudo.

Definição 1.2 (Vetor de médias populacional). 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 1.3 (Matriz de covariâncias populacional). 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, $\sigma_{jk} = \text{Cov}(X_j, X_k) = E[(X_j - \mu_j)(X_k - \mu_k)]$.

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

Na diagonal estão as **variâncias**, $\sigma_{jj} = \text{Var}(X_j)$, que medem a dispersão de cada variável; fora dela estão as **covariâncias**, σ_{jk} , que medem a tendência de associação linear entre as variáveis X_j e X_k . 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 1.4 (Matriz de correlações populacional). A **matriz de correlações populacional**, denotada por ρ , é uma versão reescalada da matriz de covariâncias, com elementos $\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}}$.

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

Escrevendo $\mathbf{D} = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}})$ para a matriz diagonal dos desvios padrão, a passagem de uma matriz para a outra é imediata:

$$\rho = \mathbf{D}^{-1}\Sigma\mathbf{D}^{-1}, \quad \Sigma = \mathbf{D}\rho\mathbf{D}$$

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

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

1. Esperança de uma Combinação Linear:

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

2. Covariância de uma Combinação Linear:

$$\text{Cov}(\mathbf{Ax} + \mathbf{c}) = \mathbf{A} \text{Cov}(\mathbf{x}) \mathbf{A}^T = \mathbf{A}\Sigma\mathbf{A}^T$$

Na prática, não temos acesso a esses parâmetros populacionais $(\boldsymbol{\mu}, \Sigma, \rho)$, mas podemos usar dados observados para obter estimativas confiáveis deles. 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 neste capítulo. A coleção de todas essas observações forma a nossa **matriz de dados**.

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

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

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

Na prática, os parâmetros populacionais não são conhecidos e precisam ser estimados a partir da matriz de dados. Embora nosso foco principal não seja a inferência, três estimadores usuais aparecem ao longo de todo o livro:

Definição 1.6 (Vetor de médias amostral). O estimador usual de $\boldsymbol{\mu}$ é o **vetor de médias amostral**, $\bar{\mathbf{x}}$. Escrevendo $\mathbf{1}_n$ para o vetor $n \times 1$ com todas as entradas iguais a 1, ele é dado por

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{X}^T \mathbf{1}_n$$

Elemento a elemento, cada componente $\bar{x}_j$ é a média das observações da 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 1.7 (Matriz de covariâncias amostral). O estimador usual de $\boldsymbol{\Sigma}$ é a **matriz de covariâncias amostral**, $\mathbf{S}$. Escrevendo $\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ para a **matriz de centralização**, que subtrai de cada coluna de $\mathbf{X}$ a sua média, temos

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}^T \mathbf{H} \mathbf{X} \quad \text{ou equivalentemente,} \quad \mathbf{S} = \frac{1}{n-1} (\mathbf{X}^T \mathbf{X} - n \bar{\mathbf{x}} \bar{\mathbf{x}}^T)$$

Elemento a elemento, s_{jk} é a variância amostral (quando $j = k$) ou a covariância amostral (quando $j \neq k$):

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad \text{resultando em} \quad \mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

i Nota

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

Definição 1.8 (Matriz de correlações amostral). O estimador usual de ρ é a **matriz de correlações amostral**, $\mathbf{R}$. Escrevendo $\mathbf{D}_s = \text{diag}(\sqrt{s_{11}}, \dots, \sqrt{s_{pp}})$ para a matriz diagonal dos desvios padrão amostrais, definimos $\mathbf{R}$ por:

$$\mathbf{R} = \mathbf{D}_s^{-1} \mathbf{S} \mathbf{D}_s^{-1}$$

Cada entrada r_{jk} de $\mathbf{R}$ é obtida padronizando a covariância amostral:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} \quad \text{resultando em} \quad \mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

1.2 Outras estruturas de dados multivariados

A matriz de dados $\mathbf{X}$ é a estrutura de dados costumeira para representar dados multivariados. No entanto, existem outras estruturas importantes que serão úteis durante este livro. Em particular, temos algumas estruturas para dados categóricos.

Definição 1.9 (Tabela de contingência). Sejam X e Y duas variáveis categóricas com I e J categorias, respectivamente. Uma tabela de contingência (ou matriz de contingência) $\mathbf{N}$ de dimensão $I \times J$ é uma matriz onde cada elemento n_{ij} representa a contagem (frequência absoluta) de observações que pertencem simultaneamente à i -ésima categoria de X e à j -ésima categoria de Y .

$X \setminus Y$	Categoria 1	Categoria 2	$\cdots$	Categoria J	Total
Categoria 1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1.}$
Categoria 2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Categoria I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.J}$	n

As somas nas margens da tabela fornecem os totais de cada categoria: $n_{i.} = \sum_{j=1}^J n_{ij}$ é o total marginal da i -ésima linha (categoria i de X) e $n_{.j} = \sum_{i=1}^I n_{ij}$ é o total marginal da j -ésima coluna (categoria j de Y). A soma de todas as contagens resulta no número total de observações da amostra, $n = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$.

Definição 1.10 (Matriz de correspondência). A **matriz de correspondência**, denotada por P , é obtida dividindo cada elemento da tabela de contingência N pelo total geral de observações n . Cada elemento $p_{ij} = n_{ij}/n$ representa a frequência relativa conjunta das categorias i e j .

$$P = \frac{1}{n}N$$

$X \setminus Y$	Categoria 1	Categoria 2	$\cdots$	Categoria J	Total
Categoria 1	p_{11}	p_{12}	$\cdots$	p_{1J}	$p_{1.}$
Categoria 2	p_{21}	p_{22}	$\cdots$	p_{2J}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Categoria I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	$p_{I.}$
Total	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.J}$	1

As somas marginais da matriz de correspondência formam os vetores de **proporções marginais**.

Quando existem três ou mais variáveis categóricas, duas outras estruturas são úteis: a matriz indicadora e a tabela de Burt.

Definição 1.11 (Matriz indicadora). Considere uma amostra de n observações sobre Q variáveis categóricas, onde a q -ésima variável tem J_q categorias, com $J = \sum_{q=1}^Q J_q$ categorias no total. A **matriz indicadora** (ou matriz de variáveis *dummy*), denotada por Z , é uma matriz binária de dimensão $n \times J$ particionada em Q blocos:

$$Z = (Z_1 \quad Z_2 \quad \cdots \quad Z_Q)$$

onde cada bloco Z_q tem dimensão $n \times J_q$. O elemento (i, j) de Z_q é igual a 1 se a i -ésima observação pertence à j -ésima categoria da variável q , e 0 caso contrário. Como cada observação assume exatamente uma categoria por variável, cada linha do bloco Z_q tem uma única entrada igual a 1. Por isso, cada linha da matriz completa Z soma Q (o número de variáveis), e a soma de todos os elementos de Z é igual a nQ . As somas das colunas de Z correspondem às frequências marginais de cada categoria.

Definição 1.12 (Tabela de Burt). A **tabela de Burt** (ou matriz de Burt), denotada por $\mathbf{B}$, é uma matriz simétrica de dimensão $J \times J$ obtida pelo produto $\mathbf{B} = \mathbf{Z}^T \mathbf{Z}$. Ela é estruturada em $Q \times Q$ blocos formados pelas tabelas de contingência entre todas as variáveis tomadas duas a duas:

$$\mathbf{B} = \begin{pmatrix} \mathbf{N}_{11} & \mathbf{N}_{12} & \cdots & \mathbf{N}_{1Q} \\ \mathbf{N}_{21} & \mathbf{N}_{22} & \cdots & \mathbf{N}_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{N}_{Q1} & \mathbf{N}_{Q2} & \cdots & \mathbf{N}_{QQ} \end{pmatrix}$$

onde cada bloco fora da diagonal, $\mathbf{N}_{qk} = \mathbf{Z}_q^T \mathbf{Z}_k$ ($q \neq k$), é a tabela de contingência entre as variáveis q e k . Quando $q = k$, o bloco diagonal $\mathbf{N}_{qq}$ é uma matriz diagonal contendo as frequências marginais da variável q .

2 Álgebra matricial

Grande parte dos métodos estatísticos multivariados consiste em olhar para uma nuvem de pontos, girar seus eixos, medir distâncias ou projetá-la em um subespaço de menor dimensão. Em linguagem algébrica, essas operações geométricas nada mais são do que multiplicações por matrizes, autovalores e decomposições. Este capítulo reúne as ferramentas dessa linguagem que usaremos ao longo de todo o livro.

Definição 2.1 (Produto interno e norma). Dados dois vetores $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, o **produto interno** (ou produto escalar) entre eles é o escalar definido por:

$$\mathbf{u}^T \mathbf{v} = \sum_{j=1}^p u_j v_j$$

A **norma euclidiana** (ou comprimento) de um vetor $\mathbf{u}$ é a raiz quadrada do produto interno de $\mathbf{u}$ consigo mesmo:

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}^T \mathbf{u}} = \sqrt{\sum_{j=1}^p u_j^2}$$

Um vetor é dito **unitário** (ou normalizado) quando $\|\mathbf{u}\| = 1$.

Definição 2.2 (Vetores ortogonais e ortonormais). Dois vetores $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ são ditos **ortogonais** se seu produto interno for nulo:

$$\mathbf{u}^T \mathbf{v} = 0$$

Geometricamente, vetores ortogonais formam um ângulo de 90° entre si. Um conjunto de vetores $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ é dito **ortonormal** se todos os seus elementos forem mutuamente ortogonais e tiverem norma unitária:

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & \text{se } i = j, \\ 0, & \text{se } i \neq j. \end{cases}$$

Definição 2.3 (Inversa de uma matriz). Uma matriz quadrada A de dimensão $p \times p$ é dita **invertível** (ou **não singular**) se existir uma matriz A^{-1} de mesma dimensão tal que:

$$AA^{-1} = A^{-1}A = I$$

A matriz A^{-1} é única e chamada de **matriz inversa** de A . Quando essa matriz não existe, A é dita **singular** (ou não invertível).

Proposição 2.1 (Propriedades da matriz inversa). *Sejam A e B matrizes invertíveis de dimensão $p \times p$ e $c \in \mathbb{R}$ um escalar não nulo. O inverso matricial satisfaz:*

1. **Inversa do produto:** $(AB)^{-1} = B^{-1}A^{-1}$.
2. **Inversa da transposta:** $(A^T)^{-1} = (A^{-1})^T$.
3. **Inversa da inversa:** $(A^{-1})^{-1} = A$.
4. **Multiplicação por escalar:** $(cA)^{-1} = \frac{1}{c}A^{-1}$.

Definição 2.4 (Matriz ortogonal). Uma matriz quadrada Q de dimensão $p \times p$ é dita **ortogonal** se sua transposta coincide com sua inversa:

$$Q^T Q = Q Q^T = I$$

Proposição 2.2 (Propriedades de matrizes ortogonais). *Seja Q uma matriz ortogonal de dimensão $p \times p$. Então:*

1. **Colunas e linhas ortonormais:** as colunas de Q formam uma base ortonormal de $\mathbb{R}^p$, e o mesmo vale para suas linhas.
2. **Preservação de comprimentos e distâncias:** para quaisquer vetores $u, v \in \mathbb{R}^p$,

$$(Qu)^T(Qv) = u^T Q^T Q v = u^T v$$

o que implica $\|Qu\| = \|u\|$ e $\|Qu - Qv\| = \|u - v\|$.

Definição 2.5 (Traço de uma matriz). Para uma matriz quadrada A de dimensão $p \times p$, o **traço** de A , denotado por $\text{tr}(A)$, é a soma dos elementos de sua diagonal principal:

$$\text{tr}(A) = \sum_{i=1}^p a_{ii}$$

Proposição 2.3 (Propriedades do traço). *Sejam A e B matrizes quadradas de ordem $p \times p$ e $c, d \in \mathbb{R}$ escalares. O traço satisfaz as seguintes propriedades:*

1. **Linearidade:** $\text{tr}(cA + dB) = c \text{tr}(A) + d \text{tr}(B)$.

2. **Invariância sob transposição:** $\text{tr}(\mathbf{A}^T) = \text{tr}(\mathbf{A})$.
3. **Propriedade cíclica:** se $\mathbf{A}$ tem dimensão $n \times p$ e $\mathbf{B}$ tem dimensão $p \times n$, então

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$$

Em geral, o traço do produto de várias matrizes é invariante sob permutações cíclicas, por exemplo $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB}) = \text{tr}(\mathbf{BCA})$, desde que as dimensões permitam as multiplicações.

Definição 2.6 (Norma de Frobenius). Para uma matriz $\mathbf{A} = (a_{ij})$ de dimensão $n \times p$, a **norma de Frobenius** de $\mathbf{A}$, denotada por $\|\mathbf{A}\|_F$, é a raiz quadrada da soma dos quadrados de todos os seus elementos:

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^p a_{ij}^2} = \sqrt{\text{tr}(\mathbf{A}^T \mathbf{A})}$$

Definição 2.7 (Determinante de uma matriz). Para uma matriz quadrada $\mathbf{A} = (a_{ij})$ de ordem $p \times p$, o **determinante** de $\mathbf{A}$, denotado por $\det(\mathbf{A})$, é o escalar definido indutivamente por:

- Para $p = 1$, $\det(\mathbf{A}) = a_{11}$;
- Para $p \geq 2$, pela expansão em cofatores ao longo da primeira linha:

$$\det(\mathbf{A}) = \sum_{j=1}^p (-1)^{1+j} a_{1j} \det(\mathbf{A}_{-1,-j})$$

onde $\mathbf{A}_{-1,-j}$ é a submatriz $(p-1) \times (p-1)$ obtida ao remover a primeira linha e a j -ésima coluna de $\mathbf{A}$.

Proposição 2.4 (Propriedades do determinante). Sejam $\mathbf{A}$ e $\mathbf{B}$ matrizes quadradas de ordem $p \times p$ e $c \in \mathbb{R}$ um escalar. O determinante satisfaz:

1. **Multiplicatividade:** $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.
2. **Inversão:** se $\mathbf{A}$ é invertível, $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$.
3. **Invariância sob transposição:** $\det(\mathbf{A}^T) = \det(\mathbf{A})$.
4. **Multiplicação por escalar:** $\det(c\mathbf{A}) = c^p \det(\mathbf{A})$.

Definição 2.8 (Posto de uma matriz). O **posto** de uma matriz $\mathbf{A}$ de dimensão $n \times p$, denotado por $\text{posto}(\mathbf{A})$, é o número de colunas (ou linhas) linearmente independentes de $\mathbf{A}$.

Proposição 2.5 (Propriedades do posto). Para uma matriz $\mathbf{A}$ de dimensão $n \times p$:

1. $\text{posto}(\mathbf{A}) \leq \min(n, p)$.

2. Se $\mathbf{A}$ é quadrada de ordem $p \times p$, dizemos que $\mathbf{A}$ tem **posto completo** se $\text{posto}(\mathbf{A}) = p$. Uma matriz quadrada é invertível se e somente se tem posto completo, o que equivale a $\det(\mathbf{A}) \neq 0$.

Definição 2.9 (Forma quadrática). Uma forma quadrática é uma função polinomial que contém apenas termos de grau dois. Dado um vetor $\mathbf{x}$ de dimensão $p \times 1$ e uma matriz numérica simétrica $\mathbf{A}$ de dimensão $p \times p$, o produto definido por:

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

é uma forma quadrática.

Definição 2.10 (Matriz positiva-definida). Uma matriz simétrica $\mathbf{A}$ é dita **positiva-definida** se $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ para todo vetor não nulo $\mathbf{x}$. Se $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$ para todo $\mathbf{x}$, a matriz é dita **positiva-semidefinida**.

Definição 2.11 (Matriz idempotente). Uma matriz quadrada $\mathbf{A}$ é dita **idempotente** se

$$\mathbf{A}^2 = \mathbf{A}$$

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

Definição 2.12 (Autovalores e autovetores). Seja $\mathbf{A}$ uma matriz quadrada de ordem $p \times p$. Um escalar λ é dito um **autovalor** de $\mathbf{A}$ associado ao **autovetor** não nulo $\mathbf{e} \in \mathbb{R}^p$ se:

$$\mathbf{A} \mathbf{e} = \lambda \mathbf{e}$$

Essa equação pode ser reescrita como $(\mathbf{A} - \lambda \mathbf{I}) \mathbf{e} = \mathbf{0}$. Para que exista uma solução não nula $\mathbf{e}$, a matriz $\mathbf{A} - \lambda \mathbf{I}$ deve ser singular.

Teorema 2.1 (Decomposição espectral). Toda matriz simétrica $\mathbf{A}$ de dimensão $p \times p$ com autovalores reais $\lambda_1, \dots, \lambda_p$ e autovetores ortonormais correspondentes $\mathbf{e}_1, \dots, \mathbf{e}_p$, pode ser decomposta como:

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

onde $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ é a matriz diagonal dos autovalores e $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_p)$ é a matriz ortogonal ($\mathbf{E}^T \mathbf{E} = \mathbf{E} \mathbf{E}^T = \mathbf{I}$) cujas colunas são os autovetores ortonormais.

Proposição 2.6 (Propriedades dos autovalores). Para uma matriz simétrica $\mathbf{A}$ de ordem $p \times p$ com autovalores $\lambda_1, \dots, \lambda_p$ e decomposição espectral $\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$:

1. **Soma dos autovalores (Traço):** Pela propriedade cíclica do traço e pela ortogonalidade de $\mathbf{E}$,

$$\sum_{i=1}^p \lambda_i = \text{tr}(\mathbf{\Lambda}) = \text{tr}(\mathbf{\Lambda}\mathbf{E}^T\mathbf{E}) = \text{tr}(\mathbf{E}\mathbf{\Lambda}\mathbf{E}^T) = \text{tr}(\mathbf{A})$$

2. **Produto dos autovalores (Determinante):** Pela propriedade multiplicativa do determinante,

$$\prod_{i=1}^p \lambda_i = \det(\mathbf{\Lambda}) = \det(\mathbf{E}) \det(\mathbf{\Lambda}) \det(\mathbf{E}^T) = \det(\mathbf{E}\mathbf{\Lambda}\mathbf{E}^T) = \det(\mathbf{A})$$

3. **Caracterização de matrizes positiva-definidas:** $\mathbf{A}$ é positiva-definida (positiva-semidefinida) se e somente se todos os seus autovalores satisfazem $\lambda_i > 0$ ($\lambda_i \geq 0$) para todo $i = 1, \dots, p$.
4. **Matrizes idempotentes:** Se $\mathbf{A}$ é simétrica e idempotente ($\mathbf{A}^2 = \mathbf{A}$), seus autovalores são exclusivamente 0 ou 1, o que implica $\text{posto}(\mathbf{A}) = \text{tr}(\mathbf{A}) = \sum_{i=1}^p \lambda_i$.

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

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

1. **Autovalores:** Resolvendo a equação característica $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$, isto é, $(2 - \lambda)^2 - 1 = 0$, encontramos $\lambda_1 = 3$ e $\lambda_2 = 1$.
2. **Autovetores:**

- Para $\lambda_1 = 3$: Resolvendo $(\mathbf{A} - 3\mathbf{I})\mathbf{e}_1 = \mathbf{0}$, obtemos $\mathbf{e}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$.
- Para $\lambda_2 = 1$: Resolvendo $(\mathbf{A} - 1\mathbf{I})\mathbf{e}_2 = \mathbf{0}$, obtemos $\mathbf{e}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$.

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

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

Assim, $\mathbf{A}$ multiplica por 3 a direção de $\mathbf{e}_1$ e por 1 a direção de $\mathbf{e}_2$. Note que:

- $\text{tr}(\mathbf{A}) = 2 + 2 = 4$ e $\lambda_1 + \lambda_2 = 3 + 1 = 4$;
- $\det(\mathbf{A}) = 2 \cdot 2 - 1 \cdot 1 = 3$ e $\lambda_1 \lambda_2 = 3 \cdot 1 = 3$.

Definição 2.13 (Matriz raiz quadrada). Seja $\mathbf{A}$ uma matriz simétrica e positiva-semidefinida de ordem $p \times p$, com decomposição espectral $\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$, onde $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ com $\lambda_i \geq 0$. A **matriz raiz quadrada** de $\mathbf{A}$ é definida por:

$$\mathbf{A}^{1/2} = \mathbf{E}\mathbf{\Lambda}^{1/2}\mathbf{E}^T$$

onde $\mathbf{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$, satisfazendo $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$. Se $\mathbf{A}$ for positiva-definida ($\lambda_i > 0$), sua **inversa da raiz quadrada** é:

$$\mathbf{A}^{-1/2} = \mathbf{E}\mathbf{\Lambda}^{-1/2}\mathbf{E}^T$$

onde $\mathbf{\Lambda}^{-1/2} = \text{diag}(1/\sqrt{\lambda_1}, \dots, 1/\sqrt{\lambda_p})$, satisfazendo $\mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2} = \mathbf{I}$.

Definição 2.14 (Decomposição em valores singulares). Toda matriz $\mathbf{A}$ de dimensão $I \times J$ com posto $r = \text{posto}(\mathbf{A})$ pode ser escrita como:

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

onde $\mathbf{U}$ é uma matriz ortogonal $I \times I$ cujas colunas são os **vetores singulares à esquerda**, $\mathbf{V}$ é uma matriz ortogonal $J \times J$ cujas colunas são os **vetores singulares à direita**, e $\mathbf{\Lambda}$ é uma matriz retangular $I \times J$ que contém os **valores singulares** λ_k em sua diagonal principal, ordenados de modo que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$, e zeros nas demais posições. Os valores singulares não nulos são as raízes quadradas dos autovalores positivos de $\mathbf{A}^T\mathbf{A}$ e $\mathbf{A}\mathbf{A}^T$.

Enquanto a decomposição espectral se aplica a matrizes simétricas, a SVD é um resultado mais geral, que permite decompor qualquer matriz retangular. Para uma matriz simétrica e positiva-semidefinida $\mathbf{A}$, a SVD coincide com a decomposição espectral: seus valores singulares são os próprios autovalores, e seus vetores singulares à esquerda e à direita são os autovetores ($\mathbf{U} = \mathbf{V} = \mathbf{E}$).

Teorema 2.2 (Teorema de Eckart-Young-Mirsky). *Seja $\mathbf{A}$ uma matriz de dimensão $I \times J$ com posto r e decomposição em valores singulares $\mathbf{A} = \sum_{k=1}^r \lambda_k \mathbf{u}_k \mathbf{v}_k^T$, em que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$. Para qualquer inteiro $q < r$, a melhor aproximação de posto q de $\mathbf{A}$ segundo a norma de Frobenius é a soma truncada da SVD:*

$$\hat{\mathbf{A}}_q = \sum_{k=1}^q \lambda_k \mathbf{u}_k \mathbf{v}_k^T$$

Isto é, para qualquer matriz $\mathbf{B}$ com $\text{posto}(\mathbf{B}) \leq q$,

$$\|\mathbf{A} - \hat{\mathbf{A}}_q\|_F \leq \|\mathbf{A} - \mathbf{B}\|_F$$

e o erro quadrático mínimo de aproximação é a soma dos quadrados dos valores singulares descartados:

$$\|\mathbf{A} - \hat{\mathbf{A}}_q\|_F^2 = \sum_{k=q+1}^r \lambda_k^2$$

3 A distribuição normal multivariada

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

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

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp \left\{ -\frac{1}{2} \underbrace{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}_{d_M^2(\mathbf{x}, \boldsymbol{\mu})} \right\}$$

A densidade combina dois elementos geométricos. A constante normalizadora contém a raiz quadrada da **variância generalizada**, $\det(\boldsymbol{\Sigma})^{1/2}$, que quantifica o hipervolume ocupado pela distribuição no espaço p -dimensional. O termo no expoente é a **distância de Mahalanobis ao quadrado** entre $\mathbf{x}$ e o vetor de médias $\boldsymbol{\mu}$,

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

Diferentemente da distância euclidiana usual, $\|\mathbf{x} - \boldsymbol{\mu}\|^2 = (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu})$, que atribui o mesmo peso a todas as direções e assume variáveis não correlacionadas com mesma escala, a distância de Mahalanobis pondera cada direção pela sua variabilidade e desconta as correlações mútuas. Usando a inversa da raiz quadrada $\boldsymbol{\Sigma}^{-1/2}$, temos $d_M^2(\mathbf{x}, \boldsymbol{\mu}) = \|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|^2$, de modo que pontos com a mesma distância de Mahalanobis ao centro formam contornos de densidade constante e recebem a mesma probabilidade.

3.1 Propriedades da distribuição normal multivariada

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

1. **Transformações lineares.** Se $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então, para uma matriz constante $\mathbf{A}$ de dimensão $q \times p$ e um vetor constante $\mathbf{b}$ de dimensão $q \times 1$,

$$\mathbf{Ax} + \mathbf{b} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

Em particular, toda combinação linear escalar $\mathbf{a}^T \mathbf{x}$ tem distribuição normal univariada $N(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a})$.

2. **Distribuições marginais.** Qualquer subconjunto das componentes de um vetor NMV também tem distribuição normal multivariada. Por exemplo, se $\mathbf{x} = (X_1, X_2, X_3)^T$ é NMV, então $(X_1, X_3)^T$ também é NMV.
3. **Covariância zero e independência.** Em geral, não correlação não implica independência. Para vetores conjuntamente normais, porém, dois subconjuntos são independentes quando toda a matriz de covariâncias cruzadas entre eles é nula.

Proposição 3.1 (Distribuição de formas quadráticas). Se $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\boldsymbol{\Sigma}$ positiva-definida, então:

1. A distância de Mahalanobis ao quadrado segue uma distribuição qui-quadrado com p graus de liberdade:

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

2. Se $\mathbf{P}$ é uma matriz de projeção simétrica e idempotente de dimensão $p \times p$ com posto $r = \text{tr}(\mathbf{P}) \leq p$, então:

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1/2} \mathbf{P} \boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_r^2$$

i Nota

A propriedade $d_M^2(\mathbf{x}, \boldsymbol{\mu}) \sim \chi_p^2$ é a base para o diagnóstico de normalidade multivariada na prática. Dada uma amostra $\mathbf{x}_1, \dots, \mathbf{x}_n$, calculamos as distâncias amostrais $d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$. Se os dados forem aproximadamente normais, o gráfico de quantis amostrais de d_i^2 contra os quantis teóricos da distribuição χ_p^2 (*gráfico quantil-quantil qui-quadrado*) deve se alinhar próximo a uma reta.

3.2 Visualizando a normal bivariada

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

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

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

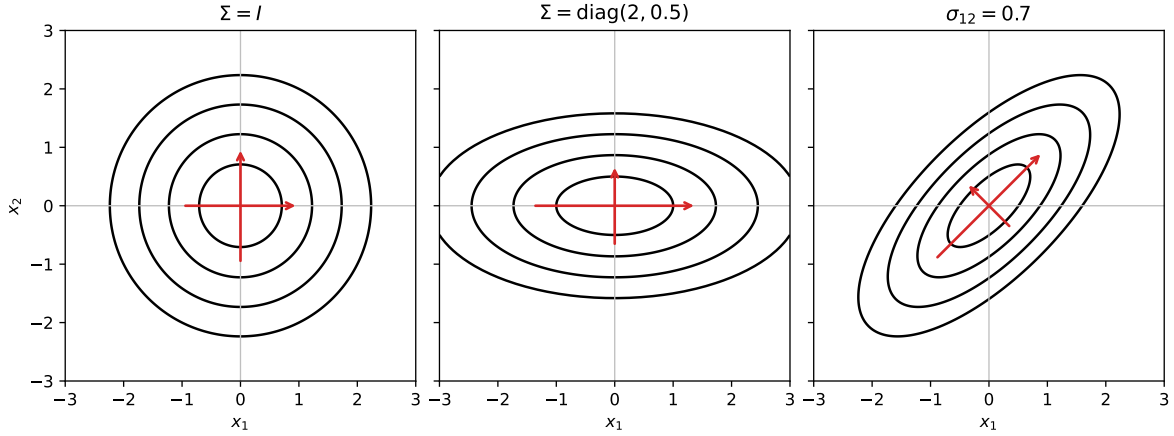


Figura 3.1: Contornos de densidade de normais bivariadas. Os vetores vermelhos indicam os autovetores da matriz de covariâncias.

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

A normalidade multivariada é um modelo de referência importante para métodos de classificação e inferência baseados em verossimilhança. Já métodos puramente descritivos e geométricos podem ser aplicados sem assumir normalidade, alterando-se principalmente a interpretação inferencial dos resultados.

3.3 Exercícios

Exercício 3.1. Seja $\mathbf{x} = (X_1, X_2)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}$$

- Determine a distribuição marginal de X_1 e a distribuição marginal de X_2 .
- Encontre a distribuição da combinação linear $Y = X_1 - 2X_2$.
- Determine o valor da constante c para que a combinação linear $Z = X_1 + cX_2$ seja independente de Y .

Exercício 3.2. Considere um vetor aleatório $\mathbf{x} \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$, com $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

a) Calcule a matriz inversa Σ^{-1} e a distância de Mahalanobis ao quadrado $d_M^2(\mathbf{x}, \mathbf{0})$ para o ponto $\mathbf{x} = (1, 1)^T$.

b) Determine a equação da elipse que delimita a região de 95% de probabilidade para $\mathbf{x}$, sabendo que o quantil de 95% da distribuição χ_2^2 é aproximadamente 5,99.

Exercício 3.3. Seja $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ com Σ positiva-definida. Mostre que o vetor padronizado $\mathbf{z} = \Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ tem distribuição $N_p(\mathbf{0}, \mathbf{I})$. Em seguida, use esse resultado para provar que a distância de Mahalanobis ao quadrado $d_M^2(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{z}^T \mathbf{z}$ segue uma distribuição χ_p^2 .

Parte II

Parte II: Métodos multivariados

4 Análise de componentes principais

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

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

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

Em forma matricial, podemos escrever $\mathbf{y} = \mathbf{U}^T \mathbf{x}$, em que $\mathbf{y} = (Y_1, \dots, Y_p)^T$ e as colunas da matriz $\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_p]$ definem as direções das novas coordenadas. Os CPs satisfazem duas propriedades:

- São não correlacionados entre si, isto é, $\text{Cov}(Y_k, Y_l) = 0$ para todo $k \neq l$.
- Suas variâncias são decrescentes: $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \dots \geq \text{Var}(Y_p)$.

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

Exemplo 4.1. Suponha que estamos estudando o atraso até o destino de dois ônibus municipais. Sejam X_1 e X_2 os atrasos dos dois ônibus, com $\mathbf{x} = (X_1, X_2)^T \sim N_2(\mathbf{0}, \Sigma)$, onde

$$\Sigma = \begin{bmatrix} 1,0 & 0,7 \\ 0,7 & 1,0 \end{bmatrix}$$

Ou seja, os atrasos médios são zero, os dois atrasos têm a mesma variância populacional, e a alta covariância indica atrasos conjuntos, que podem ser explicados por más condições de trânsito na cidade. A Figura 4.1 (esquerda) mostra uma elipse de concentração dessa distribuição, ou seja, o

conjunto de pontos $\mathbf{x}$ que satisfazem $\mathbf{x}^T \Sigma^{-1} \mathbf{x} = c^2$, para uma constante c qualquer, fixada apenas para fins de visualização.

Os eixos X_1 e X_2 são perfeitamente válidos, mas não acompanham a orientação principal da elipse. A posição da elipse indica que a dispersão é muito maior ao longo de uma direção diagonal. Ao longo desse capítulo, vamos propor e justificar o uso de novos eixos que acompanhem a direção de maior dispersão dos dados, representadas na Figura 4.1 (direita), os chamados componentes principais.

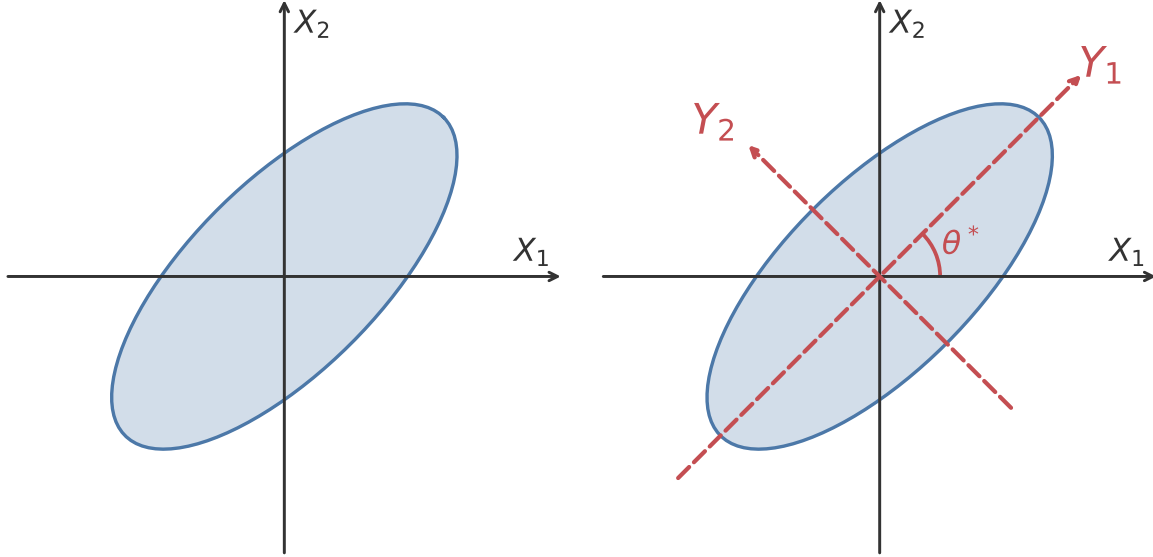


Figura 4.1: Elipse de concentração de Σ nos eixos originais e nos novos eixos Y_1 e Y_2 .

4.1 Variância como medida de informação

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. Em um caso extremo, se uma variável tiver variância zero, todas as observações seriam idênticas, ou seja, não teríamos informação útil sobre elas.

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

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

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

Exemplo 4.2 (Interpretação e solução geométrica). Retomando o Exemplo 4.1, a Figura 4.1 (direita) mostra que a busca pelo eixo mais informativo Y_1 pode ser traduzida na busca por um ângulo ideal θ^* que define uma rotação do eixo original X_1 . Para um ângulo θ qualquer, a rotação pode ser expressa pela matriz de rotação

$$[\mathbf{u}_1(\theta) \quad \mathbf{u}_2(\theta)] = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

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

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

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

$$\begin{aligned} \text{Var}(Y_1) &= \text{Var}(\mathbf{u}_1(\theta)^T \mathbf{x}) \\ &= \mathbf{u}_1(\theta)^T \Sigma \mathbf{u}_1(\theta) \\ &= \begin{bmatrix} \cos \theta & \sin \theta \end{bmatrix} \begin{bmatrix} 1,0 & 0,7 \\ 0,7 & 1,0 \end{bmatrix} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \\ &= \underbrace{\cos^2 \theta + \sin^2 \theta}_1 + \underbrace{1,4 \sin \theta \cos \theta}_{0,7 \sin 2\theta} \\ &= 1 + 0,7 \sin 2\theta \end{aligned}$$

Portanto, a variância do primeiro eixo é uma função do ângulo θ . A ACP escolhe a rotação que maximiza essa função. Nesse exemplo simples, podemos observar isso graficamente. A Figura 4.2 ilustra a solução.

A figura mostra que o ângulo ideal é $\theta^* = 45,0$, com esse ângulo, temos $\text{Var}(Y_1) = \cos^2(45^\circ) + 1,4 \sin(45^\circ) \cos(45^\circ) + \sin^2(45^\circ) = 1,7$. As variâncias dos atrasos nos eixos originais, por sua vez, são $\text{Var}(X_1) = 1,0$ e $\text{Var}(X_2) = 1,0$, de modo que a variância total é $\text{tr}(\Sigma) = 2,0$.

Note que Y_1 tem variância maior que X_1 e X_2 . A variância residual é $2,0 - 1,7 = 0,3$. É aqui que introduzimos o segundo novo eixo Y_2 . Ele é perpendicular a Y_1 e captura essa variância residual, de modo que $\text{Var}(Y_2) = 0,3$. Fica claro como os novos eixos Y_1 e Y_2 definem um par de coordenadas que redistribui a variabilidade. O primeiro eixo Y_1 captura a maior parte da informação, enquanto o segundo Y_2 captura a variância residual.

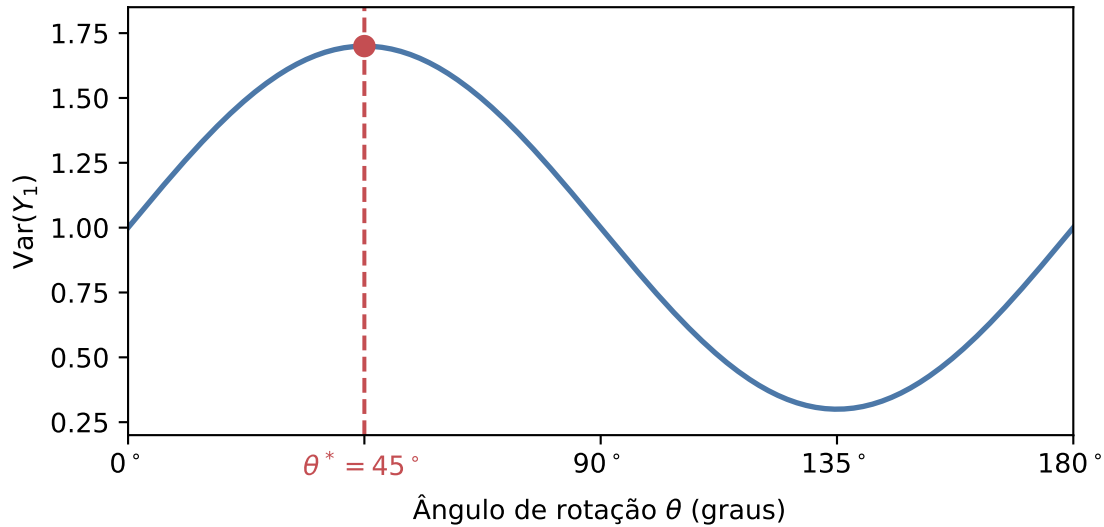


Figura 4.2: Variância da primeira coordenada rotacionada em função do ângulo. O máximo identifica a direção do primeiro componente principal.

Enquanto essa solução geométrica é intuitiva em duas dimensões, ela não se estende de maneira conveniente a dezenas ou centenas de variáveis. Em breve, veremos como a solução analítica, baseada em autovalores, resolve o problema de forma geral.

4.2 Centralização e escalonamento

! Importante

Antes de aplicar a ACP, precisamos centralizar e avaliar a necessidade de escalonar as variáveis.

Centralização. A ACP busca eixos que descrevam a dispersão interna dos dados em torno do seu centro de massa $\mu = E[x]$. Geometricamente, no entanto, os novos eixos lineares são subespaços vetoriais que sempre passam pela origem $(0, \dots, 0)$.

Se os dados estiverem distantes da origem e não forem centralizados, a direção de maior magnitude a partir do zero não refletirá o formato da nuvem, mas apenas a posição do seu centro de massa. Em termos matemáticos, a magnitude quadrática média de uma projeção $Y = u^T x$ em relação à origem decompõe-se em

$$E[Y^2] = \text{Var}(u^T x) + (u^T \mu)^2$$

A primeira parcela, $\text{Var}(\mathbf{u}^T \mathbf{x})$, mede a dispersão real dos dados em torno da média, enquanto a segunda, $(\mathbf{u}^T \boldsymbol{\mu})^2$, mede a distância da origem até a média ao longo da direção $\mathbf{u}$. Se a média $\boldsymbol{\mu}$ for grande em relação à variabilidade interna, esse segundo termo dominará o cálculo. Com isso, o primeiro eixo resultante apenas ligará a origem ao centro da nuvem, ignorando a verdadeira estrutura de variabilidade dos dados.

Ao centralizarmos o vetor aleatório fazendo $\mathbf{x} - \boldsymbol{\mu}$, garantimos que o centro da nuvem coincida com a origem ($E[\mathbf{x} - \boldsymbol{\mu}] = \mathbf{0}$). O termo $(\mathbf{u}^T \boldsymbol{\mu})^2$ desaparece e a análise passa a depender exclusivamente da matriz de covariâncias $\boldsymbol{\Sigma}$, tornando a ACP invariante a translações. Por conveniência, assumimos daqui em diante que $\mathbf{x}$ já representa o vetor centralizado, com média zero.

Escalonamento. A variância depende diretamente da unidade de medida. Imagine que registramos o atraso de um transporte: medido em horas, a variância pode ser 1; se medirmos o mesmo atraso em minutos, a variância salta para $1 \times 60^2 = 3600$. O fenômeno observado não se tornou mais informativo por ter sido medido em minutos, mas a sua variância numérica explodiu. Como a ACP busca a direção de maior variabilidade, ela daria peso demais a essa variância inflada pela unidade de medida.

Para evitar esse efeito, quando as variáveis têm unidades distintas ou ordens de grandeza incomparáveis, padronizamos cada uma delas dividindo-a por seu desvio padrão. Todas as variáveis passam a ter variância 1 e tornam-se adimensionais. Com $\mathbf{D}$ a matriz diagonal dos desvios padrão (introduzida em Definição 1.4), a transformação de padronização é $\mathbf{D}^{-1}(\mathbf{x} - \boldsymbol{\mu})$. A matriz de variâncias e covariâncias desse vetor de variáveis padronizadas é precisamente a matriz de correlações $\boldsymbol{\rho}$ de $\mathbf{x}$.

Ou seja, aplicar a ACP sobre as variáveis padronizadas equivale a fazer a decomposição sobre a matriz de correlações $\boldsymbol{\rho}$, em vez da matriz de covariâncias $\boldsymbol{\Sigma}$. Quando as variáveis já compartilham a mesma unidade e suas diferenças de dispersão são relevantes, as escalas originais devem ser mantidas.

4.3 Solução dos componentes principais

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

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

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

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

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

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

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

Chegamos à equação de autovalores e autovetores. A direção candidata $\mathbf{u}_1$ que resolve o problema deve ser um autovetor da matriz de covariâncias Σ . Para encontrar a variância, multiplicamos a equação à esquerda por $\mathbf{u}_1^T$ e usamos a restrição

$$\text{Var}(Y_1) = \mathbf{u}_1^T \Sigma \mathbf{u}_1 = \lambda_1 \underbrace{\mathbf{u}_1^T \mathbf{u}_1}_1 = \lambda_1$$

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

i Nota

A matriz de covariâncias Σ é, por construção, simétrica e positiva semidefinida. Conforme discutido em Teorema 2.1, o **Teorema Espectral** garante que seus autovalores são reais e não negativos e que existe uma base ortonormal de autovetores.

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

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

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

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

Pré-multiplicamos a derivada por e_1^T . Como e_1 é autovetor de Σ , temos $e_1^T \Sigma = \lambda_1 e_1^T$, e as restrições cuidam do resto

$$\begin{aligned} 0 &= 2e_1^T \Sigma u_2 - 2\lambda_2 e_1^T u_2 - \phi e_1^T e_1 \\ &= 2(\lambda_1 - \lambda_2) \underbrace{e_1^T u_2}_0 - \phi \underbrace{e_1^T e_1}_1 \\ &= -\phi \end{aligned}$$

Logo, $\phi = 0$, e a derivada se reduz a $\Sigma u_2 = \lambda_2 u_2$. A segunda direção também deve, portanto, ser um autovetor de Σ . Entre os autovetores ortogonais a e_1 , a maior variância disponível é o segundo maior autovalor, λ_2 . Denotando o autovetor unitário correspondente por e_2 , a solução é $u_2 = e_2$.

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

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

i Nota

A dedução por multiplicadores de Lagrange é direta e estende-se naturalmente aos componentes seguintes. Outra abordagem clássica restringe a norma de u_1 por meio do quociente,

$$\text{Var}(Y_1) = \max_{u_1} \frac{u_1^T \Sigma u_1}{u_1^T u_1}$$

Trata-se do quociente de Rayleigh. Para qualquer matriz simétrica A , o máximo da forma quadrática $x^T A x$, sujeito a $x^T x = 1$, coincide com o **maior autovalor** de A , e o vetor que atinge esse máximo é o autovetor associado. Como a matriz de covariâncias Σ é simétrica, o resultado se aplica diretamente ao nosso problema.

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

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

Prova. Pela Definição 4.2, a variância total das variáveis originais é $\text{tr}(\Sigma)$. Usando a decomposição espectral $\Sigma = E \Lambda E^T$ e a propriedade cíclica do traço, obtemos

$$\begin{aligned}
\sum_{j=1}^p \text{Var}(X_j) &= \text{tr}(\Sigma) \\
&= \text{tr}(\mathbf{E}\mathbf{\Lambda}\mathbf{E}^T) \\
&= \text{tr}\left(\mathbf{\Lambda} \underbrace{\mathbf{E}^T\mathbf{E}}_{\mathbf{I}}\right) \\
&= \sum_{k=1}^p \lambda_k \\
&= \sum_{k=1}^p \text{Var}(Y_k)
\end{aligned}$$

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

A Proposição 4.1 mostra que a soma das variâncias dos componentes principais é sempre igual à variância total das variáveis originais. Essa quantidade é fixa, mas como ela se distribui entre os componentes depende diretamente da estrutura de correlação entre as variáveis originais. Quando as variáveis estão fortemente correlacionadas, essa distribuição tende a ser desigual, com um ou dois autovalores concentrando a maior parte da variância e os componentes seguintes em papel secundário. Esse é o cenário do Exemplo 4.1, em que a correlação de 0,7 alonga a elipse de concentração ao longo de uma direção, permitindo que Y_1 sozinho capture a maior parte da informação.

No outro extremo, se as variáveis originais são pouco ou nada correlacionadas, essa distribuição tende a se tornar mais uniforme entre os componentes. No caso limite de variáveis não correlacionadas e com a mesma variância, a elipse de concentração se torna um círculo, e nenhuma direção acumula mais variância do que outra. A ACP deixa de trazer qualquer utilidade nesse caso.

4.4 Interpretação dos componentes

As colunas da matriz $\mathbf{E} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \cdots \ \mathbf{e}_p]$ contêm os coeficientes que definem os componentes principais. O elemento e_{jk} (linha j , coluna k) é frequentemente chamado de **carga** (*loading*) e indica o peso da variável original X_j na formação do componente Y_k . A magnitude de e_{jk} reflete a importância relativa da variável j no componente k , mas fica difícil comparar esses coeficientes diretamente quando as variáveis têm unidades e variâncias distintas.

Uma medida mais interpretável e padronizada é a **correlação entre as variáveis originais e os componentes principais**, $\text{Corr}(X_j, Y_k)$. Ela indica o quanto cada variável original está linearmente associada a cada componente, numa escala de -1 a 1 .

Proposição 4.2 (Correlação entre variáveis originais e componentes principais). *Seja $\mathbf{y} = \mathbf{E}^T \mathbf{x}$ o vetor com todos os componentes principais de $\mathbf{x}$, com $\lambda_k > 0$ e $\sigma_{jj} = \text{Var}(X_j) > 0$ para todos j e k . Então*

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

ou, reunindo todas as correlações em uma única matriz,

$$\text{Corr}(\mathbf{x}, \mathbf{y}) = \mathbf{D}^{-1} \mathbf{E} \mathbf{\Lambda}^{1/2}$$

em que $\mathbf{D} = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}})$ é a matriz diagonal dos desvios padrão de $\mathbf{x}$.

Prova. Todas as covariâncias entre as variáveis originais e os componentes saem de uma única conta. Usando a decomposição espectral $\mathbf{\Sigma} = \mathbf{E} \mathbf{\Lambda} \mathbf{E}^T$ e a ortogonalidade de $\mathbf{E}$,

$$\text{Cov}(\mathbf{x}, \mathbf{y}) = \text{Cov}(\mathbf{x}, \mathbf{E}^T \mathbf{x}) = \mathbf{\Sigma} \mathbf{E} = \mathbf{E} \mathbf{\Lambda}$$

O elemento (j, k) dessa matriz é $\text{Cov}(X_j, Y_k) = \lambda_k e_{jk}$. Como $\text{Var}(X_j) = \sigma_{jj}$ e $\text{Var}(Y_k) = \lambda_k$, basta padronizar

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

Essa padronização, feita de uma vez para toda a matriz, corresponde a dividir cada linha j por $\sqrt{\sigma_{jj}}$ e cada coluna k por $\sqrt{\lambda_k}$, isto é, a multiplicar $\mathbf{E} \mathbf{\Lambda}$ por $\mathbf{D}^{-1}$ à esquerda e por $\mathbf{\Lambda}^{-1/2}$ à direita

$$\text{Corr}(\mathbf{x}, \mathbf{y}) = \mathbf{D}^{-1} \mathbf{E} \mathbf{\Lambda} \mathbf{\Lambda}^{-1/2} = \mathbf{D}^{-1} \mathbf{E} \mathbf{\Lambda}^{1/2}$$

□

Na prática, os componentes são nomeados a partir dessa matriz, e a coluna k de $\mathbf{D}^{-1} \mathbf{E} \mathbf{\Lambda}^{1/2}$ diz, em uma escala comum a todas as variáveis, quais delas Y_k de fato resume.

4.5 Componentes principais amostrais

Toda a derivação anterior foi formulada em termos populacionais, supondo conhecidos a matriz de covariâncias Σ e seus autovalores e autovetores. Foi exatamente essa hipótese que nos permitiu desenhar a elipse de concentração exata do Exemplo 4.1. Na prática, onde não conhecemos Σ , observamos uma amostra de indivíduos e precisamos estimar essa estrutura a partir dela.

Nesse caso, basta substituir Σ pela matriz de covariâncias amostral S . Os componentes principais são obtidos da mesma maneira, decompondo a matriz de covariâncias (ou a matriz de correlações no caso padronizado) amostral e as quantidades resultantes são estimativas de seus análogos populacionais:

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

A teoria e a interpretação permanecem as mesmas. Para simplificar a notação, omitimos o circunflexo ao longo deste capítulo, lembrando que na prática lidamos sempre com estimativas amostrais.

Exemplo 4.3 (Uma amostra da população). Retomamos o exemplo dos atrasos Exemplo 4.1, mas agora assumindo Σ desconhecido. Observamos uma amostra de $n = 160$ viagens para cada ônibus. A Figura 4.3 (esquerda) mostra a nuvem desses 160 pontos nos eixos originais X_1 e X_2 . Como não conhecemos Σ , calculamos a matriz de covariâncias amostral S e a decomparamos para estimar os eixos dos componentes principais, $\hat{e}_1$ e $\hat{e}_2$; a Figura 4.3 (direita) mostra esses eixos estimados.

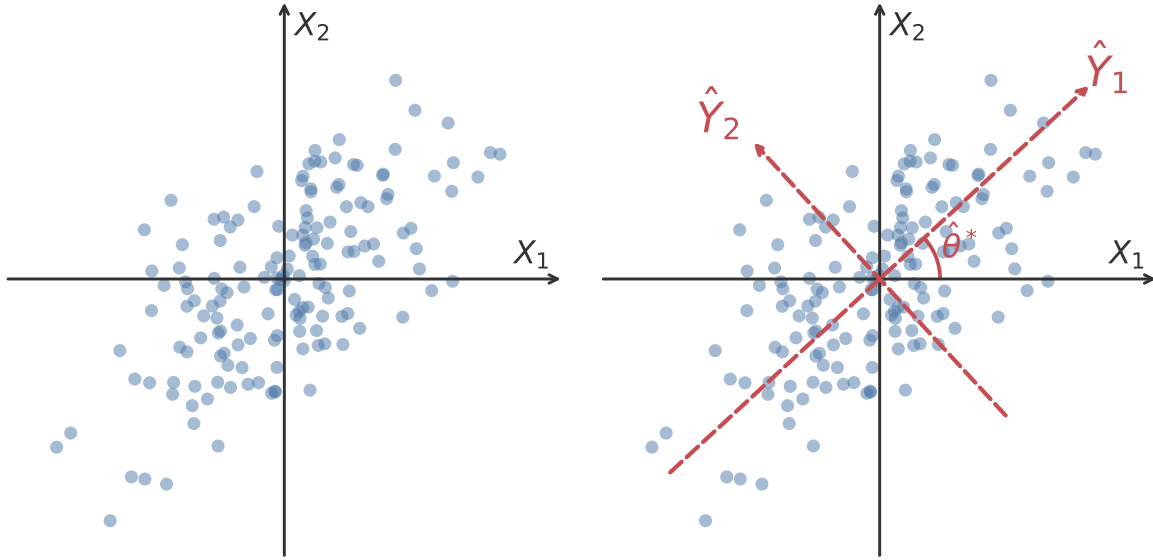


Figura 4.3: Amostra de 160 observações nos eixos originais e nos eixos estimados $\hat{Y}_1$ e $\hat{Y}_2$.

A interpretação geométrica da ACP não muda, ainda buscamos os eixos que maximizam a variância, mas agora segundo a amostra, ou seja, seguimos a direção de maior dispersão da nuvem de pontos. Nesse caso, poderíamos fazer uma estimativa do ângulo ideal de rotação θ^* . Para esses dados gerados, temos $\hat{\theta}^* \approx 42.7$. O valor é próximo, mas não idêntico, ao ângulo populacional exato do Exemplo 4.1.

4.6 Escores

Até agora, tratamos os componentes principais como direções, vetores $e_1, \dots, e_p$ que definem o novo sistema de eixos. Quando o interesse está em reduzir a dimensionalidade dos dados ou em representar graficamente as observações nesse novo sistema, precisamos também calcular a posição de cada indivíduo da amostra ao longo de cada novo eixo.

Definição 4.3 (Matriz de escores). Dada a matriz de dados centralizados $\mathbf{X}$, de dimensão $n \times p$, e a matriz ortogonal de autovetores $\mathbf{E} = [e_1 \ e_2 \ \dots \ e_p]$, a **matriz de escores** $\mathbf{T}$, de dimensão $n \times p$, é obtida pela projeção das observações no novo sistema de eixos

$$\mathbf{T} = \mathbf{X}\mathbf{E} = [\mathbf{t}_1 \ \mathbf{t}_2 \ \dots \ \mathbf{t}_p]$$

em que cada coluna $\mathbf{t}_k = \mathbf{X}\mathbf{e}_k$ ($n \times 1$) contém as coordenadas de todas as observações no k -ésimo componente principal Y_k .

A matriz de escores $\mathbf{T}$ representa os próprios dados originais reescritos no novo sistema de coordenadas. Suas linhas continuam correspondendo às n observações, e suas colunas, aos componentes principais $Y_1, \dots, Y_p$. Essa representação é a base para a redução de dimensionalidade e é usada diretamente em gráficos de dispersão e em análises multivariadas posteriores (como agrupamento ou classificação).

Para facilitar ainda mais a interpretação, existem representações visuais úteis para a ACP. A mais comum é o **biplot**. Ele combina os escores das observações nos componentes principais com vetores associados às variáveis originais. Assim, os pontos representam as observações, enquanto as setas mostram as direções em que as variáveis aumentam. Voltaremos ao biplot no Exemplo 4.6.

Exemplo 4.4 (Resolução manual). Para acompanhar como esses cálculos acontecem passo a passo a partir de uma matriz de dados, vamos resolver um exemplo inteiramente à mão, com uma amostra pequena cujos números foram escolhidos para simplificar as contas.

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

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

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

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

A matriz de covariâncias amostral é, portanto,

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

Neste exemplo, vamos assumir que X_1 e X_2 têm escalas comparáveis e, por isso, não padronizaremos os dados, ou seja, vamos aplicar a decomposição diretamente em $\mathbf{S}$.

Autovalores e Autovetores. Os autovalores resolvem a equação característica

$$\det \begin{bmatrix} 2.5 - \lambda & 1.5 \\ 1.5 & 2.5 - \lambda \end{bmatrix} = (2.5 - \lambda)^2 - 1.5^2 = \lambda^2 - 5\lambda + 4 = 0$$

As raízes são os autovalores $\lambda_1 = 4$ e $\lambda_2 = 1$. Note que $\lambda_1 + \lambda_2 = 5 = \text{tr}(\mathbf{S})$, como esperado.

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

$$\lambda_1 = 4 : \quad \begin{bmatrix} -1.5 & 1.5 \\ 1.5 & -1.5 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies -1.5e_{11} + 1.5e_{21} = 0 \implies e_{21} = e_{11}$$

$$\text{Normalizando com } e_{11}^2 + e_{21}^2 = 1 : \quad \mathbf{e}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\lambda_2 = 1 : \quad \begin{bmatrix} 1.5 & 1.5 \\ 1.5 & 1.5 \end{bmatrix} \begin{bmatrix} e_{12} \\ e_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies 1.5e_{12} + 1.5e_{22} = 0 \implies e_{22} = -e_{12}$$

$$\text{Normalizando com } e_{12}^2 + e_{22}^2 = 1 : \quad \mathbf{e}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Os autovetores e_1 e e_2 são ortogonais entre si ($e_1^T e_2 = 0$) e têm norma 1, formando a base do novo sistema de coordenadas.

Componentes principais. As combinações lineares que definem os novos eixos são

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

em que X_1 e X_2 representam as variáveis centralizadas.

Banco de dados transformado. Reunindo os autovetores em $E = [e_1 \ e_2]$, obtemos a matriz de dados expressa no sistema de coordenadas dos componentes principais

$$T = XE = X \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -4 & 0 \\ 0 & -2 \\ 0 & 0 \\ 0 & 2 \\ 4 & 0 \end{bmatrix}$$

As linhas de T continuam representando as cinco observações, enquanto suas colunas agora correspondem aos escores em Y_1 e Y_2 .

4.7 Redução de dimensão

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

Na prática, escolhemos reter $q < p$ componentes principais, descartando $Y_{q+1}, \dots, Y_p$. As propriedades dos componentes principais garantem que esse descarte maximiza a informação preservada.

Para que essa redução seja adequada, é preciso escolher q cuidadosamente. Poucos componentes tornam a análise mais simples e fácil de interpretar, enquanto mais componentes preservam mais da variância original. Entre os critérios mais comuns para a escolha de q , temos:

- O **critério da variância explicada acumulada** calcula a proporção da variância total explicada por cada componente e acumula essa proporção,

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

escolhendo 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.

- O **critério do autovalor (ou critério de Kaiser)**, aplicável quando a ACP é baseada na matriz de correlações, sugere reter apenas os componentes cujos autovalores (λ_k) são maiores que 1. Nesse caso, as variáveis originais são padronizadas para ter variância 1, e um componente com autovalor menor que 1 está explicando menos variabilidade do que uma única variável original. Reter tal componente não traria uma “economia” de informação, tornando-o um candidato à exclusão.
- O **gráfico do cotovelo (ou scree plot)**, uma alternativa mais visual, ilustrado na Figura 4.4, mostra os autovalores em função do número de componentes principais, com um “cotovelo” indicando o ponto de corte. A ideia é reter os componentes que aparecem antes do cotovelo, pois eles concentram a maior parte da variância total.

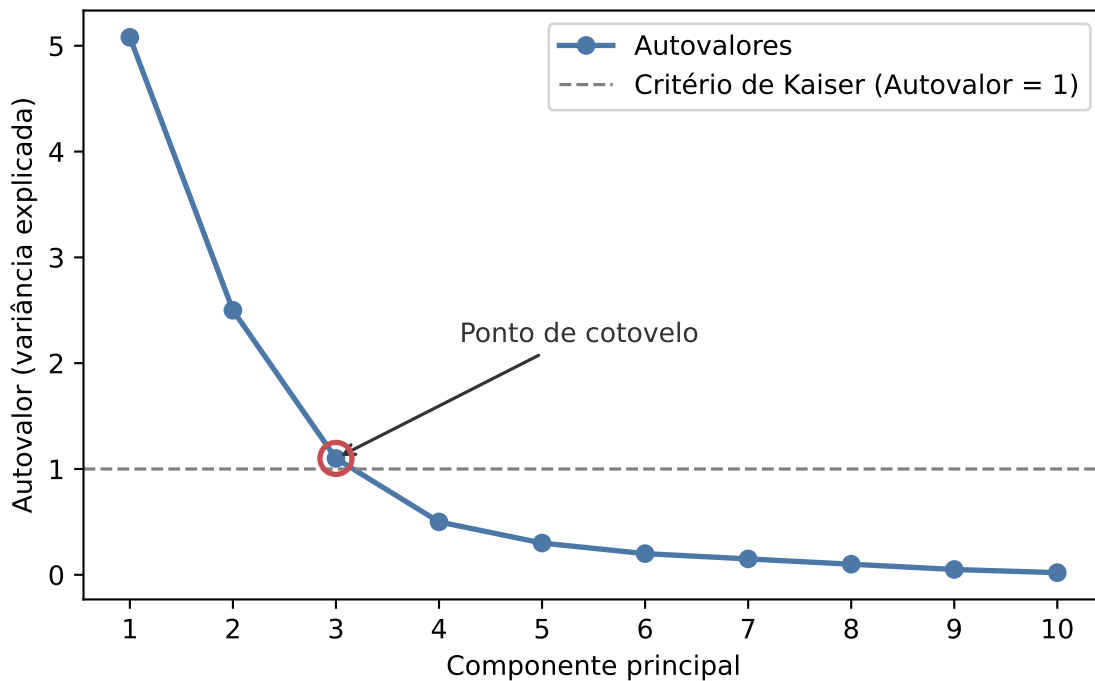


Figura 4.4: Exemplo de um *scree plot*. O cotovelo em $k = 3$ sugere a retenção de 3 componentes.

4.8 Reconstrução e projeção ortogonal

A projeção dada pela ACP é reversível. Isto é, podemos reconstruir os dados originais a partir dos seus escores e autovetores. Como a matriz de autovetores $\mathbf{E}$ é ortogonal, podemos multiplicar a matriz de escores $\mathbf{T} = \mathbf{X}\mathbf{E}$ à direita por $\mathbf{E}^T$ para recuperar perfeitamente os dados centralizados

$$\mathbf{X} = \mathbf{T}\mathbf{E}^T = \sum_{k=1}^p \mathbf{t}_k \mathbf{e}_k^T$$

Essa equação decompõe a matriz de dados em uma soma de p matrizes de posto um, onde cada termo $\mathbf{t}_k \mathbf{e}_k^T$ representa a contribuição exclusiva do k -ésimo componente principal. Agora, quando reduzimos a dimensionalidade retendo apenas $q < p$ componentes, a aproximação de posto q da matriz de dados centralizados é

$$\widehat{\mathbf{X}}_q = \mathbf{T}_q \mathbf{E}_q^T = \sum_{k=1}^q \mathbf{t}_k \mathbf{e}_k^T$$

onde $\mathbf{T}_q = \mathbf{X}\mathbf{E}_q$ ($n \times q$) reúne os escores mantidos e $\mathbf{E}_q = [\mathbf{e}_1 \cdots \mathbf{e}_q]$ ($p \times q$) contém as respectivas direções principais. Pelo Teorema de Eckart-Young-Mirsky (Teorema 2.2), essa reconstrução $\widehat{\mathbf{X}}_q$ é a melhor aproximação de posto q da matriz de dados centralizados segundo a norma de Frobenius.

Proposição 4.3 (Erro de reconstrução). *Se $\widehat{\mathbf{X}}_q$ é a aproximação de posto q da matriz de dados centralizados $\mathbf{X}$, obtida pelos q primeiros componentes principais, então o erro quadrático de reconstrução normalizado equivale à soma dos autovalores dos componentes descartados*

$$\frac{1}{n-1} \|\mathbf{X} - \widehat{\mathbf{X}}_q\|_F^2 = \sum_{k=q+1}^p \lambda_k$$

Prova. Como $\mathbf{X} - \widehat{\mathbf{X}}_q = \sum_{k=q+1}^p \mathbf{t}_k \mathbf{e}_k^T$, em que os autovetores $\mathbf{e}_k$ formam um conjunto ortonormal e os vetores de escores $\mathbf{t}_k$ são mutuamente ortogonais, a ortogonalidade dos termos garante que

$$\|\mathbf{X} - \widehat{\mathbf{X}}_q\|_F^2 = \sum_{k=q+1}^p \|\mathbf{t}_k \mathbf{e}_k^T\|_F^2 = \sum_{k=q+1}^p \|\mathbf{t}_k\|^2 \|\mathbf{e}_k\|^2 = \sum_{k=q+1}^p \|\mathbf{t}_k\|^2$$

Dividindo por $n-1$, e lembrando que $\frac{1}{n-1} \|\mathbf{t}_k\|^2 = \frac{1}{n-1} \mathbf{t}_k^T \mathbf{t}_k = \lambda_k$ é a variância amostral do k -ésimo componente principal

$$\frac{1}{n-1} \|\mathbf{X} - \widehat{\mathbf{X}}_q\|_F^2 = \sum_{k=q+1}^p \frac{\|\mathbf{t}_k\|^2}{n-1} = \sum_{k=q+1}^p \lambda_k$$

□

Exemplo 4.5 (Reconstrução dos dados manuais). Retomando o Exemplo 4.4, suponha que desejamos comprimir os dados mantendo apenas o primeiro componente principal ($q = 1$). O vetor de escores retido é $\mathbf{t}_1 = \frac{1}{\sqrt{2}}(-4, 0, 0, 0, 4)^T$ e o autovetor correspondente é $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(1, 1)^T$. A reconstrução aproximada dos dados originais é

$$\widehat{\mathbf{X}}_1 = \mathbf{t}_1 \mathbf{e}_1^T = \frac{1}{2} \begin{bmatrix} -4 \\ 0 \\ 0 \\ 0 \\ 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} = \begin{bmatrix} -2 & -2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 2 & 2 \end{bmatrix}$$

A matriz de resíduos (erro de aproximação) é a diferença entre os dados originais e os reconstruídos

$$\mathbf{X} - \widehat{\mathbf{X}}_1 = \begin{bmatrix} -2 & -2 \\ -1 & 1 \\ 0 & 0 \\ 1 & -1 \\ 2 & 2 \end{bmatrix} - \begin{bmatrix} -2 & -2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -1 & 1 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \end{bmatrix}$$

O erro quadrático médio de reconstrução é

$$\frac{1}{4} \|\mathbf{X} - \widehat{\mathbf{X}}_1\|_F^2 = \frac{1}{4} ((-1)^2 + 1^2 + 1^2 + (-1)^2) = \frac{4}{4} = 1$$

Note que esse erro coincide exatamente com a variância descartada $\lambda_2 = 1$. O primeiro componente preservou 80% da variância total ($\lambda_1 / \text{tr}(\mathbf{S}) = 4/5$), e a perda de informação foi exatamente a variância do segundo componente.

4.9 Exemplo prático

Exemplo 4.6 (Medidas corporais dos pinguins). Vamos exemplificar o uso da ACP com o conjunto de dados [Palmer Penguins](#), que traz, para cada pinguim, o comprimento e a profundidade do bico, o comprimento da nadadeira e a massa corporal. As três primeiras estão em milímetros; a última, em gramas. Por isso, uma ACP baseada diretamente na matriz de covariâncias daria peso excessivo à massa corporal. Começamos removendo os valores ausentes e padronizando cada variável.

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

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

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

Agora fazemos a decomposição espectral de $\mathbf{R}$. As colunas da matriz $\mathbf{E}$ são os autovetores e definem as combinações lineares; a matriz $\mathbf{T}$ contém os escores das observações. A Tabela 4.2 apresenta a variância explicada, enquanto a Figura 4.5 mostra os mesmos autovalores graficamente.

i Nota

Existem bibliotecas que facilitam a aplicação da ACP, como `sklearn.decomposition.PCA`. Como os cálculos são extremamente simples, vamos apenas usar a decomposição espectral fornecida pelo NumPy.

Tabela 4.2: Variância explicada pelos componentes principais.

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

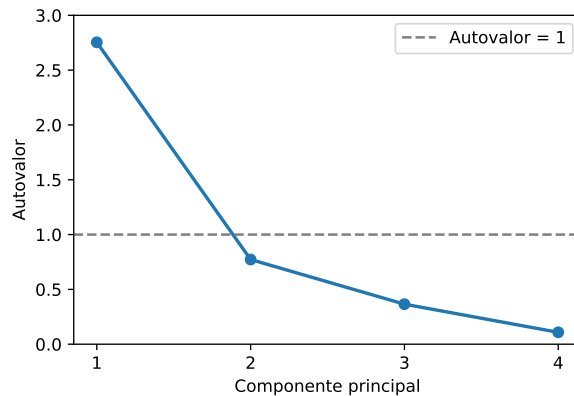


Figura 4.5: Gráfico do cotovelo da ACP dos pinguins.

O primeiro componente tem variância 2.75 e concentra 68.8% da variância total. Os dois primeiros, juntos, concentram 88.2%. Se o interesse do problema fosse a redução da dimensionalidade, seria

razoável manter dois componentes, pois eles preservam a maior parte da variância e permitem analisar os dados em um espaço bidimensional. Aqui, vamos focar na interpretação de CP1 e CP2.

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

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

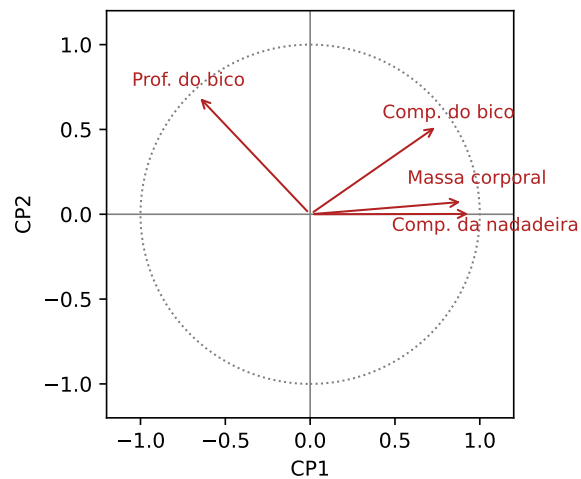


Figura 4.6: Círculo de correlações das medidas corporais dos pinguins.

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

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

O círculo de correlações, mostrado na Figura 4.6, permite avaliar as relações entre as variáveis e sua representação no plano. O comprimento ao quadrado de cada seta é a soma das correlações ao quadrado com CP1 e CP2: quanto mais próxima da circunferência estiver sua ponta, melhor a variável está representada. Para variáveis bem representadas, ângulos pequenos indicam correlação positiva, ângulos próximos de 180° indicam correlação negativa e ângulos próximos de 90° indicam correlação próxima de zero. Aqui, comprimento da nadadeira e massa corporal têm setas próximas, enquanto profundidade do bico e comprimento da nadadeira formam um ângulo obtuso, coerente com sua correlação negativa. Como usamos apenas dois componentes, essas leituras são aproximadas.

O banco de dados original contém também as espécies dos pinguins. Elas são três: Adelie, Chinstrap e Gentoo. Vamos construir um biplot, colorindo os pontos de acordo com a espécie. No biplot da Figura 4.7, os pontos mostram os escores brutos nos dois primeiros componentes e as setas indicam a direção em que cada medida padronizada aumenta na reconstrução. As setas usam os coeficientes dos autovetores, ampliados por um fator comum de 3 para facilitar a visualização. Essa escala difere da

usada no círculo de correlações, de modo que os ângulos entre as setas deste biplot não devem ser interpretados como correlações.

 Dica

Note que a variável relativa à espécie do animal não entrou no cálculo da ACP, ela aparece apenas na etapa interpretativa para completar nossas conclusões. Nesse caso, dizemos que a espécie é uma **variável exógena**.

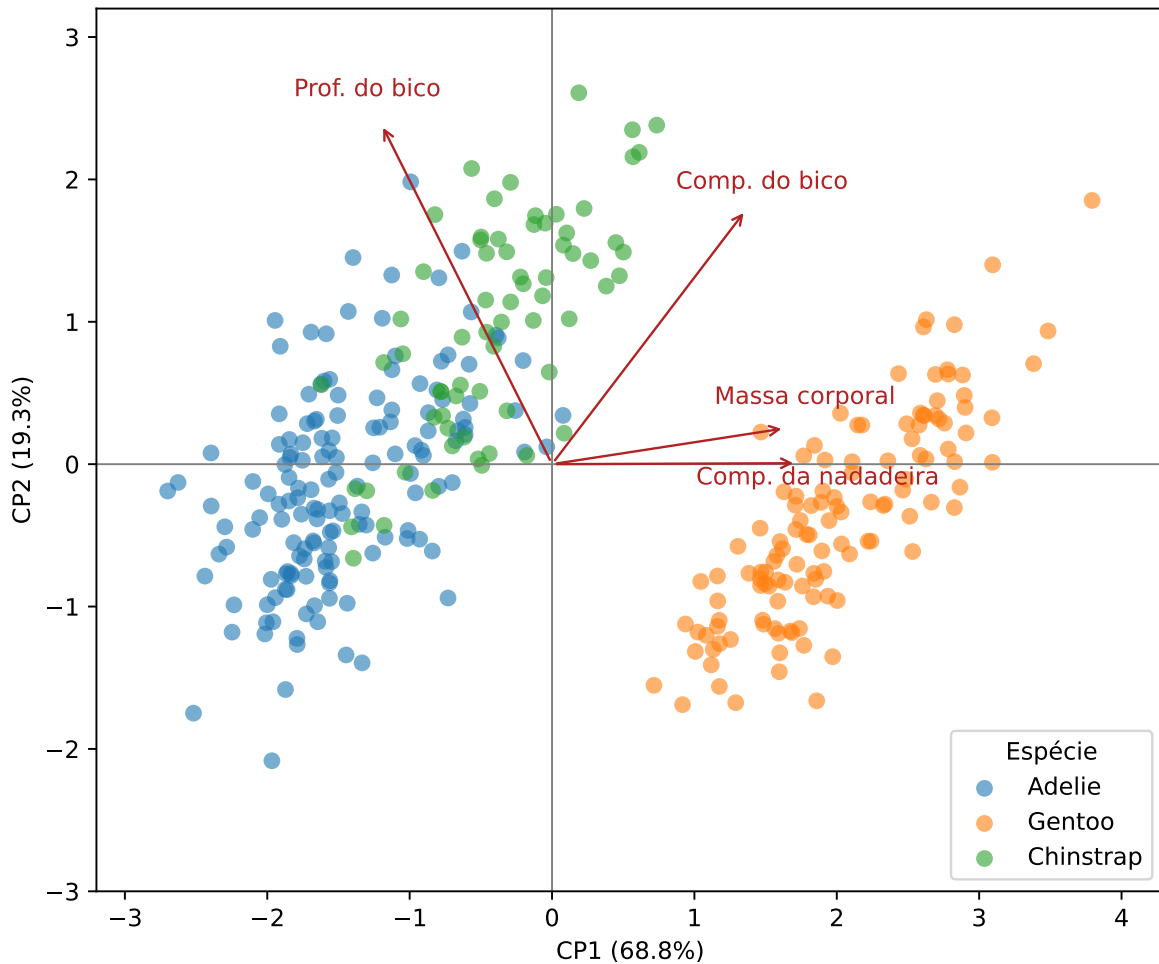


Figura 4.7: Biplot das medidas corporais dos pinguins, com os escores brutos nos dois primeiros componentes.

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

esquerda, ilustrando que esses pinguins são mais leves e menores.

Os Chinstrap ocupam a parte superior, associada a valores altos nas duas medidas do bico. Na amostra, seus bicos são, em média, mais compridos que os dos Adelie, mas têm profundidade muito semelhante. Em comparação com os Gentoo, a diferença mais evidente está na maior profundidade do bico.

Também podemos reconstruir as variáveis na escala física original a partir dos dois primeiros componentes. Desfazendo a padronização, a Tabela 4.4 compara as medidas originais com os valores reconstruídos para as três primeiras observações da amostra.

Tabela 4.4: Comparação entre valores originais e reconstruídos com 2 componentes principais para os 3 primeiros pinguins.

Pinguim	Medida	Original	Reconstruído (2 CPs)	Diferença
Pinguim 1	Comp. do bico	39.1	39.5	0.4
	Prof. do bico	18.7	18.7	0.0
	Comp. da nadadeira	181.0	186.0	5.0
	Massa corporal	3750.0	3395.5	-354.5
Pinguim 2	Comp. do bico	39.5	39.3	-0.2
	Prof. do bico	17.4	17.5	0.1
	Comp. da nadadeira	186.0	190.3	4.3
	Massa corporal	3800.0	3599.0	-201.0
Pinguim 3	Comp. do bico	40.3	40.0	-0.3
	Prof. do bico	18.0	18.0	0.0
	Comp. da nadadeira	195.0	189.8	-5.2
	Massa corporal	3250.0	3590.1	340.1

Como retemos 88.2% da variância acumulada com dois componentes, as medidas reconstruídas se aproximam bastante dos valores originais, descartando apenas o ruído e a variação secundária das duas últimas dimensões.

4.10 Exercícios

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

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

$$\text{Cov}(\mathbf{y}) = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Exercício 4.2. Um vendedor de ferramentas automotivas coletou dados a respeito de um determinado produto de seu portfólio, avaliando a qualidade das peças (X_1), a facilidade de uso (X_2) e a qualidade da embalagem (X_3). Após analisar os dados, ele calculou a seguinte matriz de covariância amostral

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

- a) Apenas olhando para essa matriz, quantos componentes você espera que tenham variância positiva? Como espera que as três variáveis participem desses componentes?
- b) Faça a decomposição da matriz para obter os componentes principais. Depois, compare os resultados com as suposições feitas no item (a).

Exercício 4.3. Dois grupos fizeram uma ACP sobre a mesma matriz de dados. Para um dos componentes, o primeiro grupo obteve o autovetor $\mathbf{e}_k$ e o vetor de escores $\mathbf{t}_k = \mathbf{X}\mathbf{e}_k$. O segundo encontrou $-\mathbf{e}_k$ e o vetor de escores com sinal invertido, $-\mathbf{t}_k$. Os gráficos produzidos pelos grupos aparecem refletidos, e cada grupo acredita que o outro cometeu um erro.

- a) Mostre que, se $\mathbf{e}_k$ é um autovetor de Σ associado a λ_k , então $-\mathbf{e}_k$ também é.
- b) Compare a variância explicada, as correlações $\text{Corr}(X_j, Y_k)$ entre as variáveis originais e o componente (Proposição 4.2) e os escores obtidos pelos dois grupos. O que muda e o que permanece igual?
- c) Afinal, os grupos encontraram componentes diferentes?

Exercício 4.4. Considere um vetor aleatório bidimensional $\mathbf{x} = (X_1, X_2)^T$, em que X_1 e X_2 são padronizadas e $\text{Corr}(X_1, X_2) = a$, com $0 < a \leq 1$.

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

Exercício 4.5. Dois pesquisadores aplicaram ACP à matriz de covariâncias

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

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

$$\mathbf{v}_1 = \begin{bmatrix} \cos \phi \\ \sin \phi \\ 0 \end{bmatrix} \quad \text{e} \quad \mathbf{v}_2 = \begin{bmatrix} -\sin \phi \\ \cos \phi \\ 0 \end{bmatrix}$$

- a) Verifique que $\mathbf{v}_1$ e $\mathbf{v}_2$ são autovetores unitários e ortogonais associados ao autovalor 4.
- b) Algum dos pesquisadores está errado? Explique como a repetição dos autovalores afeta a ACP.

Exercício 4.6. Considere a matriz de dados centralizados $\mathbf{X}$ ($n \times p$) e a aproximação de posto q dada por $\widehat{\mathbf{X}}_q = \mathbf{X}\mathbf{P}_q$, em que $\mathbf{P}_q = \mathbf{E}_q\mathbf{E}_q^T$ é construída a partir dos q primeiros autovetores ortonormais de $\mathbf{S}$.

- a) Mostre que a matriz $\mathbf{P}_q$ é simétrica e idempotente ($\mathbf{P}_q^2 = \mathbf{P}_q$). Qual é a interpretação geométrica dessa transformação sobre as observações?
- b) Mostre que a matriz residual de reconstrução, $\mathbf{X} - \widehat{\mathbf{X}}_q$, é ortogonal às direções mantidas, isto é, $(\mathbf{X} - \widehat{\mathbf{X}}_q)\mathbf{E}_q = \mathbf{0}$. O que isso indica sobre a informação descartada?

Exercício 4.7 (Composição química de vinhos). O conjunto [Wine, da UCI Machine Learning Repository](#), reúne 178 vinhos produzidos na mesma região da Itália a partir de três cultivares. Para cada vinho, foram registradas 13 medidas de sua composição química. Que diferenças na composição química uma ACP permite identificar entre esses vinhos? Os três cultivares se distinguem no espaço dos componentes? Justifique todas as decisões tomadas.

4.11 Tópicos avançados

Até agora, esse capítulo apresentou a teoria clássica da ACP. Agora, vamos explorar algumas técnicas avançadas e discussões que não cabem no primeiro contato com a técnica. Esta seção apresenta extensões da ACP com um menor nível de detalhe.

4.11.1 Soluções alternativas da ACP

A solução usual da ACP pode ser obtida pela decomposição em valores singulares (SVD), definida em Definição 2.14, da matriz de dados centralizados. Isso elimina a etapa de cálculo da matriz de covariância e também garante melhor estabilidade numérica. Essa estratégia é usada por boa parte das bibliotecas, como o `scikit-learn`. A seguir, formalizamos essa solução.

Proposição 4.4 (ACP pela SVD). *Seja $\mathbf{X}$ ($n \times p$) a matriz de dados centralizados e $\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T$ sua SVD completa, com $\mathbf{U}$ ($n \times n$) e $\mathbf{V}$ ($p \times p$) ortogonais. A matriz $\mathbf{\Delta}$ ($n \times p$) contém os valores singulares positivos $\delta_1 \geq \dots \geq \delta_r > 0$ na diagonal e zeros nas demais posições, em que $r \leq \min(n-1, p)$ é o posto de $\mathbf{X}$. Sejam $\mathbf{E}$ e $\mathbf{\Lambda}$ as matrizes de autovetores e autovalores de $\mathbf{S}$, respectivamente, e $\mathbf{T} = \mathbf{X}\mathbf{E}$ a matriz de escores. Segue que,*

$$\mathbf{E} = \mathbf{V}, \quad \mathbf{\Lambda} = \frac{1}{n-1} \mathbf{\Delta}^T \mathbf{\Delta}, \quad \mathbf{T} = \mathbf{U}\mathbf{\Delta}$$

isto é, as direções principais são as colunas de $\mathbf{V}$. Para $k \leq r$, temos $\lambda_k = \delta_k^2/(n-1)$ e $\mathbf{t}_k = \delta_k \mathbf{u}_k$. Os demais autovalores e escores são nulos.

Prova. Como as colunas de $\mathbf{U}$ são ortonormais e $\mathbf{\Delta}^T \mathbf{\Delta} = \text{diag}(\delta_1^2, \dots, \delta_r^2, 0, \dots, 0)$,

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}^T \mathbf{X} = \frac{1}{n-1} \mathbf{V} \mathbf{\Delta}^T \mathbf{U}^T \mathbf{U} \mathbf{\Delta} \mathbf{V}^T = \mathbf{V} \left(\frac{\mathbf{\Delta}^T \mathbf{\Delta}}{n-1} \right) \mathbf{V}^T$$

é uma decomposição espectral de $\mathbf{S}$, com $\mathbf{V}$ ortogonal e autovalores positivos $\delta_k^2/(n-1)$ em ordem decrescente, seguidos dos autovalores nulos. Podemos, portanto, escolher $\mathbf{E} = \mathbf{V}$ e $\mathbf{\Lambda} = \mathbf{\Delta}^T \mathbf{\Delta}/(n-1)$. Os escores seguem de $\mathbf{T} = \mathbf{X}\mathbf{E} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \mathbf{V} = \mathbf{U}\mathbf{\Delta}$. $\square$

A proposição mostra a equivalência entre a solução por decomposição espectral vista no capítulo e a solução por SVD. Ela explica também por que o Teorema de Eckart-Young, que é um resultado teórico sobre a SVD, pôde ser usado na Proposição 4.3, já que a soma de matrizes de posto um $\sum_k \mathbf{t}_k \mathbf{e}_k^T$ é exatamente a SVD truncada de $\mathbf{X}$.

Corolário 4.1 (ACP pela matriz de produtos internos). *Nas condições da Proposição 4.4, seja $\mathbf{X}\mathbf{X}^T$ a matriz $n \times n$ de produtos internos entre observações, cujas entradas são $\mathbf{x}_i^T \mathbf{x}_j$. Então*

$$\mathbf{X}\mathbf{X}^T = \mathbf{U} (\mathbf{\Delta}\mathbf{\Delta}^T) \mathbf{U}^T$$

é uma decomposição espectral de $\mathbf{X}\mathbf{X}^T$.

Prova. Como $\mathbf{V}$ é ortogonal, $\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \mathbf{V} \mathbf{\Delta}^T \mathbf{U}^T = \mathbf{U} (\mathbf{\Delta}\mathbf{\Delta}^T) \mathbf{U}^T$, com $\mathbf{U}$ ortogonal e $\mathbf{\Delta}\mathbf{\Delta}^T$ diagonal, contendo $\delta_1^2 \geq \dots \geq \delta_r^2 > 0$ e zeros nas posições restantes. Os autovalores positivos de $\mathbf{X}\mathbf{X}^T$ são, portanto, os δ_k^2 , e pela Proposição 4.4 valem $\delta_k^2 = (n-1)\lambda_k$. Os escores seguem de $\mathbf{T} = \mathbf{U}\mathbf{\Delta}$, cuja k -ésima coluna é $\delta_k \mathbf{u}_k$ para $k \leq r$, sendo nulas as demais. $\square$

Note que a matriz de escores da proposição Proposição 4.4 depende exclusivamente de $\mathbf{\Delta}$ e $\mathbf{U}$, e o corolário mostra que essas matrizes podem ser obtidas pela decomposição de $\mathbf{X}\mathbf{X}^T$. Ou seja, a ACP pela matriz de produtos internos é um atalho para obter os escores, sem passar pelas projeções.

4.11.2 ACP com kernel

A ACP busca um subespaço linear que preserve a variância dos dados. Quando a estrutura dos dados é curva, uma redução linear pode não descrevê-la bem. Uma alternativa é usar a ACP com kernel. Ela aproveita o Corolário 4.1, segundo o qual os escores dependem da amostra apenas pelos produtos internos $\mathbf{x}_i^T \mathbf{x}_j$.

O **truque do kernel** consiste em trocar o produto interno usual por uma função de **kernel** $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, que mede produtos internos em um espaço transformado de dimensão muito maior, sem precisar construí-lo. Nesse espaço, relações não lineares passam a ser tratadas como lineares. A escolha usual para capturar estruturas curvas é o kernel gaussiano

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

onde γ é um parâmetro fixado. Basta então montar e decompor a matriz de kernel $\mathbf{K}$, com $k_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, devidamente centralizada. Os escores no espaço do kernel são dados por $\mathbf{T} = \mathbf{U}\mathbf{\Delta}$, onde $\mathbf{U}$ contém os autovetores da matriz de kernel centralizada e $\mathbf{\Delta} = \text{diag}(\delta_1, \dots, \delta_r)$ reúne as raízes quadradas de seus autovalores.

i Nota

A ACP com o kernel linear $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, recupera exatamente a ACP linear usual deste capítulo.

A Figura 4.8 mostra um caso em que a diferença é gritante. Os dados são dois anéis concêntricos de raios 1 e 3, e a informação que interessa é justamente a qual anel cada ponto pertence.

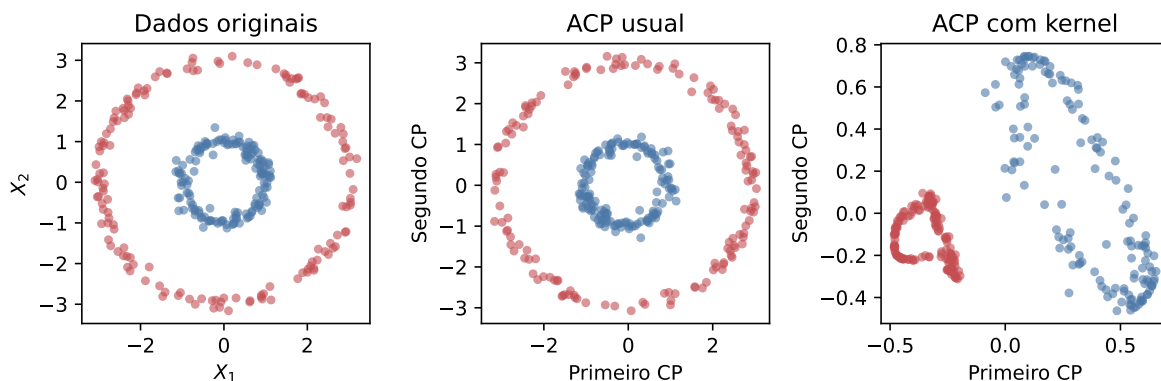


Figura 4.8: Dois anéis concêntricos, os escores da ACP usual e os escores da ACP com kernel gaussiano ($\gamma = 0,5$).

Os dois anéis têm o mesmo centro e, em cada anel, a variância populacional é a mesma em todas as direções. Uma projeção linear sobre um único eixo produz sobreposição entre os grupos. Os escores do painel central são apenas os dados originais girados. Já no espaço induzido pelo kernel gaussiano os anéis ficam separados sobre o primeiro componente, e uma única coordenada passa a distinguir perfeitamente os dois grupos.

Na ACP com kernel, os componentes são lineares no espaço transformado, mas em geral são funções não lineares das variáveis originais, calculadas por meio do kernel. Assim, não temos cargas lineares dessas variáveis com a interpretação usual da ACP.

4.11.3 Estabilidade da solução

Os componentes de uma amostra são estimativas, e como toda estimativa variam de amostra para amostra. Sob normalidade, há um resultado assintótico que descreve essa variação para as direções estimadas.

Proposição 4.5 (Distribuição assintótica dos autovetores). *Seja $x_1, \dots, x_n$ uma amostra aleatória de uma $N_p(\mu, \Sigma)$, em que Σ tem autovalores distintos $\lambda_1 > \dots > \lambda_p > 0$ e autovetores ortonormais associados $e_1, \dots, e_p$. Sejam $\hat{e}_1, \dots, \hat{e}_p$ os autovetores de S , com o sinal fixado de modo que $\hat{e}_k^T e_k > 0$. Então, para cada k , a distribuição de $\sqrt{n}(\hat{e}_k - e_k)$ converge, quando $n \rightarrow \infty$, para uma $N_p(0, \Omega_k)$, em que*

$$\Omega_k = \lambda_k \sum_{j \neq k} \frac{\lambda_j}{(\lambda_j - \lambda_k)^2} e_j e_j^T$$

Não vamos demonstrar a proposição, que depende de ferramentas assintóticas fora do escopo do livro. O que interessa aqui é como ler Ω_k . Para n grande, $\text{Cov}(\hat{e}_k) \approx \Omega_k/n$, e cada parcela dessa soma traz a distância entre dois autovalores elevada ao quadrado no denominador. Quando λ_j está próximo de λ_k , a variância de $\hat{e}_k$ na direção de e_j explode, e a direção estimada gira livremente dentro do plano gerado por e_j e e_k , passando a refletir mais o ruído amostral do que a estrutura da população. A hipótese de autovalores distintos, aliás, garante a unicidade das direções, a menos do sinal, e o Exercício 4.5 mostra que, com $\lambda_j = \lambda_k$, qualquer base do plano serve. A proposição diz que a transição entre os dois cenários é contínua, e que autovalores apenas próximos já produzem soluções instáveis.

Para visualizar essa incerteza sem recorrer a aproximações assintóticas, o caminho mais simples é reamostrar os dados, recalculando as cargas e desenhando os intervalos resultantes sobre o círculo de correlações, tomando o cuidado de alinhar o sinal de cada réplica à solução original, pelo motivo discutido no Exercício 4.3.

Autovalores próximos são uma causa de instabilidade difícil de perceber a olho nu. Uma segunda causa, bem mais visível, são os outliers. A Definição 4.1 constrói os componentes maximizando a variância, e a variância é um dos funcionais menos resistentes da estatística. A Figura 4.9 retoma a amostra de atrasos do Exemplo 4.3 e acrescenta oito observações contaminadas, cerca de 5% do total.

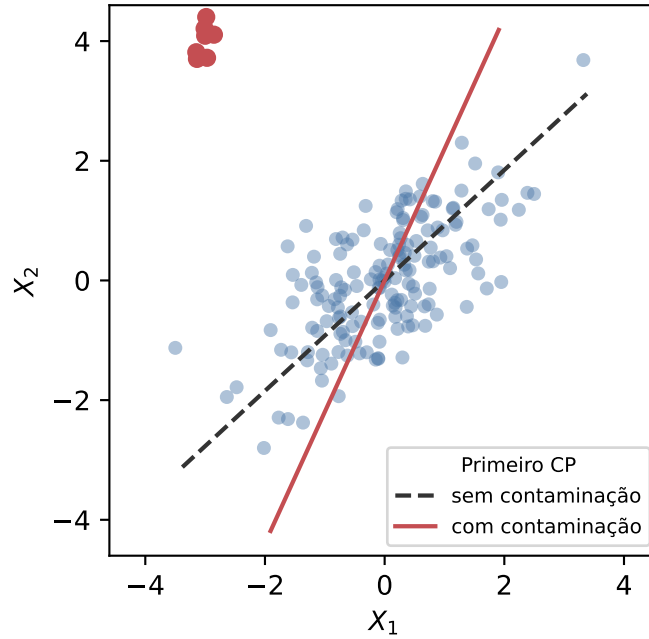


Figura 4.9: Efeito de 8 observações contaminadas sobre a direção do primeiro componente principal.

A direção do primeiro componente gira de 43° para 66° , e o segundo autovalor salta de 0.36 para 1.42. O problema está no próprio funcional maximizado, e os métodos de **ACP robusta** substituem a variância por uma medida de escala resistente σ_{rob} , como o desvio absoluto mediano

$$\max_{\mathbf{u}} \sigma_{\text{rob}}(\mathbf{u}^T \mathbf{x}) \quad \text{sujeito a} \quad \mathbf{u}^T \mathbf{u} = 1$$

4.11.4 Análise paralela

Mesmo sob independência completa entre as variáveis, os autovalores amostrais não são todos iguais a 1: as correlações amostrais não são exatamente nulas, e a ACP se aproveita delas. A Figura 4.10 mostra o efeito em $p = 20$ variáveis independentes observadas em $n = 100$ indivíduos, isto é, em ruído puro.

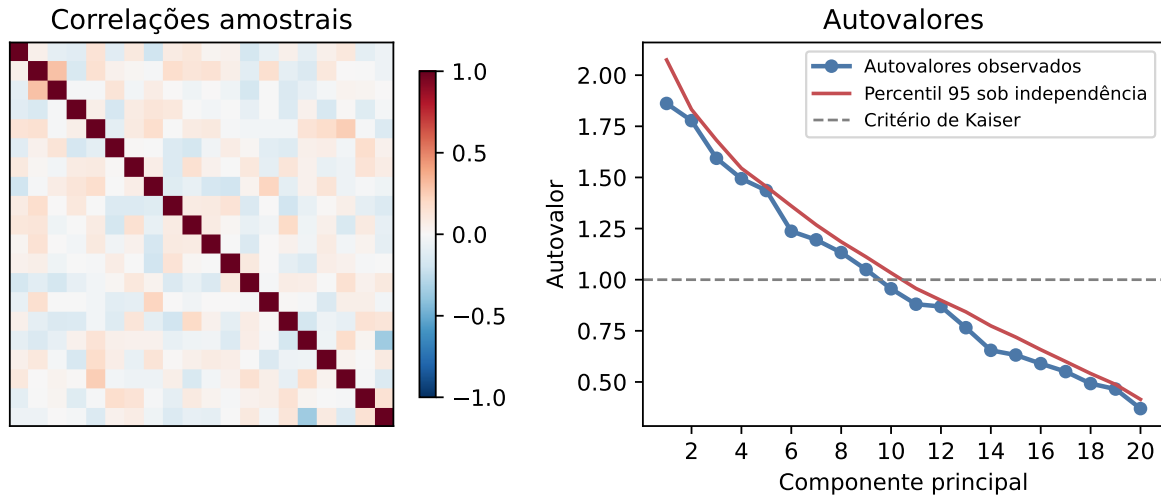


Figura 4.10: À esquerda, as correlações amostrais de 20 variáveis independentes ($n = 100$). À direita, os autovalores correspondentes, comparados ao critério de Kaiser e à distribuição nula obtida por permutação.

Fora da diagonal, as correlações são pequenas mas não nulas, e isso basta para que 9 autovalores fiquem acima de 1. Em dados sem estrutura alguma, o critério de Kaiser reteria quase metade dos componentes e o gráfico do cotovelo daria margem a qualquer leitura. Esses critérios descrevem a distribuição da variância, mas não a comparam com uma referência obtida sob independência.

A **análise paralela** faz essa comparação usando os próprios dados. Permuta-se cada coluna de forma independente, o que destrói as correlações e preserva as distribuições marginais, e recalculam-se os autovalores muitas vezes. O resultado é a distribuição dos autovalores que se observaria sob independência, e retêm-se apenas os componentes que superam um percentil alto dessa distribuição, como o percentil 95 desenhado na figura. Nesta simulação de ruído puro, o critério não retém nenhum componente. Esse resultado não é garantido em toda amostra, pois as comparações também estão sujeitas à variação amostral.

Podemos agora refazer a escolha de q do Exemplo 4.6 com esse critério.

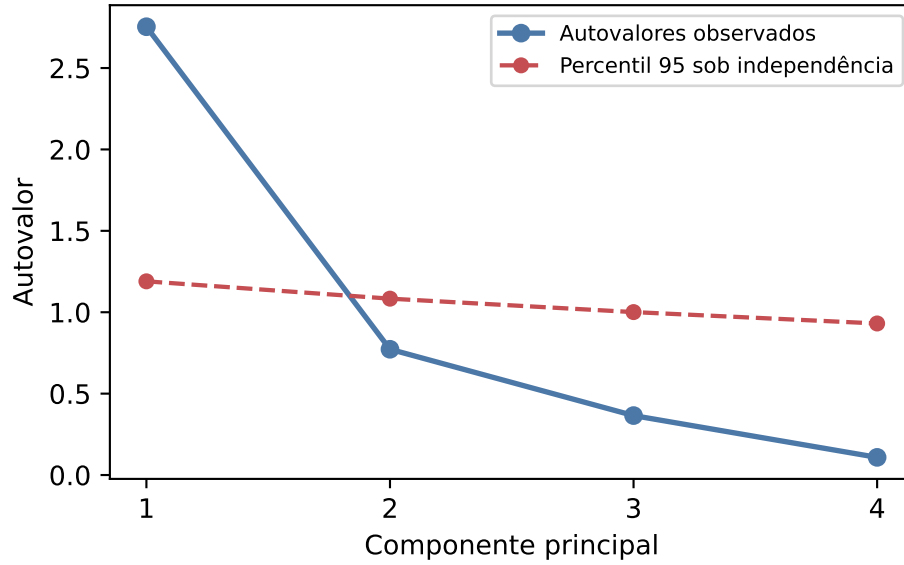


Figura 4.11: Análise paralela aplicada às medidas corporais dos pinguins.

Apenas o primeiro componente supera o percentil 95 dos autovalores correspondentes nas permutações, de modo que a análise paralela sugere reter um componente. O segundo, que responde por 19.3% da variância total, não supera esse limiar, mas isso não demonstra que ele seja apenas ruído. CP2 continua útil para descrever diferenças nas medidas do bico e construir a representação bidimensional.

4.11.5 Outras reduções de dimensão

A **ACP probabilística** transforma em modelo a separação entre variância retida e variância descartada da Proposição 4.3. Ao reter q componentes, a ACP usual apenas abandona os $p - q$ últimos autovalores; aqui supõe-se que todos eles valem o mesmo, um patamar de ruído σ^2 idêntico em todas as direções não retidas. Em termos da matriz de covariâncias,

$$\Sigma = E_q (\Lambda_q - \sigma^2 I_q) E_q^T + \sigma^2 I_p$$

com $\Lambda_q = \text{diag}(\lambda_1, \dots, \lambda_q)$. Somando a hipótese de normalidade, $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, o modelo ganha uma verossimilhança, e a estimação por máxima verossimilhança devolve as direções principais de sempre e $\hat{\sigma}^2 = \sum_{k>q} \lambda_k / (p - q)$, o erro de reconstrução repartido igualmente entre as direções abandonadas.

Com um modelo probabilístico, ganhamos critérios como o AIC e o BIC para escolher q , a estimação por algoritmo EM, que acomoda valores ausentes, e a possibilidade de simular observações ou usar o

modelo como componente de uma mistura. Permitir que cada variável tenha o seu próprio nível de ruído, trocando $\sigma^2 \mathbf{I}_p$ por uma matriz diagonal qualquer, leva à análise fatorial, tema do Capítulo 6.

A **ACP esparsa** (*sparse PCA*) é uma alternativa mais interpretável. Como os autovetores raramente têm coordenadas nulas, cada componente é uma combinação de todas as p variáveis, o que dificulta a leitura quando p é grande. Zerar as cargas pequenas a olho quebra a ortogonalidade e a otimalidade da solução; a ACP esparsa faz isso de maneira controlada, acrescentando uma restrição ℓ_1 ao problema de maximização original

$$\max_{\mathbf{u}} \mathbf{u}^T \mathbf{S} \mathbf{u} \quad \text{sujeito a} \quad \|\mathbf{u}\|_2 = 1, \quad \|\mathbf{u}\|_1 \leq c$$

Quanto menor c , mais coordenadas são levadas exatamente a zero. Em troca, os componentes resultantes deixam de ser ortogonais e não correlacionados, e a própria noção de variância explicada precisa ser redefinida.

O **t-SNE** e o **UMAP** abandonam a ideia de projeção. Em vez de procurar um subespaço, procuram diretamente uma configuração de pontos em duas dimensões que preserve as vizinhanças locais dos dados. São excelentes para visualizar agrupamentos, mas não produzem cargas nem matriz de projeção, dependem fortemente de hiperparâmetros e são estocásticos, de modo que execuções distintas geram figuras distintas. Sobretudo, as distâncias entre grupos bem separados na figura e os tamanhos relativos dos agrupamentos não são interpretáveis.

5 Análise de correspondência

A Análise de Correspondência (AC) é uma família de técnicas exploratórias para a análise de dados categóricos. Ela facilita a visualização das associações existentes entre os níveis de diferentes variáveis categóricas. A **Análise de Correspondência Simples (ACS)** trabalha com duas variáveis categóricas a partir de uma tabela de contingência. A extensão para três ou mais variáveis é a **Análise de Correspondência Múltipla (ACM)**.

A análise de correspondência costuma ser apresentada como uma “ACP para dados categóricos”, porque também busca uma representação de baixa dimensão que possa ser desenhada e interpretada. A principal saída de uma análise de correspondência é o **mapa de correspondência**, um gráfico em que níveis com perfis de frequência semelhantes aparecem próximos uns dos outros e níveis com perfis distintos aparecem afastados. É lendo esse mapa que identificamos padrões de associação, como quais níveis de uma variável tendem a aparecer junto com quais níveis de outra.

Definição 5.1 (Tabela de contingência). Sejam X e Y duas variáveis categóricas com I e J níveis, respectivamente. Uma **tabela de contingência** N de dimensão $I \times J$ é uma matriz em que cada elemento n_{ij} representa a contagem de observações que pertencem simultaneamente ao i -ésimo nível de X e ao j -ésimo nível de Y .

As somas nas margens da tabela fornecem os totais de cada nível, com $n_{i.} = \sum_{j=1}^J n_{ij}$ para a i -ésima linha e $n_{.j} = \sum_{i=1}^I n_{ij}$ para a j -ésima coluna. A soma de todas as contagens resulta no número total de observações da amostra, $n = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$.

Definição 5.2 (Matriz de correspondência). A **matriz de correspondência**, denotada por P , reúne as proporções amostrais conjuntas obtidas pela divisão de cada elemento de N pelo total geral de observações n

$$P = \frac{1}{n} N$$

Cada elemento $p_{ij} = n_{ij}/n$ representa a frequência relativa conjunta dos níveis i e j . As proporções marginais das linhas, $r_i = \sum_{j=1}^J p_{ij}$, e das colunas, $c_j = \sum_{i=1}^I p_{ij}$, são reunidas nos vetores

$$\mathbf{r} = (r_1, \dots, r_I)^T \quad \text{e} \quad \mathbf{c} = (c_1, \dots, c_J)^T$$

Exemplo 5.1 (Preferências por serviços de streaming). Considere uma pesquisa sobre a preferência por serviços de *streaming*. Foram entrevistadas 300 pessoas, classificadas pela faixa etária e pelo serviço escolhido. Os resultados estão na Tabela 5.1.

Tabela 5.1: Contagens observadas de preferência por serviço de *streaming* segundo a faixa etária.

Faixa etária	Netflix	Prime Video	Max	Disney+	Total
Jovem	45	15	10	30	100
Adulto	30	45	15	10	100
Idoso	15	20	55	10	100
Total	90	80	80	50	300

Dividindo cada contagem da Tabela 5.1 pelo total de $n = 300$ observações, obtemos a matriz de correspondência $\mathbf{P}$

$$\mathbf{P} = \frac{1}{300}\mathbf{N} = \begin{pmatrix} 45/300 & 15/300 & 10/300 & 30/300 \\ 30/300 & 45/300 & 15/300 & 10/300 \\ 15/300 & 20/300 & 55/300 & 10/300 \end{pmatrix} = \begin{pmatrix} 0,150 & 0,050 & 0,033 & 0,100 \\ 0,100 & 0,150 & 0,050 & 0,033 \\ 0,050 & 0,067 & 0,183 & 0,033 \end{pmatrix} \quad (5.1)$$

cujas somas marginais fornecem os vetores $\mathbf{r}$ e $\mathbf{c}$

$$\mathbf{r} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \quad \text{e} \quad \mathbf{c} = \begin{pmatrix} 90/300 \\ 80/300 \\ 80/300 \\ 50/300 \end{pmatrix} \approx \begin{pmatrix} 0,300 \\ 0,267 \\ 0,267 \\ 0,167 \end{pmatrix}$$

As três faixas etárias têm a mesma proporção de pessoas na amostra ($r_i = 1/3$ para cada). Na amostra completa, 30% preferem Netflix ($c_1 = 0,300$), 26,7% preferem Prime Video ($c_2 \approx 0,267$), 26,7% preferem Max ($c_3 \approx 0,267$) e 16,7% preferem Disney+ ($c_4 \approx 0,167$). Essas proporções descrevem a popularidade geral de cada serviço, mas as preferências mudam bastante de uma faixa etária para outra. Entre os jovens, 45% escolhem Netflix e 30% Disney+. Entre os adultos, 45% ficam com o Prime Video, e entre os idosos 55% escolhem a Max.

Nosso objetivo será transformar essas comparações isoladas em uma descrição conjunta das associações presentes na tabela.

5.1 O modelo de independência

Definição 5.3 (Independência em tabelas de contingência). Duas variáveis categóricas são independentes se as probabilidades conjuntas forem iguais ao produto das probabilidades marginais, $p_{ij} = r_i c_j$ para todos os pares (i, j) . Em notação matricial,

$$\mathbf{P} = \mathbf{r} \mathbf{c}^T$$

Quando há dependência, o afastamento em relação a essa situação fica registrado na matriz de desvios

$$\mathbf{\Delta} = \mathbf{P} - \mathbf{r} \mathbf{c}^T$$

Cada elemento $\delta_{ij} = p_{ij} - r_i c_j$ compara a proporção observada com sua proporção esperada sob independência. Valores positivos indicam que a combinação aparece mais do que o esperado, e valores negativos, que aparece menos. Sob independência, $\mathbf{\Delta} = \mathbf{0}$.

Uma maneira ingênua de resumir o desvio total seria a soma dos quadrados de seus elementos, $\sum_{i=1}^I \sum_{j=1}^J (p_{ij} - r_i c_j)^2$. Essa soma, porém, ignora a frequência esperada de cada célula, e um mesmo desvio pesa mais numa célula rara do que numa célula frequente. Tome por exemplo a célula *Jovem-Disney+* da matriz na Equação 5.1. Sob independência, esperaríamos a proporção $r_1 c_4 = \frac{1}{3} \times 0,167 \approx 0,056$, e por isso um desvio de 0,05 nessa célula já é forte, representando cerca de 90% de respostas a mais do que o esperado.

Por outro lado, a célula *Jovem-Netflix* tem proporção esperada $r_1 c_1 = \frac{1}{3} \times 0,300 = 0,100$. Nesse caso, o mesmo desvio de 0,05 representaria 50% de respostas a mais do que o esperado nessa combinação. A solução é ponderar o quadrado de cada desvio pelo inverso de sua proporção esperada sob independência. Em notação matricial, isso pede as matrizes diagonais das marginais

$$\mathbf{D}_r = \text{diag}(\mathbf{r}) \quad \text{e} \quad \mathbf{D}_c = \text{diag}(\mathbf{c})$$

Assumimos que $r_i > 0$ e $c_j > 0$ para todos os níveis. Um nível sem observações na amostra deve ser removido da tabela antes do ajuste.

Definição 5.4 (Matriz de desvios padronizados). A **matriz de desvios padronizados** é definida por

$$\mathbf{S} = \mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r} \mathbf{c}^T) \mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1/2} \mathbf{\Delta} \mathbf{D}_c^{-1/2}$$

Elemento a elemento, isso dá

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

A multiplicação à esquerda por $D_r^{-1/2}$ divide cada linha por $\sqrt{r_i}$, enquanto a multiplicação à direita por $D_c^{-1/2}$ divide cada coluna por $\sqrt{c_j}$. O sinal do desvio é mantido, mas sua magnitude passa a ser avaliada em relação à variabilidade esperada da célula.

Definição 5.5 (Inércia total). A **inércia total** da tabela de contingência é a soma dos quadrados dos desvios padronizados, coincidindo com o quadrado da norma de Frobenius (Definição 2.6) da matriz S

$$\phi^2 = \sum_{i=1}^I \sum_{j=1}^J s_{ij}^2 = \text{tr}(S^T S) = \|S\|_F^2$$

Como soma de quadrados, $\phi^2 \geq 0$, e $\phi^2 = 0$ se, e somente se, $p_{ij} = r_i c_j$ para todas as células. A inércia total serve como uma medida padronizada de quão distante a distribuição conjunta está da hipótese de independência.

Relação com o teste qui-quadrado de independência

A hipótese de independência em tabelas de contingência costuma ser avaliada pela estatística qui-quadrado de Pearson (χ^2). A inércia total nada mais é do que essa estatística dividida pelo tamanho amostral

$$\phi^2 = \frac{\chi^2}{n} \quad \text{ou} \quad \chi^2 = n\phi^2$$

A inércia total funciona, então, como uma medida de associação que não depende do tamanho amostral. O teste responde com um sim ou um não global sobre a existência de dependência. A análise de correspondência vai além e decompõe essa mesma associação para mostrar, no mapa, quais níveis se atraem e quais se repelem.

5.2 Perfis e distâncias

Definição 5.6 (Perfis de linha e de coluna). Dada a matriz de correspondência P com vetores marginais r e c , os **perfis-linha** são as linhas da matriz $D_r^{-1}P$ e os **perfis-coluna** são as colunas da matriz PD_c^{-1} .

O perfil do nível-linha i tem entradas p_{ij}/r_i , para $j = 1, \dots, J$, e contém a distribuição condicional das colunas para esse nível. De forma análoga, o perfil do nível-coluna j tem entradas p_{ij}/c_j , para $i = 1, \dots, I$, com a distribuição condicional das linhas.

O **perfil médio** das linhas é a média dos perfis-linha ponderada pelas marginais r_i . Como $\sum_i r_i (p_{ij}/r_i) = \sum_i p_{ij} = c_j$, ele coincide com a marginal das colunas c^T . Da mesma forma, o perfil médio das colunas é a marginal das linhas r .

Exemplo 5.2. No Exemplo 5.1, a matriz de perfis-linha $D_r^{-1}P$ é obtida dividindo cada linha pelo seu total marginal (100). Como as três faixas etárias têm a mesma marginal, o perfil médio é a média simples das três linhas, e na Netflix, por exemplo, $(0,45 + 0,30 + 0,15)/3 = 0,30 = c_1$. Os perfis-linha e o perfil médio são

$$D_r^{-1}P = \begin{pmatrix} 0,45 & 0,15 & 0,10 & 0,30 \\ 0,30 & 0,45 & 0,15 & 0,10 \\ 0,15 & 0,20 & 0,55 & 0,10 \end{pmatrix} \quad \text{e} \quad c^T \approx (0,300 \quad 0,267 \quad 0,267 \quad 0,167)$$

A primeira linha de $D_r^{-1}P$ descreve o perfil dos jovens, dos quais 45% preferem Netflix, 15% Prime Video, 10% Max e 30% Disney+. De forma análoga, dividindo cada coluna da tabela pelo seu total marginal, obtemos a matriz de perfis-coluna PD_c^{-1} , cujo perfil médio é r

$$PD_c^{-1} \approx \begin{pmatrix} 0,500 & 0,188 & 0,125 & 0,600 \\ 0,333 & 0,562 & 0,188 & 0,200 \\ 0,167 & 0,250 & 0,688 & 0,200 \end{pmatrix} \quad \text{e} \quad r \approx \begin{pmatrix} 0,33 \\ 0,33 \\ 0,33 \end{pmatrix}$$

Entre os assinantes da Netflix, por exemplo, 50% são jovens, 33,3% adultos e 16,7% idosos, como mostra a primeira coluna de PD_c^{-1} .

Fica mais fácil comparar os perfis colocando cada um ao lado do perfil médio. Na Figura 5.1, se a faixa etária não influenciasse a escolha, as barras de um mesmo serviço teriam todas a altura da barra do perfil médio, e estão longe disso. Na Max, a barra dos idosos chega a 0,55, mais que o dobro da média de 0,27, e na Netflix os jovens ficam em 0,45 contra 0,30. A Figura 5.2 conta a mesma história do outro lado da tabela. Entre os assinantes da Max, 69% são idosos, enquanto na amostra toda cada faixa etária responde por um terço.

Proposição 5.1 (Perfis sob a hipótese de independência). *Duas variáveis categóricas são independentes se, e somente se, todos os perfis-linha forem idênticos ao perfil médio das linhas ou, equivalentemente, todos os perfis-coluna forem idênticos ao perfil médio das colunas*

$$P = rc^T \iff D_r^{-1}P = 1c^T \iff PD_c^{-1} = r1^T$$

Prova. Sob independência, $P = rc^T$. Como $D_r^{-1}r = 1$, temos

$$D_r^{-1}P = (D_r^{-1}r) c^T = 1c^T$$

Reciprocamente, se $D_r^{-1}P = 1c^T$, multiplicando à esquerda por D_r obtemos $P = D_r 1c^T = rc^T$. O resultado para perfis-coluna é análogo. $\square$

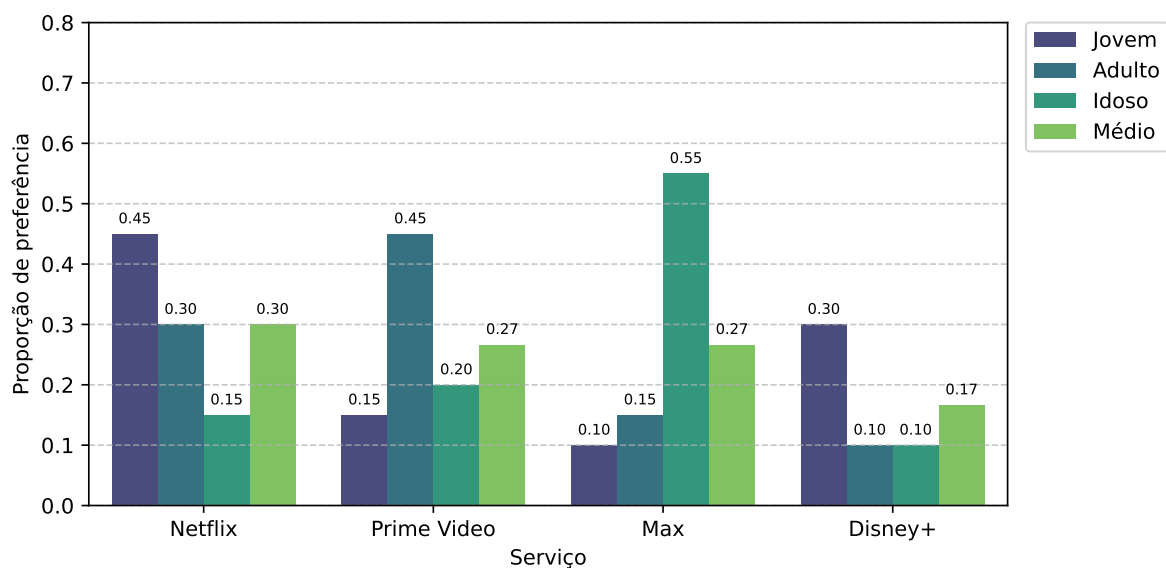


Figura 5.1: Perfis-linha das faixas etárias comparados ao perfil médio. Diferenças de altura entre as barras de um mesmo serviço indicam associação entre as variáveis.

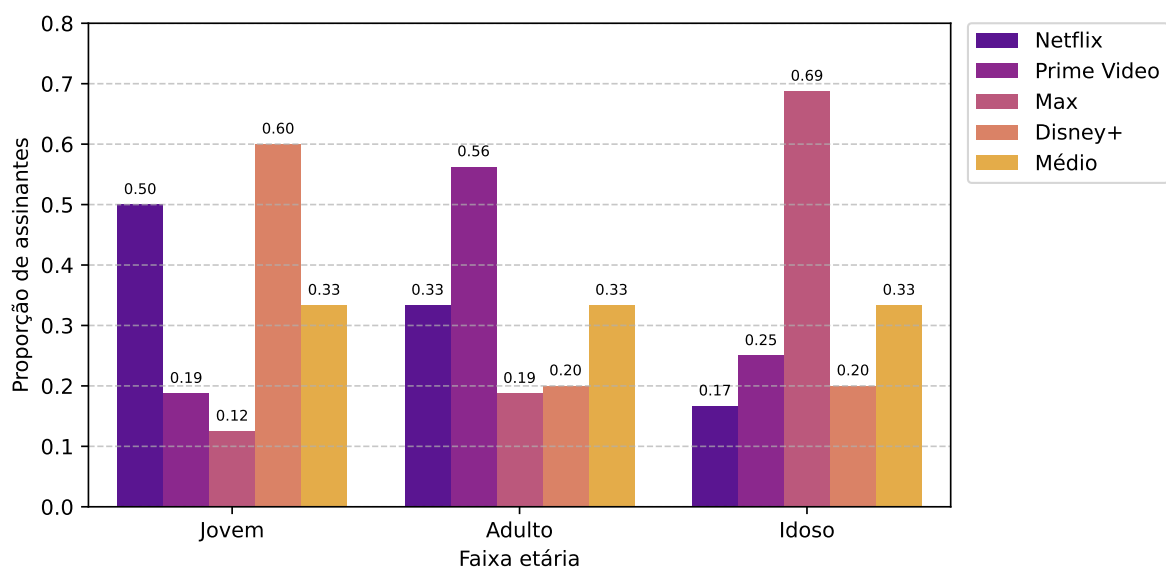


Figura 5.2: Perfis-coluna dos serviços comparados ao perfil médio. As diferenças mostram como a composição etária varia entre os assinantes de cada serviço.

Para comparar dois perfis, a distância euclidiana simples não é adequada, pois trata níveis frequentes e raros da mesma forma. A métrica natural é a distância qui-quadrado, que, pelo mesmo motivo que nos levou a padronizar os desvios, pondera a diferença em cada coordenada pelo inverso da marginal.

Definição 5.7 (Distância qui-quadrado). A **distância qui-quadrado** (χ^2) entre dois níveis-linha i e i' e entre dois níveis-coluna j e j' é definida por

$$d_{\chi^2}^2(i, i') = \sum_{j=1}^J \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2 \quad \text{e} \quad d_{\chi^2}^2(j, j') = \sum_{i=1}^I \frac{1}{r_i} \left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}} \right)^2$$

Trocando o segundo perfil pelo perfil médio, obtemos a distância qui-quadrado de cada perfil ao respectivo perfil médio

$$d_{\chi^2}^2(i, \mathbf{c}) = \sum_{j=1}^J \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - c_j \right)^2 \quad \text{e} \quad d_{\chi^2}^2(j, \mathbf{r}) = \sum_{i=1}^I \frac{1}{r_i} \left(\frac{p_{ij}}{c_j} - r_i \right)^2$$

Proposição 5.2 (Inércia total como dispersão dos perfis). A *inércia total* ϕ^2 é exatamente a média ponderada das distâncias qui-quadrado de todos os perfis em relação ao perfil médio

$$\phi^2 = \sum_{i=1}^I r_i d_{\chi^2}^2(i, \mathbf{c}) = \sum_{j=1}^J c_j d_{\chi^2}^2(j, \mathbf{r})$$

Prova. Colocando r_i em evidência no numerador da Definição 5.4, cada desvio padronizado passa a ser escrito em termos do perfil da linha i

$$s_{ij} = \frac{r_i \left(\frac{p_{ij}}{r_i} - c_j \right)}{\sqrt{r_i c_j}} = \sqrt{\frac{r_i}{c_j}} \left(\frac{p_{ij}}{r_i} - c_j \right)$$

Somando os quadrados primeiro em j e depois em i , obtemos

$$\phi^2 = \sum_{i=1}^I \sum_{j=1}^J s_{ij}^2 = \sum_{i=1}^I r_i \underbrace{\sum_{j=1}^J \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - c_j \right)^2}_{d_{\chi^2}^2(i, \mathbf{c})}$$

Para os perfis-coluna, colocamos c_j em evidência, o que dá $s_{ij} = \sqrt{c_j/r_i} (p_{ij}/c_j - r_i)$, e somamos primeiro em i . □

A inércia total tem, portanto, duas leituras equivalentes. Ela mede o afastamento da tabela em relação à independência e, ao mesmo tempo, a dispersão dos perfis em torno do perfil médio. Os perfis de uma tabela $I \times J$ vivem em um espaço de dimensão alta demais para desenhar, e o próximo passo é encontrar as poucas direções que concentram a maior parte dessa dispersão.

5.3 Análise de correspondência simples

A Análise de Correspondência Simples procura o subespaço de baixa dimensão que fica mais perto da nuvem de perfis, no mesmo espírito da ACP. Como a dispersão dos perfis é a inércia total, e a inércia total é $\|\mathbf{S}\|_F^2$, isso equivale a buscar uma aproximação de baixo posto para a matriz de desvios padronizados $\mathbf{S}$. Pelo Teorema de Eckart-Young-Mirsky (Teorema 2.2), a melhor aproximação de posto q segundo a norma de Frobenius é obtida truncando a decomposição em valores singulares (SVD), e minimizar o erro quadrático $\|\mathbf{S} - \widehat{\mathbf{S}}_q\|_F^2$ equivale a maximizar a inércia explicada pelas q primeiras dimensões. Sob independência, todos os perfis coincidem com o perfil médio e $\mathbf{S} = \mathbf{0}$, então não há nada a representar.

Como as linhas e colunas de $\mathbf{P} - \mathbf{r}\mathbf{c}^T$ somam zero, o posto $K = \text{posto}(\mathbf{S})$ não passa de $\min(I - 1, J - 1)$. A decomposição reduzida de $\mathbf{S}$ em valores singulares é

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \sum_{k=1}^K \lambda_k \mathbf{u}_k \mathbf{v}_k^T$$

em que $\mathbf{U} = (\mathbf{u}_1 \cdots \mathbf{u}_K)$ e $\mathbf{V} = (\mathbf{v}_1 \cdots \mathbf{v}_K)$ contêm os vetores singulares ortonormais à esquerda e à direita, e $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_K)$ reúne os valores singulares em ordem decrescente, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K > 0$.

Escrevendo a decomposição elemento a elemento, $s_{ij} = \sum_{k=1}^K \lambda_k u_{ik} v_{jk}$, vemos que cada desvio é reconstruído somando, ao longo das K dimensões, o produto das entradas correspondentes dos vetores singulares. O valor singular λ_k pesa o quanto cada dimensão contribui nessa soma.

Uma boa maneira de ler essa soma é tratar as linhas de $\mathbf{U}$ e de $\mathbf{V}$ como representações vetoriais dos níveis. A linha i de $\mathbf{U}$, $(u_{i1}, \dots, u_{iK})$, reúne as coordenadas do i -ésimo nível-linha nas K dimensões, e a linha j de $\mathbf{V}$, $(v_{j1}, \dots, v_{jK})$, faz o mesmo para o j -ésimo nível-coluna. Os sinais dessas coordenadas dizem para que lado cada dimensão empurra a célula (i, j) . Sinais iguais dão contribuição positiva e inflam a frequência em relação à hipótese de independência, enquanto sinais opostos dão contribuição negativa e a diminuem. A seção seguinte mostra como transformar essas coordenadas em um mapa bidimensional.

Proposição 5.3 (Decomposição da inércia total). *Se $\lambda_1, \dots, \lambda_K$ são os valores singulares não nulos de $\mathbf{S}$, então a inércia total decompõe-se como*

$$\phi^2 = \sum_{k=1}^K \lambda_k^2$$

Cada termo λ_k^2 é chamado de **inércia principal** associada à k -ésima dimensão.

Prova. Usando a ortonormalidade dos vetores singulares e a propriedade cíclica do traço,

$$\begin{aligned} \phi^2 &= \text{tr}(\mathbf{S}^T \mathbf{S}) \\ &= \text{tr} \left(\mathbf{V} \underbrace{\mathbf{\Lambda} \mathbf{U}^T \mathbf{U} \mathbf{\Lambda}}_{\mathbf{I}} \mathbf{V}^T \right) \\ &= \text{tr} \left(\mathbf{\Lambda}^2 \underbrace{\mathbf{V}^T \mathbf{V}}_{\mathbf{I}} \right) \\ &= \sum_{k=1}^K \lambda_k^2 \end{aligned}$$

□

A proporção de inércia explicada pela k -ésima dimensão é λ_k^2/ϕ^2 , e a proporção acumulada pelas q primeiras dimensões é $\sum_{k=1}^q \lambda_k^2/\phi^2$.

Exemplo 5.3. Retomando a matriz de correspondência $\mathbf{P}$ e os vetores marginais $\mathbf{r}$ e $\mathbf{c}$ calculados no Exemplo 5.1, a matriz de desvios padronizados $\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-1/2}$ e sua decomposição em valores singulares são calculadas no código a seguir.

Tabela 5.2: Valores singulares e decomposição da inércia da tabela de streaming.

	Valor singular	Inércia principal	Inércia explicada (%)
1	0.4720	0.2228	69.63
2	0.3117	0.0972	30.37

Como mostra a Tabela 5.2, a matriz tem posto $K = \min(3 - 1, 4 - 1) = 2$, com $\lambda_1 \approx 0,4720$ e $\lambda_2 \approx 0,3117$. A inércia total é

$$\phi^2 = \lambda_1^2 + \lambda_2^2 \approx 0,2228 + 0,0972 = 0,3200$$

A primeira dimensão explica 69,63% da inércia total e a segunda dimensão explica os 30,37% restantes.

5.4 Coordenadas principais

Para visualizar as relações descobertas pela SVD, usamos as coordenadas principais. Essas coordenadas têm a propriedade de preservar as distâncias qui-quadrado entre os perfis, como mostramos a seguir.

Definição 5.8 (Coordenadas principais). As **coordenadas principais das linhas** e das **colunas**, nas K dimensões da análise, são dadas por

$$\mathbf{F} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Lambda} \quad \text{e} \quad \mathbf{G} = \mathbf{D}_c^{-1/2} \mathbf{V} \mathbf{\Lambda}$$

A linha i de $\mathbf{F}$, denotada por $\mathbf{f}_i^T = (f_{i1}, \dots, f_{iK})$, reúne as coordenadas do i -ésimo nível-linha. A linha j de $\mathbf{G}$, denotada por $\mathbf{g}_j^T = (g_{j1}, \dots, g_{jK})$, reúne as coordenadas do j -ésimo nível-coluna.

Proposição 5.4 (Preservação das distâncias qui-quadrado). *A distância euclidiana ao quadrado entre dois níveis no espaço das coordenadas principais reproduz exatamente a distância qui-quadrado entre seus perfis*

$$\|\mathbf{f}_i - \mathbf{f}_{i'}\|^2 = \sum_{k=1}^K (f_{ik} - f_{i'k})^2 = d_{\chi^2}^2(i, i') \quad \text{e} \quad \|\mathbf{g}_j - \mathbf{g}_{j'}\|^2 = \sum_{k=1}^K (g_{jk} - g_{j'k})^2 = d_{\chi^2}^2(j, j')$$

Em particular, a distância euclidiana de um nível até a origem mede o desvio do seu perfil em relação ao perfil médio: $\|\mathbf{f}_i\|^2 = d_{\chi^2}^2(i, \mathbf{c})$ e $\|\mathbf{g}_j\|^2 = d_{\chi^2}^2(j, \mathbf{r})$.

Prova. Pela SVD temos $\mathbf{S}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}$, e portanto $\mathbf{F} = \mathbf{D}_r^{-1/2} \mathbf{S}\mathbf{V}$. Chame de $\tilde{\mathbf{x}}_i^T$ a linha i de $\mathbf{D}_r^{-1/2} \mathbf{S}$. Usando $s_{ij} = \sqrt{r_i/c_j} (p_{ij}/r_i - c_j)$, obtido na prova da Proposição 5.2, suas entradas são

$$\tilde{x}_{ij} = \frac{s_{ij}}{\sqrt{r_i}} = \frac{1}{\sqrt{c_j}} \left(\frac{p_{ij}}{r_i} - c_j \right)$$

Ou seja, $\tilde{\mathbf{x}}_i$ é o perfil da linha i centrado no perfil médio, com cada coordenada dividida por $\sqrt{c_j}$. Essa divisão embute os pesos $1/c_j$ da distância qui-quadrado na distância euclidiana comum

$$\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{i'}\|^2 = \sum_{j=1}^J \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2 = d_{\chi^2}^2(i, i')$$

Resta ver que a multiplicação por $\mathbf{V}$ não altera essas distâncias. Como $\mathbf{V}^T \mathbf{V} = \mathbf{I}$, vale $\mathbf{S}\mathbf{V}\mathbf{V}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \mathbf{V}\mathbf{V}^T = \mathbf{S}$, e portanto $\mathbf{V}\mathbf{V}^T \tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}_i$ para toda linha i . Com $\mathbf{f}_i = \mathbf{V}^T \tilde{\mathbf{x}}_i$, obtemos

$$\|\mathbf{f}_i - \mathbf{f}_{i'}\|^2 = (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{i'})^T \mathbf{V} \mathbf{V}^T (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{i'}) = \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_{i'}\|^2 = d_{\chi^2}^2(i, i')$$

Trocando $\tilde{\mathbf{x}}_{i'}$ pelo vetor nulo, a mesma conta dá $\|\mathbf{f}_i\|^2 = \|\tilde{\mathbf{x}}_i\|^2 = d_{\chi^2}^2(i, \mathbf{c})$. Para os níveis-coluna, basta repetir o raciocínio com $\mathbf{G} = \mathbf{D}_c^{-1/2} \mathbf{S}^T \mathbf{U}$. $\square$

As coordenadas principais são, então, os perfis centrados e reescalados escritos na base dos vetores singulares, de modo parecido com os escores da ACP. Essa construção converte a métrica qui-quadrado dos perfis na distância euclidiana comum. É isso que autoriza a leitura visual do mapa, já que a régua que usamos no papel passa a ser a régua correta do problema.

Exemplo 5.4 (Coordenadas principais de streaming). Aplicando a Definição 5.8 aos vetores e valores singulares obtidos no Exemplo 5.3, calculamos as matrizes de coordenadas principais $\mathbf{F} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Lambda}$ e $\mathbf{G} = \mathbf{D}_c^{-1/2} \mathbf{V} \mathbf{\Lambda}$, apresentadas na Tabela 5.3.

Tabela 5.3: Coordenadas principais das faixas etárias e dos serviços de streaming.

	Variável	Dimensão 1	Dimensão 2
Jovem	Faixa etária	0.503	-0.290
Adulto	Faixa etária	0.129	0.433
Idoso	Faixa etária	-0.632	-0.143
Netflix	Serviço	0.401	-0.079
Prime Video	Serviço	0.019	0.492
Max	Serviço	-0.736	-0.170
Disney+	Serviço	0.426	-0.372

5.5 Mapa de correspondência

Para construir uma representação gráfica no plano, retemos as duas primeiras dimensões principais, tomando os pontos (f_{i1}, f_{i2}) para as linhas e (g_{j1}, g_{j2}) para as colunas. O gráfico conjunto resultante é o **mapa de correspondência** (ou mapa simétrico). Como os valores singulares estão em ordem decrescente, o plano formado pelas duas primeiras direções captura a maior fração possível da inércia total, $(\lambda_1^2 + \lambda_2^2)/\phi^2$.

Embora as linhas e colunas pertençam a conjuntos distintos de níveis, suas coordenadas no mapa estão conectadas pelas **fórmulas de transição**

$$f_{ik} = \frac{1}{\lambda_k} \sum_{j=1}^J \frac{p_{ij}}{r_i} g_{jk} \quad \text{e} \quad g_{jk} = \frac{1}{\lambda_k} \sum_{i=1}^I \frac{p_{ij}}{c_j} f_{ik}$$

Ou seja, a posição de cada nível-linha é, a menos da constante $1/\lambda_k$, uma média ponderada das posições das colunas, com pesos dados pelas frequências do seu perfil (p_{ij}/r_i). De forma recíproca, cada coluna é a média ponderada das posições das linhas.

🔥 Distâncias entre linhas e colunas

No mapa simétrico, as distâncias euclidianas entre pontos da mesma nuvem (linha com linha, ou coluna com coluna) reproduzem distâncias qui-quadrado. Já a distância entre um ponto-linha e um ponto-coluna não tem interpretação clara, pois as duas nuvens vivem em espaços com métricas diferentes. O que avaliamos nesse caso é a direção em relação à origem. Pontos no mesmo quadrante e ao longo de direções próximas indicam associação positiva, com a combinação aparecendo mais do que o esperado, enquanto pontos em direções opostas indicam que ela aparece menos. Pontos próximos à origem têm perfil semelhante à média geral.

Exemplo 5.5 (Mapa de streaming). A Figura 5.3 exibe o mapa simétrico bidimensional gerado pelas coordenadas calculadas no Exemplo 5.4. Como há apenas duas dimensões não triviais ($K = \min(3 - 1, 4 - 1) = 2$), o mapa representa 100% da inércia total da tabela.

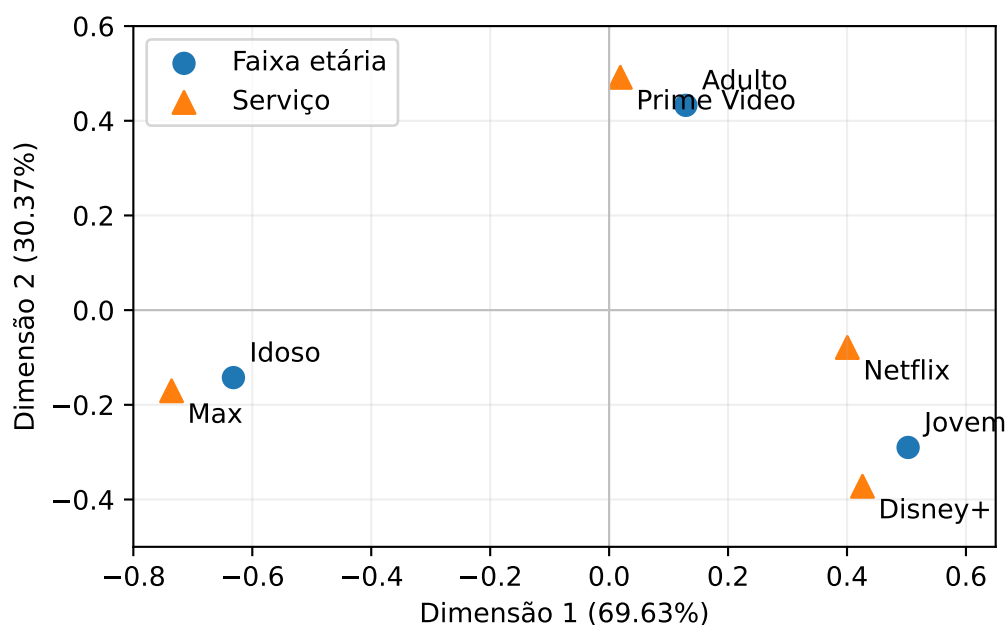


Figura 5.3: Mapa simétrico das preferências por serviço de streaming segundo a faixa etária.

Na Figura 5.3, Netflix e Disney+ aparecem próximos, sinal de que têm perfis semelhantes, enquanto as três faixas etárias ficam bem afastadas umas das outras. O ponto Jovem cai no mesmo quadrante de Netflix (45%) e Disney+ (30%), que são justamente suas escolhas predominantes, enquanto os idosos

concentram 55% das escolhas na Max e aparecem no polo oposto da primeira dimensão. O Prime Video fica praticamente sobre o eixo vertical, ou seja, a primeira dimensão quase não o distingue da média. É a segunda dimensão que o separa, levando-o para o alto do mapa ao lado dos adultos, 45% dos quais escolhem esse serviço.

Quando o espaço completo tem dimensão $K > 2$, projetar os pontos no plano gera perdas. O cosseno ao quadrado mede a fração da distância ao quadrado do perfil até a origem que o plano das duas primeiras dimensões consegue capturar,

$$\cos_{i1}^2 + \cos_{i2}^2 = \frac{f_{i1}^2 + f_{i2}^2}{\sum_{k=1}^K f_{ik}^2} \quad \text{e} \quad \cos_{j1}^2 + \cos_{j2}^2 = \frac{g_{j1}^2 + g_{j2}^2}{\sum_{k=1}^K g_{jk}^2}$$

Valores próximos de 1 indicam que a projeção do ponto é fiel. A influência de cada nível na orientação dos eixos é medida pelas **contribuições** (Ctr) para a inércia da k -ésima dimensão

$$\text{Ctr}_{ik} = \frac{r_i f_{ik}^2}{\lambda_k^2} \quad \text{e} \quad \text{Ctr}_{jk} = \frac{c_j g_{jk}^2}{\lambda_k^2}$$

A contribuição pondera a posição do nível (f_{ik}^2) pela sua marginal (r_i). Por isso, um nível raro pode ficar longe da origem e ainda assim contribuir pouco para o eixo.

Exemplo 5.6 (Cor dos cabelos e dos olhos). O conjunto de dados clássico HairEyeColor relaciona a cor do cabelo e a cor dos olhos de 592 indivíduos. A Tabela 5.4 apresenta as contagens observadas entre quatro cores de cabelo (Preto, Castanho, Ruivo, Loiro) e quatro cores de olhos (Castanho, Azul, Avelã, Verde).

Tabela 5.4: Tabela de contingência observada entre cor do cabelo e cor dos olhos.

	Castanho	Azul	Avelã	Verde
Preto	68	20	15	5
Castanho	119	84	54	29
Ruivo	26	17	14	14
Loiro	7	94	10	16

A Figura 5.4 exibe o mapa simétrico ajustado pelo prince e a Tabela 5.5 reúne a qualidade de representação ($\cos^2$) e as contribuições (Ctr) de cada nível.

As duas primeiras dimensões retêm 98.9% da inércia total da tabela, e a Tabela 5.5 confirma que praticamente todos os níveis exibem $\cos^2$ próximo de 1. A primeira dimensão ordena os níveis pela pigmentação, com os traços claros (cabelo loiro e olhos azuis) em um polo e os escuros (cabelo preto e olhos castanhos) no outro. A segunda dimensão isola o fenótipo ruivo, puxada principalmente pelo cabelo ruivo e pelos olhos verdes, com contribuições de 55,1% e 55,9%.

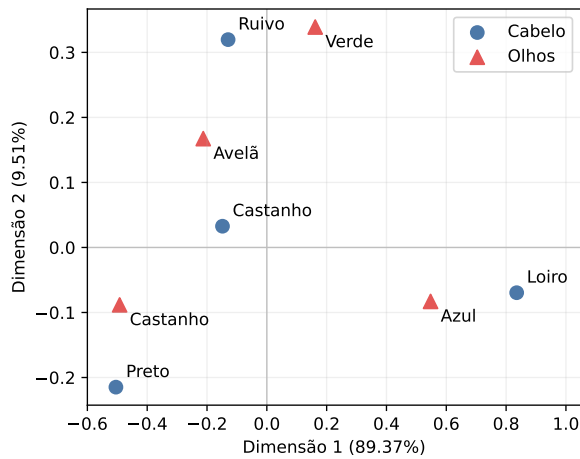


Tabela 5.5: Qualidade de representação e contribuições dos níveis.

Variável	Nível	$\cos^2$	Ctr 1 (%)	Ctr 2 (%)
Cabelo	Preto	0.990	22.2	37.9
	Castanho	0.906	5.1	2.3
	Ruivo	0.945	1.0	55.1
	Loiro	1.000	71.7	4.7
	Castanho	0.998	43.1	13.0
Olhos	Azul	1.000	52.1	11.2
	Avelã	0.879	3.4	19.8
	Verde	0.948	1.4	55.9

Figura 5.4: Mapa de correspondência simétrico entre cor do cabelo e cor dos olhos.

5.5.1 Elementos suplementares

Assim como as variáveis exógenas da ACP, podemos levar ao mapa informações que não participaram do cálculo das coordenadas. Na AC elas recebem o nome de **elementos suplementares**. Um ponto suplementar é um perfil novo passado pela fórmula de transição, sem entrar na SVD. Um perfil $(p_1^*, \dots, p_J^*)$ sobre os J níveis-coluna coloca o ponto na nuvem das linhas, e um perfil $(p_1^*, \dots, p_I^*)$ sobre os I níveis-linha coloca o ponto na nuvem das colunas

$$f_k^* = \frac{1}{\lambda_k} \sum_{j=1}^J p_j^* g_{jk} \quad \text{e} \quad g_k^* = \frac{1}{\lambda_k} \sum_{i=1}^I p_i^* f_{ik}$$

Isso serve tanto para uma variável nova, cruzada com uma das variáveis ativas, quanto para um nível que preferimos deixar fora da tabela ativa, como respostas “Não sabe” ou “Prefere não responder”. Nos dois casos, o ponto aparece no mapa sem mexer na posição dos demais.

Exemplo 5.7. Suponha que a pesquisa do Exemplo 5.1 tivesse registrado também o sexo de cada entrevistado. Como o sexo não faz parte da tabela, ele entra como variável suplementar. Para projetá-lo no mapa, cruzamos o sexo com os serviços.

Sexo	Netflix	Prime Video	Max	Disney+	Total
Feminino	72	36	36	36	180
Masculino	18	44	44	14	120

Cada sexo vira, assim, um perfil sobre os quatro serviços. O perfil feminino é (0,4; 0,2; 0,2; 0,2) e, com as coordenadas dos serviços da Tabela 5.3 e $\lambda_1 \approx 0,4720$,

$$f_1^* = \frac{0,4 \times 0,401 + 0,2 \times 0,019 + 0,2 \times (-0,736) + 0,2 \times 0,426}{0,4720} \approx 0,216$$

As demais coordenadas saem da mesma conta e estão no código a seguir.

Agora imagine que outras 30 pessoas tivessem respondido “Não sei” quando perguntadas sobre o serviço favorito, sendo 6 jovens, 6 adultos e 18 idosos. Essa resposta não expressa preferência por serviço nenhum, então faz sentido deixá-la de fora da tabela ativa. Ainda assim, “Não sei” é um nível da variável serviço, com perfil (0,2; 0,2; 0,6) sobre as faixas etárias, e a segunda fórmula o coloca na nuvem das colunas, ao lado dos serviços.

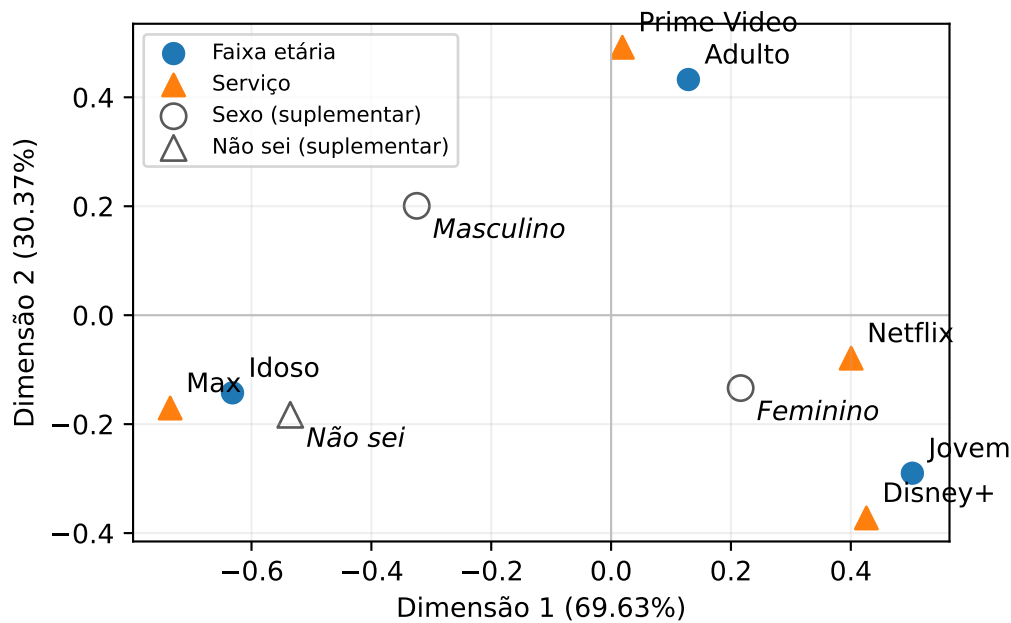


Figura 5.5: Mapa de streaming com o sexo projetado como variável suplementar e o nível Não sei como coluna suplementar.

Na Figura 5.5 os dois sexos caem bem mais perto da origem do que qualquer faixa etária, sinal de que o sexo diferencia as preferências bem menos do que a idade. O feminino puxa, de leve, para o lado de Netflix e Disney+, e o masculino para o lado de Max e Prime Video. Já o “Não sei” aparece junto dos idosos, o que é coerente com 60% de quem não soube responder estar nessa faixa.

5.6 Exercícios

Exercício 5.1. O proprietário de uma joalheria classificou 400 pedras preciosas de acordo com a qualidade do corte (Ideal, Premium, Good) e a tonalidade de cor (D, incolor; E, quase incolor; F, ligeiramente amarelada). A tabela de contingência observada é

Corte / Cor	D	E	F	Total
Ideal	90	60	30	180
Premium	40	70	40	150
Good	20	30	20	70
Total	150	160	90	400

- a) Calcule a inércia total ϕ^2 . Quantas dimensões são necessárias para representar toda essa inércia?
- b) Se a joalheria tivesse classificado 4000 pedras com exatamente as mesmas proporções, o que mudaria na inércia total, na estatística χ^2 e no p-valor do teste de independência?
- c) Usando Python, faça a análise de correspondência e interprete o mapa.

Exercício 5.2. Dois analistas fizeram a análise de correspondência da mesma tabela. Na primeira dimensão, o primeiro obteve as coordenadas principais f_{i1} e g_{j1} , e o segundo obteve $-f_{i1}$ e $-g_{j1}$. Algum deles errou? E se um terceiro analista tivesse invertido o sinal apenas das coordenadas das linhas, mantendo as das colunas? Justifique as duas respostas.

Exercício 5.3. Na análise de correspondência de uma tabela 5×5 , as duas primeiras inércias principais são $\lambda_1^2 = 0,15$ e $\lambda_2^2 = 0,03$. Um nível-linha com marginal $r_1 = 0,10$ tem, nas quatro dimensões da análise, as coordenadas principais

$$(f_{11}, f_{12}, f_{13}, f_{14}) = (0,80; 0,10; 0,30; 0,10)$$

- a) Calcule a qualidade de representação desse nível no plano das duas primeiras dimensões. A posição dele no mapa é confiável?
- b) Calcule as contribuições desse nível para as duas primeiras dimensões. Para qual dos eixos ele é mais relevante?

Exercício 5.4. Em uma pesquisa de preferência por marcas em J regiões, as marcas A e B, do mesmo fabricante, têm perfis idênticos, isto é, $p_{Aj}/r_A = p_{Bj}/r_B$ para todo j . Considere uma análise de correspondência com $K = 2$.

- a) Onde A e B aparecem no mapa de correspondência? Justifique.

b) Se A e B forem reunidas em uma única marca AB, somando suas contagens, mostre que, para qualquer outra marca C,

$$d_{\chi^2}^2(AB, C) = d_{\chi^2}^2(A, C) = d_{\chi^2}^2(B, C)$$

5.7 Análise de correspondência múltipla

Em pesquisas de opinião, formulários socioeconômicos ou fichas de cadastro raramente temos apenas duas variáveis. Em geral, temos Q perguntas categóricas respondidas por n pessoas, e queremos um único mapa que mostre ao mesmo tempo como as respostas se associam e onde cada respondente se encontra.

O biplot da ACP também representava variáveis e indivíduos no mesmo plano, e nesse sentido a ACM funciona como uma ACP para dados categóricos.

Definição 5.9 (Matriz indicadora). Considere uma amostra de n observações sobre Q variáveis categóricas, em que a q -ésima variável tem J_q níveis, com $J = \sum_{q=1}^Q J_q$ níveis no total. A **matriz indicadora** $\mathbf{Z}$ é uma matriz binária de dimensão $n \times J$ dividida em Q blocos

$$\mathbf{Z} = (\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \cdots \quad \mathbf{Z}_Q)$$

em que cada bloco $\mathbf{Z}_q$ tem dimensão $n \times J_q$. O elemento (i, j) de $\mathbf{Z}_q$ é igual a 1 se a i -ésima observação assinalou o j -ésimo nível da variável q , e 0 caso contrário.

Como cada respondente escolhe exatamente um nível por variável, cada linha de cada bloco $\mathbf{Z}_q$ contém uma única entrada 1. Por isso, cada linha da matriz completa $\mathbf{Z}$ soma Q , e a soma de todas as entradas de $\mathbf{Z}$ é nQ . As somas das colunas correspondem às frequências marginais de cada nível, $n_j = \sum_{i=1}^n z_{ij}$.

Exemplo 5.8. Uma pesquisa sobre deslocamento urbano perguntou a doze pessoas onde moram (Centro, Bairro ou Rural), como vão ao trabalho (Carro, Ônibus ou Bicicleta) e quanto tempo levam no trajeto (Curto ou Longo). São $Q = 3$ perguntas com $J_1 = 3$, $J_2 = 3$ e $J_3 = 2$ níveis, totalizando $J = 8$ níveis. Cada pergunta vira um bloco de colunas, e a resposta de cada pessoa vira um 1 na coluna do nível assinalado. A primeira pessoa, por exemplo, mora no centro, vai de bicicleta e tem trajeto curto.

$$\mathbf{Z} = \begin{pmatrix} \begin{array}{ccc|ccc|cc} \text{Centro} & \text{Bairro} & \text{Rural} & \text{Carro} & \text{Ônibus} & \text{Bicicleta} & \text{Curto} & \text{Longo} \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{array} \end{pmatrix}$$

Toda linha tem exatamente três entradas iguais a 1, uma por pergunta, e portanto soma $Q = 3$. O total geral é $nQ = 36$. Já as somas das colunas contam quantas vezes cada nível foi assinalado, (4, 5, 3 | 4, 5, 3 | 6, 6), e dentro de cada bloco elas somam $n = 12$, porque toda pessoa responde a todas as perguntas.

As definições da ACS seguem valendo, com a matriz indicadora no papel da tabela de contingência. A matriz de correspondência é $\mathbf{P} = \mathbf{Z}/(nQ)$ e, como cada linha de $\mathbf{Z}$ soma Q e a coluna j soma n_j , as marginais são

$$r_i = \frac{1}{n} \quad \text{e} \quad c_j = \frac{n_j}{nQ} = \frac{p_j}{Q}$$

em que $p_j = n_j/n$ é a proporção de respondentes que assinalaram o nível j . Todos os indivíduos pesam o mesmo, e cada nível pesa conforme sua frequência. Dentro de um bloco as marginais somam $1/Q$, então cada pergunta entra com o mesmo peso, não importa quantos níveis tenha.

5.7.1 Inércia na ACM

A matriz $\mathbf{S}$ e a inércia total são calculadas como antes, mas a leitura muda. A hipótese de independência diz que cada pessoa responde conforme as marginais, espalhando massa por todos os níveis, e nenhum respondente faz isso, já que tem zeros em todos os níveis que não assinalou. Por isso a inércia da matriz indicadora não mede associação entre as perguntas.

Proposição 5.5 (Inércia total da matriz indicadora). *Na análise de correspondência da matriz indicadora, a inércia do nível j é $(1 - p_j)/Q$ e a inércia total vale*

$$\phi^2 = \frac{J - Q}{Q}$$

qualquer que seja a associação entre as variáveis.

Prova. Substituindo $p_{ij} = z_{ij}/(nQ)$, $r_i = 1/n$ e $c_j = p_j/Q$ na Definição 5.4, obtemos

$$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} = \frac{(z_{ij} - p_j)/(nQ)}{\sqrt{p_j/(nQ)}} = \frac{z_{ij} - p_j}{\sqrt{nQ p_j}}$$

A coluna j de $\mathbf{Z}$ é uma variável binária com média p_j , e sua soma de quadrados em torno da média é $\sum_i (z_{ij} - p_j)^2 = np_j(1 - p_j)$. A inércia do nível j é, então,

$$\sum_{i=1}^n s_{ij}^2 = \frac{np_j(1 - p_j)}{nQ p_j} = \frac{1 - p_j}{Q}$$

Em cada uma das Q perguntas as proporções somam 1, então $\sum_j p_j = Q$ e, somando sobre os J níveis, $\phi^2 = (J - Q)/Q$. $\square$

A inércia depende apenas do número de níveis e da frequência de cada um, e não de como as perguntas se relacionam. A prova mostra ainda que $\mathbf{S}$ é a matriz indicadora centrada, com cada coluna dividida por $\sqrt{nQ p_j}$, o que reforça a leitura da ACM como uma ACP das variáveis indicadoras.

5.7.2 Solução e mapa de correspondência

A partir daqui, tudo segue como na ACS. Decompomos $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ e obtemos as coordenadas principais (Definição 5.8), agora com uma linha de $\mathbf{F}$ por respondente e uma linha de $\mathbf{G}$ por nível. O número de dimensões não triviais é $K = J - Q$.

Como $p_{ij}/r_i = z_{ij}/Q$, o perfil de um respondente é a lista dos Q níveis que ele assinalou, e as fórmulas de transição viram médias simples

$$f_{ik} = \frac{1}{\lambda_k} \frac{1}{Q} \sum_{j \in C(i)} g_{jk} \quad \text{e} \quad g_{jk} = \frac{1}{\lambda_k} \frac{1}{n_j} \sum_{i \in I(j)} f_{ik}$$

em que $C(i)$ reúne os Q níveis assinalados pelo respondente i e $I(j)$ reúne os n_j respondentes que assinalaram o nível j . A menos do fator $1/\lambda_k$, cada pessoa fica no centro dos níveis que marcou e cada nível fica no centro das pessoas que o marcaram. Essa é a **propriedade baricêntrica**, e graças a ela podemos desenhar pessoas e níveis no mesmo mapa.

Exemplo 5.9. O código a seguir aplica a decomposição ao questionário do Exemplo 5.8 e desenha as doze pessoas junto dos oito níveis.

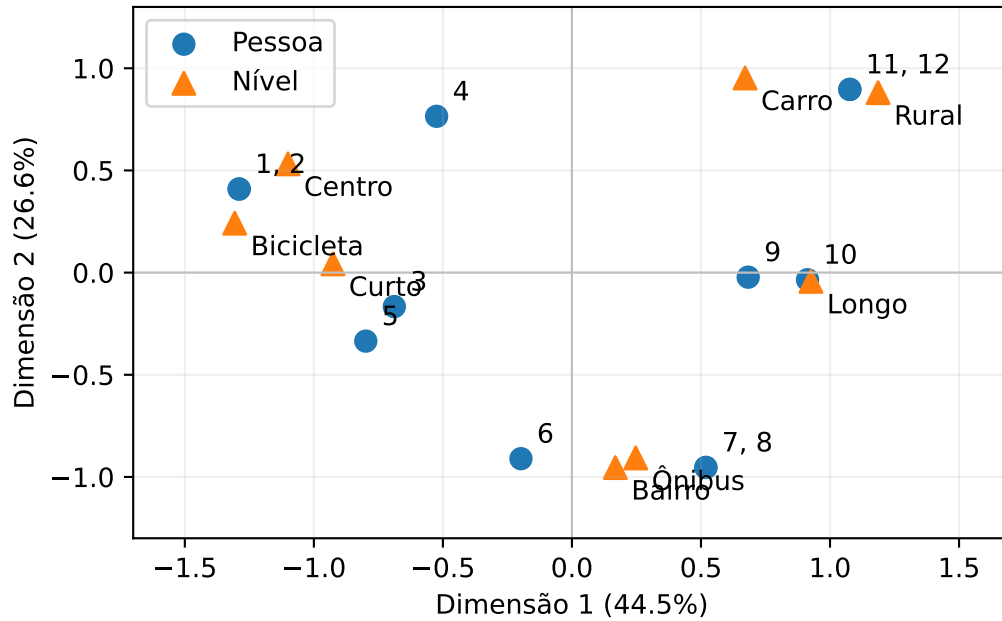


Figura 5.6: Mapa de ACM do questionário de deslocamento, com as doze pessoas e os oito níveis no mesmo plano.

A análise devolve $K = J - Q = 5$ dimensões, com autovalores 0.742, 0.443, 0.229, 0.184 e 0.068, que somam a inércia $5/3$ prevista pela Proposição 5.5.

Na Figura 5.6, a primeira dimensão separa quem mora perto e chega rápido (Centro, Bicicleta e Curto) de quem mora longe e demora (Rural, Carro e Longo). As pessoas se espalham ao longo desse eixo na mesma ordem e, pela propriedade baricêntrica, cada uma fica perto dos níveis que assinalou. Quem deu exatamente as mesmas respostas, como as pessoas 1 e 2, ocupa o mesmo ponto do mapa.

A segunda dimensão não traz informação nova. Os pontos desenharam um arco, com os níveis intermediários, Bairro e Ônibus, de um lado do eixo e os extremos, Centro, Bicicleta, Rural e Carro, do outro. Esse arco, ou ferradura, é conhecido como efeito Guttman e costuma aparecer quando as respostas seguem um único gradiente.

A matriz F também serve fora do mapa. Cada respondente, que chegou como uma lista de respostas categóricas, sai da ACM como um vetor numérico, e F faz na ACM o papel da matriz de escores da ACP (Definição 4.3). A pessoa 1 do questionário de deslocamento, que mora no centro, vai de bicicleta e tem trajeto curto, vira o ponto $(-1,29; 0,41)$ nas duas primeiras dimensões. Pela propriedade baricêntrica,

esse escore é a média dos valores dos níveis que ela marcou, como uma pontuação de questionário com pesos escolhidos pelos próprios dados.

Poderíamos argumentar que $\mathbf{Z}$ já é numérica e bastaria usá-la diretamente. A distância euclidiana entre duas linhas de $\mathbf{Z}$, porém, só conta em quantas perguntas as duas pessoas discordam, e toda discordância pesa igual. A ACM usa a distância qui-quadrado entre os perfis, que na matriz indicadora, com $p_{ij}/r_i = z_{ij}/Q$ e $c_j = n_j/(nQ)$, fica

$$d_{\chi^2}^2(i, i') = \frac{n}{Q} \sum_{j=1}^J \frac{(z_{ij} - z_{i'j})^2}{n_j}$$

Cada discordância pesa pelo inverso da frequência dos níveis envolvidos. As pessoas 6 e 7 discordam apenas no trajeto, entre Curto e Longo, assinalados por seis pessoas cada, e $d_{\chi^2}^2 = 4(1/6 + 1/6) \approx 1,33$. As pessoas 1 e 3 também discordam em uma única pergunta, mas entre Bicicleta e Ônibus, assinalados por três e cinco pessoas, e $d_{\chi^2}^2 = 4(1/3 + 1/5) \approx 2,13$. Em $\mathbf{Z}$ os dois pares estariam à mesma distância, $\sqrt{2}$. Pela Proposição 5.4, a distância euclidiana entre as linhas de $\mathbf{F}$ reproduz essa distância qui-quadrado, o que justifica usar $\mathbf{F}$ como entrada de técnicas que dependem de distâncias ou de variáveis numéricas, como a análise de agrupamentos e a análise discriminante.

5.7.3 Escolha das dimensões

Como a inércia total é fixada pelo número de níveis, os percentuais de inércia explicada saem baixos em ACM, e um primeiro eixo com 15% ou 20% é normal. Por isso, em vez de olhar os percentuais, comparamos cada autovalor com a inércia média por dimensão,

$$\frac{\phi^2}{K} = \frac{(J - Q)/Q}{J - Q} = \frac{1}{Q}$$

O **critério de Benzécri** recomenda interpretar apenas os eixos que superam essa média, $\lambda_k^2 > 1/Q$. No questionário de deslocamento, $1/Q = 0,333$ e apenas os dois primeiros autovalores, 0,742 e 0,443, superam esse limiar, então as três últimas dimensões podem ser descartadas. A segunda passa no critério, mas já sabemos que ela descreve apenas o arco, e a leitura se concentra na primeira.

Pela Proposição 5.5, a inércia de um nível, $(1 - p_j)/Q$, cresce conforme ele fica raro e se aproxima de $1/Q$, a inércia média por dimensão. Como quase ninguém marcou esse nível, sua inércia não se dilui em um eixo compartilhado, e a decomposição acaba reservando um eixo só para separar aquele punhado de pessoas. O remédio é fundir esses níveis com outros afins ou declará-los suplementares.

Exemplo 5.10 (Surto alimentar). O conjunto `poison` registra um episódio de intoxicação alimentar em um acampamento com 55 crianças. Para cada criança foram anotados 5 sintomas clínicos e o consumo de 6 alimentos, todos como variáveis binárias, além do diagnóstico geral.

Tabela 5.6: Frequências das respostas afirmativas no surto alimentar.

	Sim	Não
Variável		
Náusea	12	43
Vômito	22	33
Dor abdominal	37	18
Febre	35	20
Diarreia	35	20
Batata	52	3
Peixe	54	1
Maionese	45	10
Abobrinha	50	5
Queijo	48	7
Sorvete	51	4

As onze variáveis, sintomas e alimentos, entram como ativas, e o diagnóstico geral fica como suplementar, para que não influencie os eixos e sirva de conferência da leitura. Note, na Tabela 5.6, que peixe, batata, abobrinha e sorvete foram consumidos por quase todo mundo. O nível Não dessas variáveis é, portanto, raro, e vamos ver o efeito disso na análise.

Tabela 5.7: Autovalores da ACM do surto alimentar e percentuais de inércia explicada.

	Autovalor	Inércia explicada (%)	Acumulada (%)
1	0.335	33.5	33.5
2	0.129	12.9	46.4
3	0.107	10.7	57.2
4	0.096	9.6	66.8
5	0.079	7.9	74.6
6	0.071	7.1	81.8
7	0.060	6.0	87.8
8	0.056	5.6	93.3
9	0.041	4.1	97.5
10	0.013	1.3	98.8
11	0.012	1.2	100.0

Com $Q = 11$ variáveis ativas e $J = 22$ níveis, a análise tem $K = 11$ dimensões e o limiar de Benzécri é $1/11 \approx 0,091$. A Tabela 5.7 mostra que as quatro primeiras dimensões superam esse limiar e somam 66.8% da inércia. A primeira se destaca, com 33.5% sozinha, enquanto as três seguintes ficam próximas umas das outras e do próprio limiar.

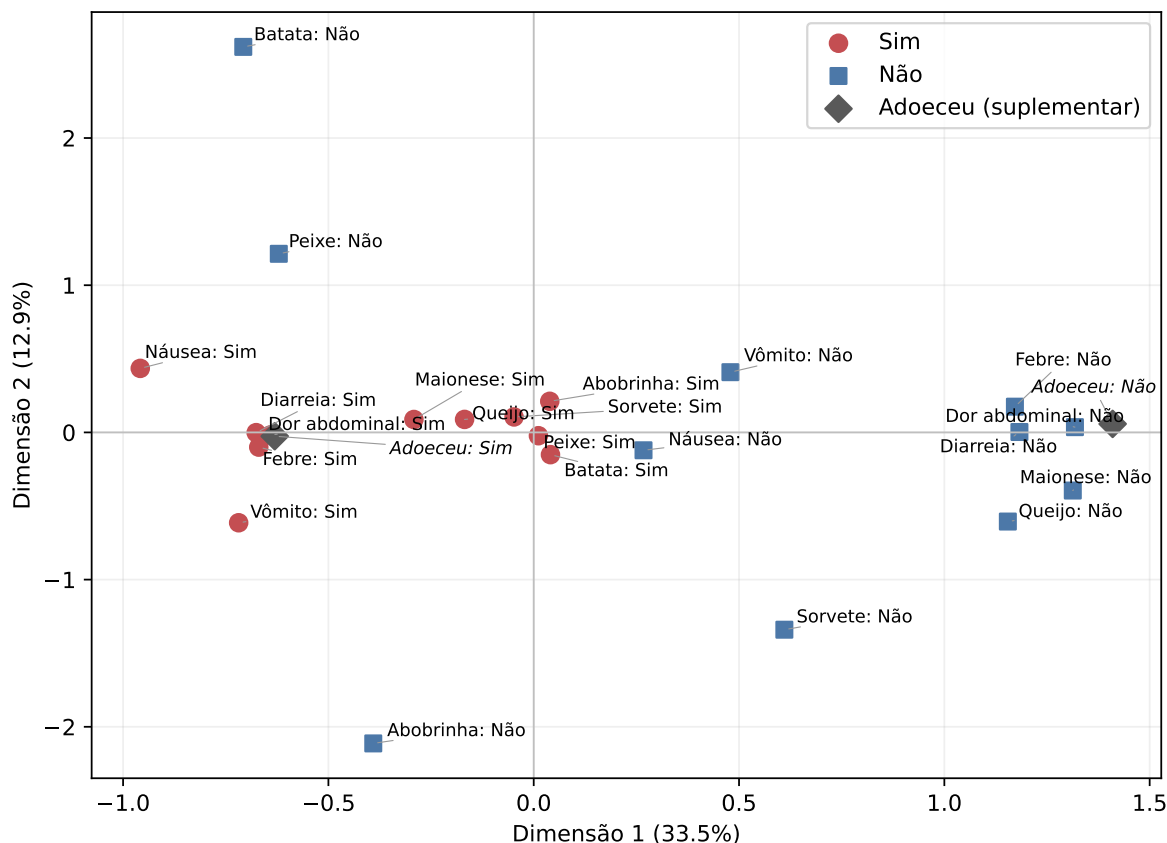


Figura 5.7: Mapa de ACM do surto alimentar, com sintomas e alimentos como variáveis ativas e o diagnóstico geral projetado como suplementar.

Na Figura 5.7, a primeira dimensão funciona como um gradiente de contaminação. Os Sim dos cinco sintomas se agrupam de um lado e os Não do outro, e o diagnóstico geral confirma a leitura, com Adoeceu: Sim e Adoeceu: Não nos extremos do eixo, sem ter participado da construção dele. Entre os alimentos, a maioria dos níveis fica perto da origem nessa dimensão, ou seja, ter comido ou não aquele alimento pouco tem a ver com os sintomas. A exceção é Maionese: Não, que cai do lado das crianças saudáveis, acompanhada de Queijo: Não. De fato, das 10 crianças que não comeram maionese, 8 não adoeceram, e a maionese aparece como o foco provável do surto.

A segunda dimensão mostra o efeito dos níveis raros. Batata: Não e Abobrinha: Não, assinalados por 3 e 5 crianças, ocupam os extremos do eixo e respondem juntos por 55% da contribuição dele. O eixo existe basicamente para separar essas oito crianças das demais e não descreve nenhum padrão do surto.

As dimensões 3 e 4 não aparecem no mapa, mas as contribuições contam a mesma história. Na terceira, Sorvete: Não, Batata: Não e Peixe: Não respondem por 23%, 21% e 21% da contribuição. Na

quarta, Peixe: Não, assinalado por uma única criança, responde sozinho por 53%. Para limpar essas dimensões, bastaria declarar os níveis raros como suplementares.

5.7.4 Tabela de Burt

Muitos programas oferecem uma alternativa à matriz indicadora, que é aplicar a análise de correspondência à tabela de Burt.

Definição 5.10 (Tabela de Burt). A **tabela de Burt** é a matriz simétrica $J \times J$

$$B = Z^T Z = \begin{pmatrix} Z_1^T Z_1 & \cdots & Z_1^T Z_Q \\ \vdots & \ddots & \vdots \\ Z_Q^T Z_1 & \cdots & Z_Q^T Z_Q \end{pmatrix}$$

Fora da diagonal, o bloco $Z_q^T Z_{q'}$ é a tabela de contingência entre as variáveis q e q' . Na diagonal, $Z_q^T Z_q$ é uma matriz diagonal com as frequências n_j dos níveis da variável q .

Exemplo 5.11. No questionário do Exemplo 5.8, a tabela de Burt reúne as três tabelas de contingência entre moradia, transporte e trajeto, cada uma aparecendo duas vezes por simetria. As linhas seguem a mesma ordem das colunas.

$$B = \begin{pmatrix} \begin{array}{ccc|ccc|cc} \text{Centro} & \text{Bairro} & \text{Rural} & \text{Carro} & \text{Ônibus} & \text{Bicicleta} & \text{Curto} & \text{Longo} \\ \hline 4 & 0 & 0 & 1 & 1 & 2 & 4 & 0 \\ 0 & 5 & 0 & 1 & 3 & 1 & 2 & 3 \\ 0 & 0 & 3 & 2 & 1 & 0 & 0 & 3 \\ \hline 1 & 1 & 2 & 4 & 0 & 0 & 1 & 3 \\ 1 & 3 & 1 & 0 & 5 & 0 & 2 & 3 \\ 2 & 1 & 0 & 0 & 0 & 3 & 3 & 0 \\ \hline 4 & 2 & 0 & 1 & 2 & 3 & 6 & 0 \\ 0 & 3 & 3 & 3 & 3 & 0 & 0 & 6 \end{array} \end{pmatrix}$$

O bloco entre moradia e transporte, por exemplo, mostra que, das quatro pessoas do centro, duas vão de bicicleta, uma de ônibus e uma de carro. Na diagonal reaparecem as frequências (4, 5, 3), (4, 5, 3) e (6, 6), as mesmas somas das colunas de Z .

Aplicar a AC à tabela de Burt não produz uma análise nova. Os eixos são os mesmos da matriz indicadora, e as inércias principais da Burt são λ_k^4 , os quadrados das inércias da indicadora. As coordenadas dos níveis só mudam de escala, ficando multiplicadas por λ_k em cada eixo, então o mapa dos níveis é o mesmo desenho, com as dimensões fracas encolhidas mais do que as fortes.

Os indivíduos, porém, desaparecem, porque B só tem níveis, e se quisermos as pessoas no mapa elas precisam entrar como pontos suplementares. Além disso, elevar as inércias ao quadrado concentra os percentuais nos primeiros eixos. No questionário de deslocamento, as duas primeiras dimensões passam de 44,5% e 26,6% para 65,7% e 23,4%. A Burt também não resolve o problema da inércia fixada pelos níveis, já que os blocos da diagonal, que cruzam cada variável com ela mesma, continuam inflando a inércia total sem medir associação alguma.

5.8 Exercícios

Exercício 5.5. Um questionário socioeconômico reúne $Q = 8$ variáveis categóricas com 3 níveis cada.

a) Calcule a inércia total da matriz indicadora, o número máximo de dimensões não triviais e o limiar de Benzécri.

b) As duas primeiras dimensões têm autovalores $\lambda_1^2 = 0,35$ e $\lambda_2^2 = 0,20$, e as demais ficam abaixo do limiar. Quanto da inércia total essas duas dimensões explicam? Esse percentual indica que a análise é ruim?

Exercício 5.6. Retome o questionário de deslocamento do Exemplo 5.8.

a) Calcule a inércia e a distância qui-quadrado ao perfil médio dos níveis Rural e Bairro. Qual dos dois fica mais longe da origem e qual carrega mais inércia?

b) Mostre que cada pergunta q contribui com $(J_q - 1)/Q$ para a inércia total. Por que perguntas com mais níveis pesam mais na análise?

Exercício 5.7 (Cogumelos). O conjunto [Mushroom, da UCI Machine Learning Repository](#), descreve 8.124 cogumelos de 23 espécies das famílias *Agaricus* e *Lepiota*. Cada cogumelo é caracterizado por 22 atributos morfológicos categóricos, como forma e cor do chapéu, odor, cor das lamelas e hábitat, e classificado como comestível ou venenoso. Use a ACM para descrever como esses atributos se associam, deixando a classificação comestível/venenoso como variável suplementar. Existe algum atributo que, sozinho, praticamente separa os dois grupos? Justifique todas as decisões tomadas, em especial o tratamento dos níveis raros e dos valores ausentes.

6 Análise fatorial

Suponha que registramos, para várias famílias, o gasto mensal em educação, em cultura e em alimentação. As três variáveis caminham juntas, pois famílias com mais renda gastam mais em tudo. Mesmo sem que a renda apareça na nossa tabela, sabemos que ela explica boa parte dos três gastos.

A análise fatorial (AF) parte dessa ideia e supõe que por trás das variáveis medidas existem algumas poucas variáveis não observáveis, os **fatores comuns**, e que é a ação delas que produz as correlações que vemos. No exemplo dos gastos, um único fator, a renda familiar, influencia os três ao mesmo tempo. Cada gasto guarda ainda uma parcela própria, que não é compartilhada com os demais e junta tanto o que é particular daquela variável quanto o erro de medida.

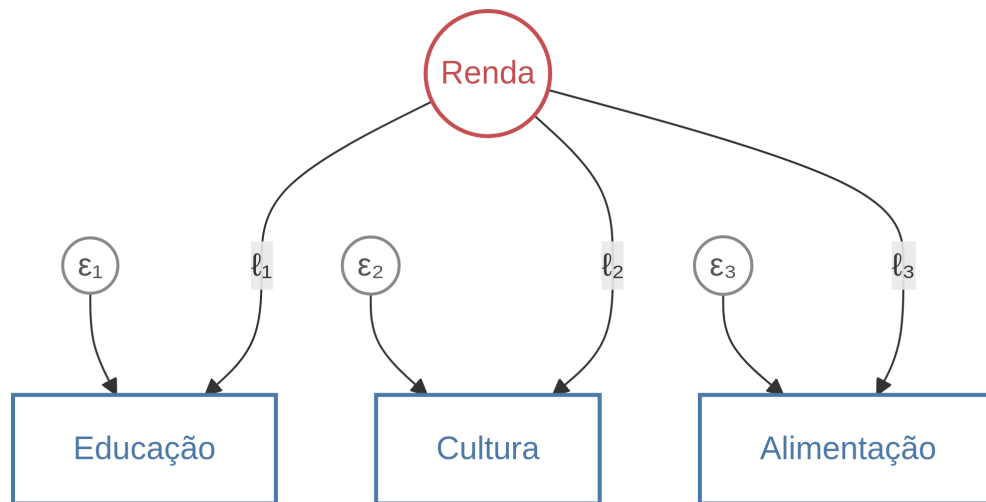


Figura 6.1: O modelo fatorial para os gastos familiares. O fator comum, não observado, responde pela parte compartilhada dos três gastos, enquanto cada termo específico fica com o que sobra.

Repare para onde apontam as setas da Figura 6.1. Todas chegam às variáveis observadas, porque são o fator e os termos específicos que as geram. As cargas ℓ_j medem a força de cada influência, e estimá-las sem nunca observar o que está no alto do diagrama é o problema deste capítulo.

6.1 O modelo fatorial ortogonal

Definição 6.1 (Modelo fatorial ortogonal). Seja $\mathbf{x}$ um vetor aleatório de p variáveis observadas, com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. O modelo fatorial ortogonal com m fatores comuns, $m < p$, escreve

$$\mathbf{x} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon} \quad (6.1)$$

em que $\mathbf{L}$ é a matriz $p \times m$ de **cargas fatoriais**, $\mathbf{F} = (F_1, \dots, F_m)^T$ reúne os fatores comuns e $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_p)^T$ reúne os **fatores específicos**. O elemento ℓ_{jk} de $\mathbf{L}$ mede quanto o fator F_k pesa sobre a variável X_j .

O modelo supõe ainda que

$$\mathbb{E}[\mathbf{F}] = \mathbf{0}, \quad \text{Cov}(\mathbf{F}) = \mathbf{I}_m$$

$$\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \text{diag}(\psi_1, \dots, \psi_p)$$

$$\text{Cov}(\mathbf{F}, \boldsymbol{\epsilon}) = \mathbf{0}$$

A exigência $\text{Cov}(\mathbf{F}) = \mathbf{I}_m$ é uma convenção. Como ninguém observa os fatores, eles não têm unidade própria, e fixamos a variância de cada um em 1. Se dobrássemos um fator e dividíssemos suas cargas por 2, nada mudaria em $\mathbf{x}$. As covariâncias nulas entre os fatores dão ao modelo o nome de ortogonal e também são uma escolha, da qual veremos como abrir mão na rotação.

Já a suposição de que $\boldsymbol{\Psi}$ é diagonal é onde está o conteúdo do modelo. Ela diz que a parte específica de uma variável não tem relação com a parte específica de outra, e então duas variáveis só podem andar juntas por meio dos fatores comuns. No exemplo dos gastos, isso significa que educação e cultura se correlacionam apenas porque ambas dependem da renda. Se as famílias que valorizam cultura também investissem mais em educação, independentemente do quanto ganham, essa associação extra não caberia num modelo que tem só a renda como fator. É uma suposição que os dados podem desmentir, e mais adiante veremos como testá-la.

Teorema 6.1 (Estrutura da matriz de covariâncias). *Sob as suposições da Definição 6.1, a matriz de covariâncias de $\mathbf{x}$ é*

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi} \quad (6.2)$$

Prova. Como $\mathbf{x} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\epsilon}$ e ambos os termos têm média zero, cada esperança abaixo é uma matriz de covariâncias, dada pelas suposições do modelo. Os dois termos cruzados desaparecem e sobra

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbb{E} [(\mathbf{LF} + \boldsymbol{\epsilon})(\mathbf{LF} + \boldsymbol{\epsilon})^T] \\ &= \mathbf{L} \underbrace{\mathbb{E} [\mathbf{FF}^T]}_{\text{Cov}(\mathbf{F})=\mathbf{I}_m} \mathbf{L}^T + \mathbf{L} \underbrace{\mathbb{E} [\mathbf{F}\boldsymbol{\epsilon}^T]}_{\text{Cov}(\mathbf{F},\boldsymbol{\epsilon})=\mathbf{0}} + \underbrace{\mathbb{E} [\boldsymbol{\epsilon}\mathbf{F}^T]}_{\text{Cov}(\boldsymbol{\epsilon},\mathbf{F})=\mathbf{0}} \mathbf{L}^T + \underbrace{\mathbb{E} [\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]}_{\text{Cov}(\boldsymbol{\epsilon})=\boldsymbol{\Psi}} \\ &= \mathbf{LL}^T + \boldsymbol{\Psi}\end{aligned}$$

□

A Equação 6.2 divide a matriz de covariâncias em duas partes. A parcela $\mathbf{LL}^T$ é compartilhada e tem posto m , bem menor que p ; a parcela $\boldsymbol{\Psi}$ é diagonal e própria de cada variável. Como $\boldsymbol{\Psi}$ não contribui para nada fora da diagonal, toda a covariância entre variáveis diferentes fica por conta de $\mathbf{LL}^T$. Para $j \neq j'$,

$$\text{Cov}(X_j, X_{j'}) = \sum_{k=1}^m \ell_{jk} \ell_{j'k}$$

Na diagonal, a mesma equação reparte a variância de cada variável. Escrevendo $\sigma_{jj} = \text{Var}(X_j)$ e tomando o elemento (j, j) de $\mathbf{LL}^T$,

$$\sigma_{jj} = \underbrace{\sum_{k=1}^m \ell_{jk}^2}_{h_j^2} + \psi_j$$

A quantidade h_j^2 é a **comunalidade** de X_j , a parte da variância que a variável divide com as demais por meio dos fatores comuns. O que resta, ψ_j , é a **variância específica**, e inclui tanto aquilo que é próprio da variável quanto o erro de medida. Quando a AF é feita sobre a matriz de correlações, todas as variâncias valem 1 e a leitura fica ainda mais direta, com $h_j^2 + \psi_j = 1$.

Proposição 6.1 (Cargas e correlações). *Sob as suposições da Definição 6.1, a carga ℓ_{jk} é a covariância entre a variável X_j e o fator F_k , e a correlação entre eles é*

$$\text{Cov}(X_j, F_k) = \ell_{jk}, \quad \text{Corr}(X_j, F_k) = \frac{\ell_{jk}}{\sqrt{\sigma_{jj}}}$$

Em particular, quando X_j está padronizada, sua carga é a própria correlação com o fator,

$$\text{Corr}(X_j, F_k) = \ell_{jk}$$

Nesse caso, as cargas ficam entre -1 e 1 .

Prova. Como $E[\mathbf{F}] = \mathbf{0}$, a definição de covariância e a substituição de $\mathbf{x} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\epsilon}$ dão

$$\begin{aligned} \text{Cov}(\mathbf{x}, \mathbf{F}) &= E[(\mathbf{x} - \boldsymbol{\mu})\mathbf{F}^T] \\ &= E[(\mathbf{L}\mathbf{F} + \boldsymbol{\epsilon})\mathbf{F}^T] \\ &= \mathbf{L} E[\mathbf{F}\mathbf{F}^T] + E[\boldsymbol{\epsilon}\mathbf{F}^T] \\ &= \mathbf{L} \text{Cov}(\mathbf{F}) + \text{Cov}(\boldsymbol{\epsilon}, \mathbf{F}) \\ &= \mathbf{L}\mathbf{I}_m + \mathbf{0} \\ &= \mathbf{L} \end{aligned}$$

Tomando o elemento (j, k) , obtemos $\text{Cov}(X_j, F_k) = \ell_{jk}$. Como $\text{Var}(F_k) = 1$, basta dividir a covariância por $\sqrt{\sigma_{jj}}$ para obter a correlação. Com X_j padronizada, $\sigma_{jj} = 1$. $\square$

A escolha da escala segue a mesma ideia da ACP (Seção 4.2). Costumamos trabalhar com a matriz de correlações quando as unidades ou as escalas das variáveis são diferentes. A matriz de covariâncias faz sentido quando as variáveis têm escalas comparáveis e queremos preservar suas diferenças de variabilidade. Para escrever o modelo na escala padronizada, seja $\mathbf{D}$ a matriz diagonal dos desvios padrão. As variáveis padronizadas $\mathbf{z} = \mathbf{D}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ satisfazem

$$\mathbf{z} = \underbrace{\mathbf{D}^{-1}\mathbf{L}}_{\mathbf{L}_z} \mathbf{F} + \underbrace{\mathbf{D}^{-1}\boldsymbol{\epsilon}}_{\boldsymbol{\epsilon}_z}$$

que é de novo um modelo fatorial, com os mesmos fatores, as cargas de cada variável divididas pelo seu desvio padrão e variâncias específicas $\boldsymbol{\Psi}_z = \mathbf{D}^{-1}\boldsymbol{\Psi}\mathbf{D}^{-1}$, que continuam formando uma matriz diagonal. A estrutura de covariâncias correspondente é

$$\boldsymbol{\rho} = \mathbf{L}_z \mathbf{L}_z^T + \boldsymbol{\Psi}_z$$

Aviso

Nos exemplos a seguir, trabalharemos com a matriz de correlações amostral $\mathbf{R}$. Nesses exemplos, escreveremos $\mathbf{L}$ e $\boldsymbol{\Psi}$ para as cargas e variâncias específicas na escala padronizada, omitindo o índice z .

Exemplo 6.1 (Gastos familiares). Retomemos os três gastos da abertura do capítulo, agora padronizados, com a matriz de correlações

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.56 & 0.48 \\ 0.56 & 1.00 & 0.42 \\ 0.48 & 0.42 & 1.00 \end{pmatrix}$$

Vamos procurar um único fator comum, $m = 1$. Nesse caso $\mathbf{L}$ é um vetor de três cargas e o modelo diz que $\rho_{jj'} = \ell_j \ell_{j'}$ para $j \neq j'$, o que dá três equações para três incógnitas

$$\ell_1 \ell_2 = 0.56, \quad \ell_1 \ell_3 = 0.48, \quad \ell_2 \ell_3 = 0.42$$

Multiplicando as duas primeiras e dividindo pela terceira, os fatores ℓ_2 e ℓ_3 se cancelam e sobra

$$\ell_1^2 = \frac{(\ell_1 \ell_2)(\ell_1 \ell_3)}{\ell_2 \ell_3} = \frac{0.56 \times 0.48}{0.42} = 0.64$$

Logo $\ell_1 = 0.8$, e as outras duas saem por substituição, $\ell_2 = 0.7$ e $\ell_3 = 0.6$. As communalidades são $h_1^2 = 0.64$, $h_2^2 = 0.49$ e $h_3^2 = 0.36$, e as variâncias específicas, $\psi_j = 1 - h_j^2$, valem 0.36, 0.51 e 0.64.

$$\mathbf{L} = \begin{pmatrix} 0.8 \\ 0.7 \\ 0.6 \end{pmatrix}, \quad \mathbf{\Psi} = \begin{pmatrix} 0.36 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.64 \end{pmatrix}$$

Multiplicando de volta, $\mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$ devolve $\mathbf{R}$ exatamente. Assim, a **matriz de resíduos**, dada pela diferença entre a matriz observada e a reproduzida pelo modelo, é nula

$$\mathbf{R} - (\mathbf{L}\mathbf{L}^T + \mathbf{\Psi}) = \mathbf{0}$$

Todos os resíduos são zero, tanto na diagonal quanto fora dela. O gasto em educação é o mais ligado à renda, com 64% da sua variância explicada pelo fator; o gasto em alimentação é o menos ligado, com 36%, o que faz sentido, já que alimentação tem um piso que independe de quanto a família ganha.

6.2 Estimação

Na prática não conhecemos Σ , só a sua estimativa $\mathbf{S}$, e o problema passa a ser encontrar $\hat{\mathbf{L}}$ e $\hat{\mathbf{\Psi}}$ tais que $\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}$ fique o mais perto possível dela. Apresentaremos os métodos em termos da matriz de covariâncias; com variáveis padronizadas, essa matriz é $\mathbf{R}$.

6.2.1 Componentes principais

A decomposição espectral (Teorema 2.1) permite escrever a matriz de covariâncias populacional em forma matricial ou como uma soma das contribuições de cada componente,

$$\Sigma = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T = \sum_{k=1}^p \lambda_k \mathbf{e}_k \mathbf{e}_k^T$$

em que $\mathbf{E}$ reúne os autovetores ortonormais nas colunas e $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ contém os autovalores em ordem decrescente. Como os autovalores de uma matriz de covariâncias são não negativos, podemos tomar $\mathbf{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$ e reescrever a decomposição como

$$\Sigma = \left(\mathbf{E}\mathbf{\Lambda}^{1/2} \right) \left(\mathbf{E}\mathbf{\Lambda}^{1/2} \right)^T$$

Essa expressão já tem a forma $\mathbf{L}\mathbf{L}^T$, e a coluna k de $\mathbf{E}\mathbf{\Lambda}^{1/2}$ é $\sqrt{\lambda_k} \mathbf{e}_k$. Se os primeiros m componentes concentram a maior parte da variabilidade, usamos essas m colunas para construir as cargas e atribuímos o que faltar na diagonal às variâncias específicas. Na amostra, fazemos essa construção a partir dos autovalores e autovetores de $\mathbf{S}$.

Definição 6.2 (Solução por componentes principais). Sejam $(\hat{\lambda}_k, \hat{\mathbf{e}}_k)$, $k = 1, \dots, p$, os pares de autovalor e autovetor de $\mathbf{S}$, com $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$. A solução por componentes principais com m fatores é

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

$$\hat{\Psi} = \text{diag} \left(\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}^T \right), \quad \hat{\psi}_j = s_{jj} - \sum_{k=1}^m \hat{\ell}_{jk}^2$$

Com variáveis padronizadas, substituímos $\mathbf{S}$ por $\mathbf{R}$ e s_{jj} por 1, usando os autovalores e autovetores de $\mathbf{R}$.

Por construção, as variâncias específicas da Definição 6.2 absorvem o que faltar na diagonal, tornando nulos os resíduos nessa posição. Para avaliar o ajuste, olhamos então para os resíduos fora da diagonal, cuja soma de quadrados pode ser limitada pelos autovalores que deixamos de fora.

Proposição 6.2 (Limite para os resíduos). *Na solução por componentes principais com m fatores da Definição 6.2, a soma dos quadrados dos resíduos fora da diagonal satisfaz*

$$\sum_{j \neq j'} \left(s_{jj'} - \sum_{k=1}^m \hat{\ell}_{jk} \hat{\ell}_{j'k} \right)^2 \leq \sum_{k=m+1}^p \hat{\lambda}_k^2$$

Prova. Considere a parte descartada da decomposição espectral,

$$\mathbf{E} = \mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}^T = \sum_{k=m+1}^p \hat{\lambda}_k \hat{\mathbf{e}}_k \hat{\mathbf{e}}_k^T$$

Para $j \neq j'$, o elemento $e_{jj'}$ é o resíduo $s_{jj'} - \sum_{k=1}^m \hat{\ell}_{jk} \hat{\ell}_{j'k}$, pois $\hat{\Psi}$ não contribui fora da diagonal. A soma dos quadrados desses elementos não pode exceder a soma dos quadrados de todos os elementos de $\mathbf{E}$. Pela ortonormalidade dos autovetores e pela Definição 2.6,

$$\sum_{j \neq j'} e_{jj'}^2 \leq \|\mathbf{E}\|_F^2 = \sum_{k=m+1}^p \hat{\lambda}_k^2$$

□

A Proposição 6.2 dá um critério prático. Se a soma dos quadrados dos autovalores descartados for pequena, a soma dos quadrados dos resíduos também será. A comparação deve levar em conta a escala das variáveis; com variáveis padronizadas, os resíduos medem diretamente as diferenças entre as correlações observadas e as reproduzidas.

Os autovalores retidos também têm leitura direta nas cargas. Somando os quadrados de uma linha de $\hat{\mathbf{L}}$, obtemos a comunalidade estimada de X_j . Somando os de uma coluna, obtemos a parcela da variância total atribuída ao fator correspondente.

Proposição 6.3 (Autovalores e variância explicada). *Na solução por componentes principais, a soma dos quadrados das cargas do fator k é o autovalor retido correspondente,*

$$\sum_{j=1}^p \hat{\ell}_{jk}^2 = \hat{\lambda}_k$$

Assim, a proporção da variância total explicada pelo fator k é $\hat{\lambda}_k / \text{tr}(\mathbf{S})$.

Prova. A coluna k de $\hat{\mathbf{L}}$ é $\sqrt{\hat{\lambda}_k} \hat{\mathbf{e}}_k$, com $\hat{\mathbf{e}}_k^T \hat{\mathbf{e}}_k = 1$. Logo,

$$\sum_{j=1}^p \hat{\ell}_{jk}^2 = \hat{\lambda}_k \hat{\mathbf{e}}_k^T \hat{\mathbf{e}}_k = \hat{\lambda}_k$$

Basta dividir por $\text{tr}(\mathbf{S})$, a variância total das variáveis. □

Exemplo 6.2. Vamos aplicar a Definição 6.2 aos gastos familiares do Exemplo 6.1, para os quais já conhecemos a resposta exata. Como os gastos estão padronizados, usamos $\mathbf{R}$ no lugar de $\mathbf{S}$. Seus autovalores são 1.975, 0.593 e 0.431. Retendo o primeiro, as cargas estimadas são $\sqrt{1.975} \hat{\mathbf{e}}_1$, o que dá

$$\hat{\mathbf{L}}_{\text{CP}} = \begin{pmatrix} 0.847 \\ 0.817 \\ 0.768 \end{pmatrix}$$

contra o valor exato $(0.8, 0.7, 0.6)^T$. Todas as cargas vieram infladas, e as communalidades também, que ficaram em 0.718, 0.668 e 0.591 no lugar de 0.64, 0.49 e 0.36. Subtraindo de $\mathbf{R}$ a matriz reproduzida pelo modelo, obtemos a matriz de resíduos

$$\mathbf{R} - (\hat{\mathbf{L}}_{\text{CP}} \hat{\mathbf{L}}_{\text{CP}}^T + \hat{\mathbf{\Psi}}) = \begin{pmatrix} 0 & -0.132 & -0.171 \\ -0.132 & 0 & -0.208 \\ -0.171 & -0.208 & 0 \end{pmatrix}$$

Na solução exata do Exemplo 6.1, todos os resíduos eram nulos. Aqui, apenas a diagonal é reproduzida exatamente; os resíduos negativos fora dela mostram que as correlações foram superestimadas. Reproduzir as variâncias, portanto, não garante reproduzir as correlações.

As cargas ficam superestimadas porque a ACP procura direções que retenham a **variância total** das variáveis, incluindo a parte específica. Ao considerar essa parcela na extração dos fatores, o método atribui a eles mais variância do que os fatores comuns explicam. Na AF, buscamos explicar a **covariância compartilhada**, distinguindo a variância explicada pelos fatores comuns da variância específica de cada variável.

O método de componentes principais exige pouco esforço computacional e funciona razoavelmente bem quando as communalidades são altas, pois a variância específica é pequena e a superestimação das cargas tende a ser menor. Também pode fornecer estimativas iniciais para os métodos iterativos.

6.2.2 Fatoração do eixo principal

No método de componentes principais, a diagonal de $\mathbf{S}$ representa toda a variância das variáveis, incluindo a parte específica. Pelo modelo fatorial, porém, $\mathbf{\Sigma} - \mathbf{\Psi} = \mathbf{L}\mathbf{L}^T$. Se retirássemos a parcela específica, restaria justamente a matriz que queremos representar pelas cargas. A **fatoração do eixo principal** segue essa ideia. Como não conhecemos $\mathbf{\Psi}$, usamos estimativas das comunalidades para construir a **matriz de covariâncias reduzida**,

$$\mathbf{S}_r = \mathbf{S} - \hat{\mathbf{\Psi}}, \quad \hat{\psi}_j = s_{jj} - \hat{h}_j^2$$

Sua diagonal contém as comunalidades estimadas. As covariâncias fora dela permanecem iguais às observadas, pois, pelo modelo, já são inteiramente atribuídas aos fatores comuns.

Para começar, precisamos de estimativas iniciais das comunalidades. Uma escolha usual é o **coeficiente de determinação** R_j^2 da regressão linear de X_j nas demais variáveis, com intercepto. Esse é o R^2 da regressão, que mede a proporção da variância de X_j explicada pelas demais variáveis. Ele também é o quadrado da correlação entre os valores observados e os previstos pela regressão.

No modelo fatorial, a parte específica de X_j não se correlaciona com as demais variáveis, então essa previsão se apoia na sua parte comum. Isso justifica usar R_j^2 como estimativa inicial da comunalidade, ainda que ele possa representar apenas parte da variância comum. Quando $\mathbf{R}$ é invertível, podemos calcular esse coeficiente diretamente pela matriz de correlações. Na escala padronizada, a variância residual da regressão é $1/[\mathbf{R}^{-1}]_{jj}$ e a variância total vale 1, de modo que

$$\hat{h}_j^2 = R_j^2 = 1 - \frac{1}{[\mathbf{R}^{-1}]_{jj}}$$

Essas estimativas entram na diagonal da **matriz de correlações reduzida**, $\mathbf{R}_r = \mathbf{R} - \hat{\mathbf{\Psi}}$. Para trabalhar na escala original, multiplicamos a proporção inicial por s_{jj} e usamos $\mathbf{S}_r$.

Decompomos a matriz reduzida e construímos as cargas com os m maiores autovalores e seus autovetores, como na Definição 6.2. Os autovalores retidos precisam ser positivos. Essas cargas, por sua vez, implicam comunalidades $\hat{h}_j^2 = \sum_{k=1}^m \hat{\ell}_{jk}^2$, que podem diferir dos palpites usados na diagonal. Atualizamos a diagonal com esses valores e repetimos a extração até que as comunalidades mudem menos que uma tolerância escolhida. Se o procedimento converge, as comunalidades usadas para extrair as cargas passam a concordar com as obtidas a partir delas. A Figura 6.2 resume esse ciclo.

Exemplo 6.3. Nos dados de gastos, começamos com as comunalidades 0.386, 0.343 e 0.264 no lugar dos uns da diagonal de $\mathbf{R}$. A primeira extração devolve cargas $(0.714, 0.670, 0.595)^T$, já mais próximas da resposta exata do que as do método de componentes principais. Seus quadrados dão novas comunalidades, aproximadamente 0.509, 0.449 e 0.354, que entram na diagonal para a próxima extração. As correlações fora da diagonal permanecem iguais.

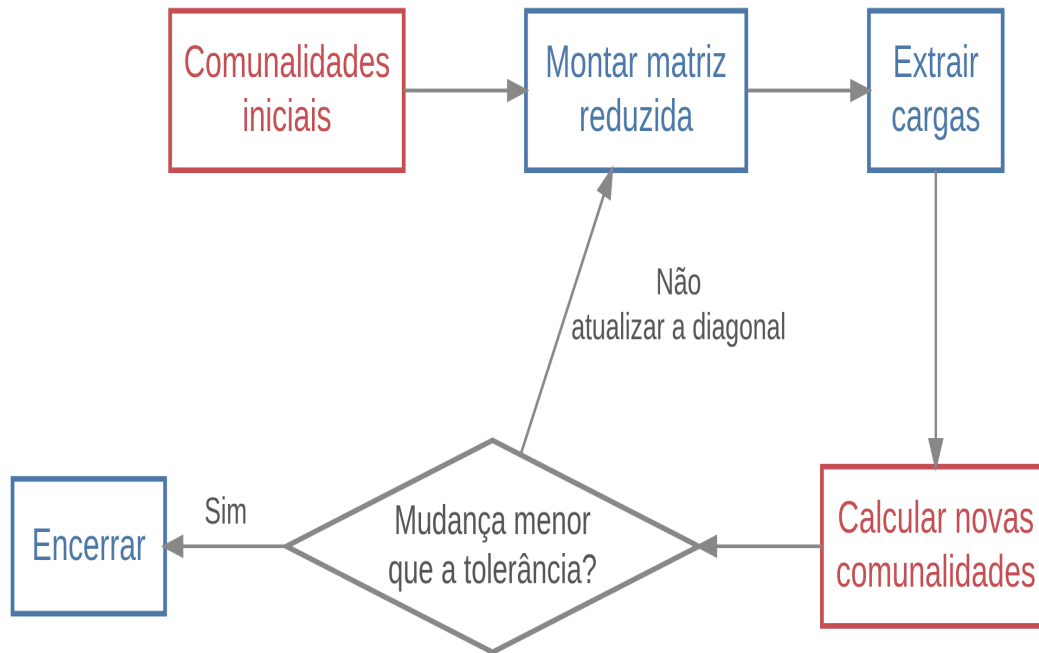


Figura 6.2: Atualização das comunalidades na fatoração do eixo principal. Apenas a diagonal da matriz reduzida muda entre as extrações.

Na décima iteração, as cargas são $(0.796, 0.703, 0.601)^T$. Continuando as atualizações, o procedimento converge para $(0.8, 0.7, 0.6)^T$, recuperando a solução exata do Exemplo 6.1, com $\hat{\psi} = (0.36, 0.51, 0.64)^T$ e todos os resíduos nulos no limite.

O código abaixo reproduz essas atualizações para um fator, usando uma tolerância de 10^{-8} para a maior mudança nas comunalidades.

```

Iteração 1: cargas = [0.714 0.67 0.595]
Iteração 10: cargas = [0.796 0.703 0.601]
Cargas finais: [0.8 0.7 0.6]
Variâncias específicas: [0.36 0.51 0.64]

```

Não há garantia de convergência, e a iteração pode levar alguma comunalidade acima da variância observada, produzindo uma variância específica negativa. Na escala padronizada, isso ocorre quando a comunalidade supera 1, um caso de Heywood, que discutiremos em breve.

6.2.3 Máxima verossimilhança

Nenhum dos dois métodos anteriores tem um modelo probabilístico por trás, e por isso nenhum deles permite testar o número de fatores. Para isso, precisamos de uma distribuição. Supondo que os fatores

e os erros sejam normais, $\mathbf{F} \sim N_m(\mathbf{0}, \mathbf{I}_m)$ e $\epsilon \sim N_p(\mathbf{0}, \Psi)$, independentes entre si, o vetor observado herda a normalidade (ver Capítulo 3) e temos $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T + \Psi)$. Somando os logaritmos das densidades de uma amostra aleatória $\mathbf{x}_1, \dots, \mathbf{x}_n$ e descartando constantes, a log-verossimilhança é

$$\ell(\boldsymbol{\mu}, \Sigma) = -\frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

Como no caso univariado, o máximo em $\boldsymbol{\mu}$ é atingido em $\bar{\mathbf{x}}$. Feita essa troca, cada parcela da soma é um número, igual ao seu próprio traço, e a propriedade cíclica do traço junta todas elas

$$\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^T \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) = \text{tr} \left(\Sigma^{-1} \underbrace{\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T}_{n\mathbf{S}_n} \right)$$

em que $\mathbf{S}_n$ é a matriz de covariâncias amostral com divisor n no lugar de $n - 1$. Sobra

$$\ell(\mathbf{L}, \Psi) = -\frac{n}{2} [\ln |\Sigma| + \text{tr} (\Sigma^{-1} \mathbf{S}_n)]$$

com $\Sigma = \mathbf{L}\mathbf{L}^T + \Psi$. Somando e subtraindo termos que não dependem dos parâmetros, maximizar ℓ equivale a minimizar

$$F(\Sigma) = \ln |\Sigma| + \text{tr} (\Sigma^{-1} \mathbf{S}_n) - \ln |\mathbf{S}_n| - p$$

Essa função mede a distância entre a matriz implicada pelo modelo e a observada, e cumpre aqui o papel que a soma de quadrados dos resíduos cumpre na regressão. Escrita em termos dos autovalores $a_1, \dots, a_p$ de $\Sigma^{-1} \mathbf{S}_n$, ela vira $F = \sum_k (a_k - \ln a_k - 1)$, e cada parcela é não negativa e só se anula quando $a_k = 1$. Logo $F \geq 0$, com igualdade apenas quando $\Sigma = \mathbf{S}_n$.

A minimização não tem forma fechada e sai de algoritmos iterativos. Há ainda uma redundância em $\mathbf{L}$, porque girar os fatores não altera $\mathbf{L}\mathbf{L}^T$ e, portanto, não altera F , como veremos na Seção 6.4. É preciso então impor alguma condição que fixe a orientação, e a escolha convencional é exigir que $\mathbf{L}^T \Psi^{-1} \mathbf{L}$ seja diagonal.

O método tem uma vantagem prática que os anteriores não têm. Se uma variável muda de unidade de medida, a solução de máxima verossimilhança acompanha a mudança, reescalando a carga e a variância específica daquela variável e mais nada. É exatamente o que o modelo prevê para a padronização, e por isso analisar $\mathbf{S}$ ou $\mathbf{R}$ leva à mesma solução, a menos da escala das cargas. Com máxima verossimilhança, a escolha entre as duas matrizes é só uma questão de leitura.

Ter uma verossimilhança permite também comparar o modelo ajustado com a alternativa irrestrita, em que Σ é qualquer matriz de covariâncias. Nessa alternativa o mínimo é $\hat{\Sigma} = S_n$, com discrepância zero, e a estatística da razão de verossimilhanças fica $2(\ell_{\text{irrestrito}} - \ell_{\text{modelo}}) = nF(\hat{\Sigma})$, isto é, n vezes a discrepância que o modelo com m fatores deixou sobrar. Pode-se mostrar que no ótimo vale $\text{tr}(\hat{\Sigma}^{-1}S_n) = p$, e $F(\hat{\Sigma})$ se reduz ao logaritmo de uma razão de determinantes. O teste para H_0 : “ m fatores bastam” usa

$$\chi^2 = \left[n - 1 - \frac{2p + 4m + 5}{6} \right] \underbrace{\ln \frac{|\hat{L}\hat{L}^T + \hat{\Psi}|}{|S_n|}}_{F(\hat{\Sigma})} \quad (6.3)$$

O fator entre colchetes é uma correção de Bartlett que faz o papel de n e melhora a aproximação em amostras moderadas. Valores grandes indicam que a matriz implicada pelo modelo se afasta demais da observada, e levam à rejeição.

Sob H_0 , a estatística segue aproximadamente uma distribuição χ^2 , e o número de graus de liberdade sai de uma contagem. A Equação 6.2 pede que $p(p+1)/2$ números distintos de Σ sejam reproduzidos por L e Ψ , que trazem $p(m+1)$ parâmetros. Descontada a redundância da rotação, sobram

$$\text{gl} = \frac{1}{2} [(p-m)^2 - (p+m)] \quad (6.4)$$

graus de liberdade. Quando $\text{gl} > 0$, o modelo impõe restrições sobre Σ que os dados podem contrariar, e faz sentido testá-lo. Quando $\text{gl} = 0$, há tantos parâmetros quanto equações e o modelo em geral reproduz Σ exatamente, sem que isso conte como evidência a seu favor. Quando $\text{gl} < 0$, pedimos fatores demais e o modelo sequer está identificado. Com $p = 25$ e $m = 5$, como no Exemplo 6.7, sobram 185 graus de liberdade, e aí o ajuste passa a ser informativo. O exemplo dos gastos está no extremo oposto.

Exemplo 6.4. Ajustando um fator por máxima verossimilhança à matriz de correlações dos gastos familiares, obtemos cargas $(0.8, 0.7, 0.6)^T$ e $\hat{\psi} = (0.36, 0.51, 0.64)^T$, a solução exata do Exemplo 6.1, que o método de componentes principais não alcançava e a fatoração do eixo principal só alcançava depois de iterar. Como $\hat{\Sigma}$ reproduz R sem erro, a discrepância $F(\hat{\Sigma})$ é zero e a estatística do teste também. Isso não conta a favor do modelo. Com $p = 3$ e $m = 1$ temos $\text{gl} = 0$, o ajuste exato já estava garantido pela contagem e o teste não tem o que avaliar.

Esse teste é sensível ao tamanho da amostra. Com n grande, diferenças irrelevantes entre $\hat{\Sigma}$ e S_n produzem estatísticas enormes, e a hipótese acaba rejeitada para todo m que se tente. Usá-lo como critério único de escolha tende a inflar o número de fatores sem ganho interpretativo. O exemplo do Big Five, adiante, mostra isso acontecendo.

A escolha do método depende das suposições que podemos fazer sobre os dados e das dificuldades encontradas no ajuste. A máxima verossimilhança costuma ser usada quando a normalidade é plausível,

pois permite testar o ajuste e tem a propriedade de invariância de escala. Mesmo sem normalidade, minimizar F continua sendo uma forma razoável de aproximar S_n , mas o teste pode deixar de ser confiável. Quando a otimização não converge ou produz casos de Heywood, a fatoração do eixo principal é uma alternativa que não exige uma distribuição específica para os dados. O método de componentes principais pode ser usado numa primeira exploração, lembrando que tende a superestimar as cargas quando as comunalidades são baixas.

6.3 Escolha do número de fatores

A escolha de m exige combinar os critérios numéricos com a interpretação dos fatores. Com poucos fatores, algumas correlações importantes podem ficar sem explicação. Com fatores demais, a solução pode se tornar difícil de interpretar e apresentar casos de Heywood. Os critérios disponíveis são quase todos herdados da ACP (Capítulo 4) e devem ser avaliados em conjunto. Os que usam apenas os autovalores de R podem, inclusive, ser aplicados antes de qualquer estimação.

O critério da **variância explicada** retém fatores até que a proporção acumulada atinja um patamar considerado suficiente. Em AF esse patamar costuma ser bem mais baixo do que em ACP, porque aqui só a parte comum entra na conta e a variância específica fica de fora por construção. O **critério de Kaiser**, aplicado à matriz de correlações, retém os fatores com autovalor maior que 1, sob o argumento de que um fator deve ao menos explicar tanto quanto uma variável isolada. Ele é simples e por isso popular, mas tem a tendência conhecida de superestimar m . O **gráfico do cotovelo** procura visualmente o ponto em que os autovalores passam a decair devagar, e tem o defeito de que cada leitor enxerga o cotovelo num lugar.

A **análise paralela** é o critério mais confiável do conjunto e deve ser o preferido quando estiver disponível. Em vez de comparar os autovalores com um limiar fixo, ela os compara com a distribuição dos autovalores que se obteria em dados sem nenhuma estrutura, mas com o mesmo n e o mesmo p . Permuta-se cada coluna da matriz de dados de forma independente, o que destrói as correlações e preserva as distribuições marginais, recalculam-se os autovalores muitas vezes, e retêm-se apenas os fatores cujos autovalores observados superam um percentil alto dessa distribuição nula. O procedimento é descrito com mais detalhe na Seção 4.11.4.

Quando a estimação é por máxima verossimilhança, o teste da Equação 6.3 ainda permite comparar modelos aninhados, com a ressalva sobre amostras grandes já registrada.

Quando os critérios indicam números diferentes de fatores, vale ajustar as soluções candidatas e comparar a interpretação de seus fatores à luz do problema estudado.

Depois de estimar, há ainda um sinal de alerta a observar. Como ψ_j é uma variância, precisa ser não negativa, e como h_j^2 é uma parcela da variância total, não pode ultrapassá-la. Nada no sistema de equações impõe isso, e soluções que violam essas restrições aparecem com alguma frequência. São os chamados **casos de Heywood**.

Exemplo 6.5 (Caso de Heywood). Suponha que, para três variáveis padronizadas, a matriz de correlações observada seja

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

Repetindo a conta do Exemplo 6.1 para a terceira variável,

$$\ell_3^2 = \frac{(\ell_1\ell_3)(\ell_2\ell_3)}{\ell_1\ell_2} = \frac{0.9 \times 0.7}{0.4} = 1.575$$

A comunalidade $h_3^2 = 1.575$ excede a variância total da variável, que é 1, e a variância específica correspondente seria $\psi_3 = 1 - 1.575 = -0.575$. Não existe modelo de um fator compatível com essas correlações.

Um número inadequado de fatores, tanto por excesso quanto por falta, pode levar a estimativas negativas de variância específica. Em amostras pequenas, a variabilidade amostral também pode produzir casos de Heywood. Quando isso acontece, vale reavaliar o número de fatores e as variáveis incluídas no modelo. Substituir a estimativa negativa por zero não resolve a dificuldade de ajuste que levou a esse resultado.

6.4 Rotação fatorial

A orientação dos fatores não é determinada pelo modelo. Podemos girar os eixos fatoriais e alterar as cargas sem mudar a matriz de covariâncias reproduzida. A rotação usa essa liberdade para procurar uma matriz de cargas mais fácil de interpretar.

Seja $\mathbf{T}$ uma matriz ortogonal $m \times m$, com $\mathbf{T}^T \mathbf{T} = \mathbf{I}_m$. Podemos definir novas cargas e novos fatores por

$$\mathbf{L}^* = \mathbf{L}\mathbf{T}, \quad \mathbf{F}^* = \mathbf{T}^T \mathbf{F}$$

O produto $\mathbf{L}^* \mathbf{F}^* = \mathbf{L}\mathbf{F}$ permanece o mesmo, assim como a parte comum da matriz de covariâncias,

$$\mathbf{L}^* (\mathbf{L}^*)^T = \mathbf{L} \mathbf{T} \mathbf{T}^T \mathbf{L}^T = \mathbf{L} \mathbf{L}^T$$

Os fatores continuam com média zero e $\text{Cov}(\mathbf{F}^*) = \mathbf{I}_m$. As variâncias específicas e as comunalidades também não mudam. Portanto, a rotação ortogonal preserva a matriz de covariâncias reproduzida e todos os resíduos do ajuste.

O que muda é a distribuição das cargas entre os fatores. Procuramos uma **estrutura simples**, com poucas cargas altas em módulo por variável e as demais próximas de zero. Assim, podemos interpretar cada fator pelo que as variáveis mais associadas a ele têm em comum. Nem todo conjunto de dados admite uma separação clara, e a rotação não garante que cada variável se associe a apenas um fator.

6.4.1 Varimax

A **varimax** escolhe uma rotação ortogonal que torna as cargas ao quadrado mais diferentes entre si dentro de cada coluna. Se um fator tem cargas parecidas em todas as variáveis, pouco sabemos sobre quais delas o caracterizam. Se tem cargas altas em um grupo e quase nulas nas demais, a interpretação fica mais clara.

Partimos das cargas estimadas $\hat{L}$. Cada matriz ortogonal T produz cargas giradas $\hat{L}^* = \hat{L}T$, com elementos $\hat{\ell}_{jk}^*$. O critério varimax eleva essas cargas ao quadrado, calcula a variância dos quadrados em cada coluna e soma sobre os fatores,

$$V = \sum_{k=1}^m \frac{1}{p} \sum_{j=1}^p \left[(\hat{\ell}_{jk}^*)^2 - \frac{1}{p} \sum_{i=1}^p (\hat{\ell}_{ik}^*)^2 \right]^2 \quad (6.5)$$

As cargas giradas mudam conforme o giro, e V muda junto. A varimax procura, entre todas as matrizes ortogonais T , aquela que deixa V o maior possível. Uma maneira de fazer a busca é girar um par de eixos por vez, mantendo os demais fixos e escolhendo o giro que mais aumenta V . Percorremos os pares de fatores e repetimos o processo até que a melhora no critério seja menor que uma tolerância. Há também algoritmos que atualizam a matriz ortogonal inteira em cada iteração. Como podem existir máximos locais, diferentes orientações iniciais podem ser comparadas.

Varimax e lasso

A busca por cargas pequenas lembra o **lasso**, que usa uma penalização L1, proporcional à soma dos módulos dos coeficientes, e pode tornar alguns deles exatamente nulos.

A varimax também favorece uma representação com poucas cargas altas, mas faz isso girando os fatores, sem acrescentar penalização à estimação. A matriz de covariâncias reproduzida permanece a mesma, e as cargas pequenas não precisam ser exatamente nulas.

Exemplo 6.6 (Testes de habilidades). Considere seis testes padronizados. Os testes X_1 , X_2 e X_3 avaliam vocabulário, compreensão de texto e fluência verbal; X_4 , X_5 e X_6 avaliam cálculo, álgebra e raciocínio numérico. Para acompanhar a análise, usaremos a matriz de correlações hipotética

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.73 & 0.64 & 0.29 & 0.18 & 0.23 \\ 0.73 & 1.00 & 0.59 & 0.36 & 0.24 & 0.28 \\ 0.64 & 0.59 & 1.00 & 0.20 & 0.11 & 0.16 \\ 0.29 & 0.36 & 0.20 & 1.00 & 0.63 & 0.55 \\ 0.18 & 0.24 & 0.11 & 0.63 & 1.00 & 0.48 \\ 0.23 & 0.28 & 0.16 & 0.55 & 0.48 & 1.00 \end{pmatrix}$$

Como as variáveis estão padronizadas, essa é também sua matriz de covariâncias. As correlações são maiores dentro de cada grupo de três testes do que entre os grupos, o que sugere experimentar dois fatores. Primeiro, vamos estimar as cargas pelo método de componentes principais da Definição 6.2.

Os dois maiores autovalores são aproximadamente 2.912 e 1.530. Multiplicando os autovetores correspondentes pelas raízes desses valores, obtemos

$$\hat{\mathbf{L}} \approx \begin{pmatrix} 0.765 & -0.480 \\ 0.797 & -0.382 \\ 0.663 & -0.539 \\ 0.712 & 0.505 \\ 0.600 & 0.602 \\ 0.620 & 0.497 \end{pmatrix}$$

As somas dos quadrados das linhas dão as comunalidades, e o que falta para 1 fica nas variâncias específicas,

$$\hat{\Psi} \approx \text{diag}(0.184, 0.219, 0.270, 0.238, 0.277, 0.369)$$

O primeiro fator tem cargas positivas em todos os testes. O segundo tem cargas negativas nos testes verbais e positivas nos numéricos. Vamos procurar uma orientação que separe melhor esses dois grupos.

Como há apenas dois fatores, basta girar o par de eixos por um ângulo θ . A matriz desse giro é

$$\mathbf{T}(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

Buscamos o ângulo θ para o qual as cargas giradas $\hat{\mathbf{L}}\mathbf{T}(\theta)$ dão o maior valor de V . O código abaixo aplica a varimax, sem normalizar as linhas, usando exatamente o critério da Equação 6.5.

Tabela 6.1: Cargas antes e depois da varimax. As communalidades e as variâncias específicas permanecem iguais.

Teste	F1 inicial	F2 inicial	F1 varimax	F2 varimax	h^2	ψ
X1	0.765	-0.480	0.893	0.137	0.816	0.184
X2	0.797	-0.382	0.853	0.232	0.781	0.219
X3	0.663	-0.539	0.854	0.026	0.730	0.270
X4	0.712	0.505	0.208	0.848	0.762	0.238
X5	0.600	0.602	0.061	0.848	0.723	0.277
X6	0.620	0.497	0.144	0.781	0.631	0.369

Depois da rotação, o primeiro fator reúne os testes verbais, com cargas entre 0.85 e 0.89, e o segundo reúne os numéricos, com cargas entre 0.78 e 0.85. As cargas nos outros fatores ficam abaixo de 0.24 em módulo. Podemos, então, interpretar os fatores como habilidade verbal e habilidade matemática.

A matriz encontrada é, aproximadamente,

$$\mathbf{T} = \begin{pmatrix} 0.757 & 0.654 \\ -0.654 & 0.757 \end{pmatrix}$$

Ela corresponde a $\theta \approx -40.8^\circ$. O critério sobe de 0.0144 para 0.2432. A Figura 6.3 mostra como esse giro aproxima os eixos dos dois grupos de testes.

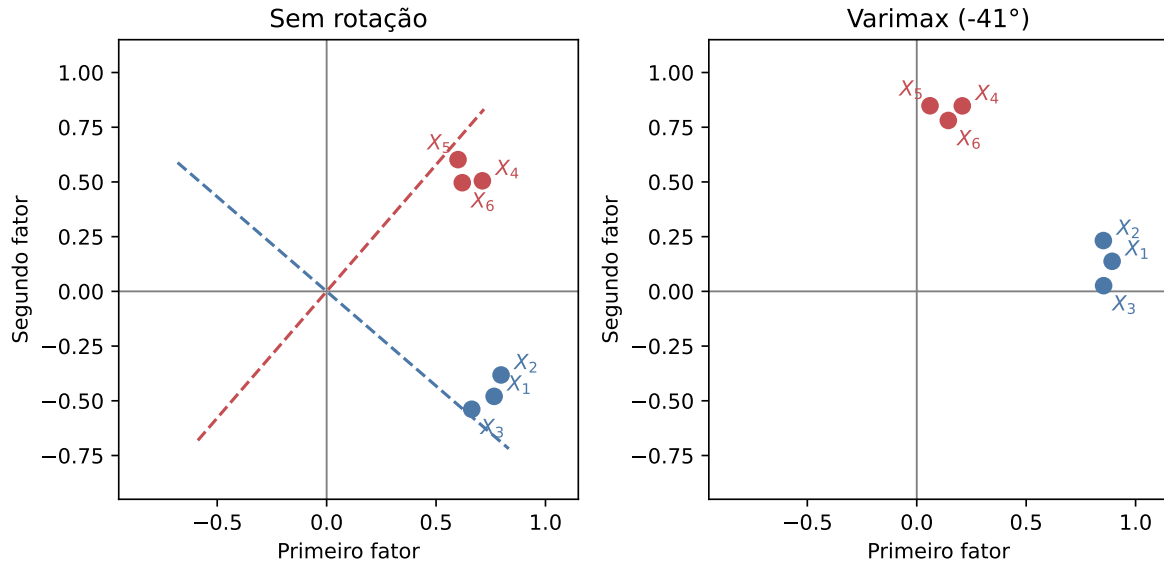


Figura 6.3: Cargas dos seis testes antes e depois da varimax. As linhas tracejadas no painel esquerdo indicam os novos eixos. No painel direito, as cargas são expressas nesses eixos.

Cada ponto representa um teste, e suas coordenadas são as cargas nos dois fatores. A rotação preserva as distâncias entre os pontos e de cada ponto à origem. O quadrado desta última distância é justamente a comunalidade.

6.4.2 Outras rotações

Entre as rotações ortogonais, a **quartimax** maximiza $\sum_{j,k} (\hat{\ell}_{jk}^*)^4$. Como a soma dos quadrados de cada linha é fixa, o critério favorece concentrar as cargas de cada variável em poucos fatores. Pode, por isso, produzir um fator com cargas altas em muitas variáveis, um fator geral. A **equimax** combina os critérios de simplificação das linhas e das colunas, ponderando-os pelo número de variáveis e de fatores.

Também podemos permitir que os fatores sejam correlacionados. No exemplo dos testes, separar habilidade verbal de habilidade matemática não implica que elas sejam não correlacionadas. Numa **rotação oblíqua**, usamos uma matriz invertível Q e definimos

$$L^* = LQ, \quad F^* = Q^{-1}F, \quad \Phi = \text{Cov}(F^*) = Q^{-1}(Q^{-1})^T$$

As colunas de Q são reescaladas para que os fatores tenham variância 1, tornando Φ uma matriz de correlações. A estrutura de covariâncias continua preservada, agora na forma

$$\Sigma = L^* \Phi (L^*)^T + \Psi, \quad L^* \Phi (L^*)^T = LL^T \quad (6.6)$$

Assim, as comunalidades permanecem as mesmas, mas passam a ser os elementos diagonais de $L^* \Phi (L^*)^T$, e não apenas as somas dos quadrados das linhas de L^* .

A **promax** parte de cargas varimax b_{jk} e constrói um alvo $u_{jk} = b_{jk}|b_{jk}|^{r-1}$, geralmente com $r = 4$. Isso equivale a elevar o módulo de cada carga à potência r , preservando o sinal, e acentua a diferença relativa entre cargas grandes e pequenas. Em seguida, ajusta uma transformação oblíqua ao alvo por mínimos quadrados e reescala os fatores para variância 1. Aplicaremos esse procedimento no Exemplo 6.7.

A **oblimin direta** otimiza um critério sobre as cargas, sem construir um alvo. Seu parâmetro modifica o critério de simplicidade, sem fixar ou limitar diretamente as correlações entre fatores. Na versão **quartimin**, esse parâmetro é zero e minimizamos $\sum_j \sum_{k < k'} (\ell_{jk}^*)^2 (\ell_{jk'}^*)^2$, que penaliza cargas simultaneamente altas em dois fatores na mesma variável.

Na solução oblíqua, precisamos distinguir a **matriz de padrões**, L^* , da **matriz de estrutura**, $L^* \Phi$. A primeira contém os coeficientes da regressão linear das variáveis nos fatores, considerando-os em conjunto. A segunda contém as covariâncias $\text{Cov}(x, F^*)$; quando as variáveis observadas também estão padronizadas, seus elementos são correlações. Para interpretar os fatores, olhamos principalmente a matriz de padrões. Uma correlação alta na matriz de estrutura pode vir da correlação com outro fator, mesmo que a carga no fator em questão seja pequena. As duas matrizes coincidem quando $\Phi = I_m$.

A escolha entre rotação ortogonal e oblíqua deve acompanhar a interpretação dos fatores. Se faz sentido admitir associação entre eles, uma rotação oblíqua permite examiná-la por meio de Φ . Um ajuste idêntico nas duas soluções é esperado e não serve para decidir entre elas.

6.5 Escores fatoriais

Com as cargas estimadas, falta situar cada indivíduo nos fatores, tal como fizemos com os escores da ACP. Esses valores servem para representar as observações graficamente e para alimentar análises posteriores, usando os fatores no lugar das variáveis originais.

Só que os escores da ACP eram $\mathbf{X}\mathbf{E}$, uma transformação determinística dos dados, enquanto $\mathbf{F}$ é uma variável aleatória que ninguém viu e que não se escreve como função das variáveis observadas. O melhor que podemos fazer é predizê-la, e métodos diferentes de predição dão respostas diferentes.

O **método da regressão**, também chamado de método de Thomson, trata o problema como a predição de $\mathbf{F}$ a partir de $\mathbf{x}$ que minimiza o erro quadrático médio. Sob o modelo fatorial, $\text{Cov}(\mathbf{F}, \mathbf{x}) = \mathbf{L}^T$ e $\text{Cov}(\mathbf{x}) = \Sigma$, de modo que a regressão linear de $\mathbf{F}$ em $\mathbf{x}$ tem coeficientes $\mathbf{L}^T \Sigma^{-1}$ e

$$\hat{\mathbf{f}}_i = \hat{\mathbf{L}}^T \left(\hat{\mathbf{L}} \hat{\mathbf{L}}^T + \hat{\Psi} \right)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (6.7)$$

Entre as predições lineares, esses escores têm o menor erro quadrático médio. Eles, porém, tendem a ficar mais próximos de zero do que os valores dos fatores que procuram predizer. Esse efeito é chamado de encolhimento e se reflete na variância dos escores, que é menor que 1, a variância de cada fator.

O **método de Bartlett** trata $\mathbf{F}$ como um parâmetro fixo a ser estimado e aplica mínimos quadrados ponderados, usando Ψ^{-1} como peso, já que variáveis com variância específica grande são medidas mais ruidosas do fator e devem pesar menos. O resultado é

$$\hat{\mathbf{f}}_i = \left(\hat{\mathbf{L}}^T \hat{\Psi}^{-1} \hat{\mathbf{L}} \right)^{-1} \hat{\mathbf{L}}^T \hat{\Psi}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (6.8)$$

Esses escores são não viesados, no sentido de que $E[\hat{\mathbf{f}} | \mathbf{F}] = \mathbf{F}$, ao custo de uma variância maior que a do método da regressão. A escolha entre os dois é a escolha usual entre viés e variância. Quando os escores vão ser correlacionados com outras variáveis, o encolhimento do método da regressão pode distorcer as conclusões, e os escores de Bartlett costumam ser preferíveis. Para representação gráfica, a diferença raramente importa.

Quando a solução é oblíqua, ambas as fórmulas se ajustam para acomodar Φ ; no caso da regressão, por exemplo, os coeficientes passam a ser $\Phi \mathbf{L}^T \Sigma^{-1}$.

i Indeterminação dos escores

Mesmo que $\mathbf{L}$ e $\mathbf{\Psi}$ fossem conhecidos exatamente, e com n infinito, os escores fatoriais continuariam indeterminados. Existe toda uma família de vetores aleatórios $\mathbf{F}$ compatíveis com o modelo e com os dados observados, e nada nas observações permite escolher entre eles. Esse resultado é conhecido como indeterminação dos escores fatoriais.

A gravidade do problema depende das comunalidades. Quando elas são altas, os candidatos ficam próximos uns dos outros e a indeterminação é inofensiva; quando são baixas, dois conjuntos de escores igualmente válidos podem até correlacionar mal entre si.

6.6 Exemplo prático

Exemplo 6.7 (Traços de personalidade). O conjunto `bfi`, do pacote `R psych`, reúne respostas de 2800 pessoas a 25 itens de personalidade, medidos em escala de 1 a 5. Os itens foram construídos para medir cinco traços, o chamado Big Five, com cinco itens cada: amabilidade (A), conscienciosidade (C), extroversão (E), neuroticismo (N) e abertura à experiência (O). Queremos saber se a análise fatorial, olhando apenas para as correlações entre os itens e sem saber nada sobre essa divisão, consegue recuperá-la.

Alguns itens são redigidos ao contrário dos demais para desencorajar respostas automáticas. O item A1, por exemplo, afirma indiferença aos sentimentos alheios, de modo que discordar dele indica *mais* amabilidade. Sem recodificar, esses itens aparecem com cargas negativas no fator do seu próprio bloco, o que não está errado, mas atrapalha a leitura, e mais adiante inverteria também o sinal das correlações entre os fatores. Subtrair de 6 alinha todos os itens no mesmo sentido.

Restaram 2436 respostas completas para 25 itens, que usaremos para ajustar o modelo.

Quantos fatores extrair? O critério de Kaiser e a análise paralela não concordam.

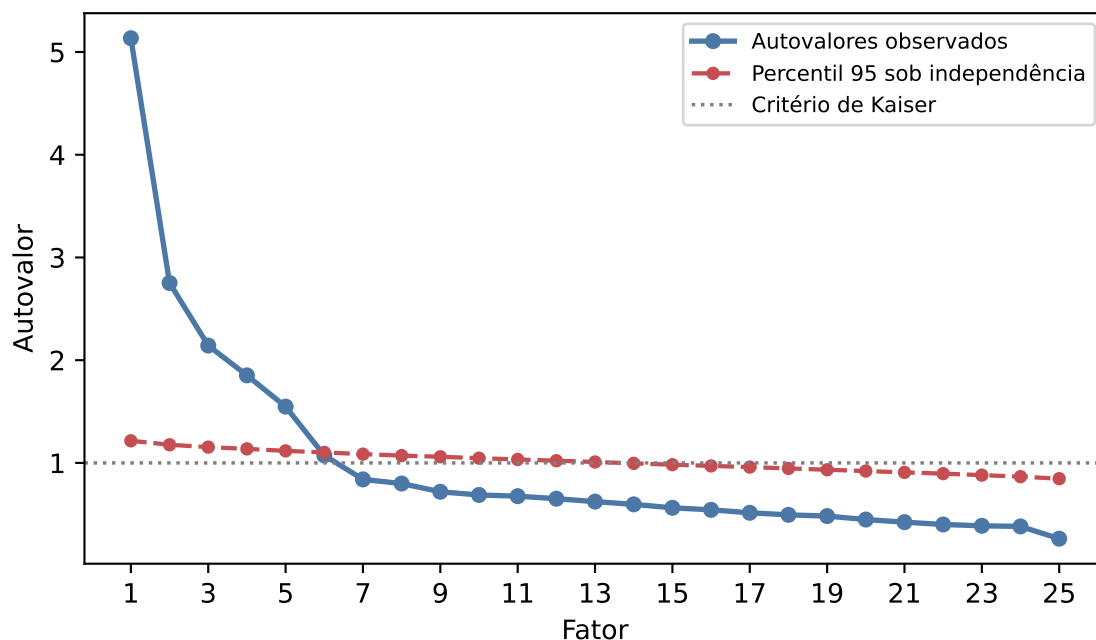


Figura 6.4: Autovalores da matriz de correlações dos 25 itens, comparados ao critério de Kaiser e ao percentil 95 dos autovalores obtidos sob independência por permutação.

O critério de Kaiser retém 6 fatores e a análise paralela retém 5. A discrepância é a esperada, com o critério de Kaiser pecando pelo excesso. O sexto autovalor vale 1.07, mal ultrapassando 1, enquanto o percentil 95 sob independência naquela posição é 1.10. Ficamos com cinco fatores, decisão que a análise paralela sustenta e que coincide com a estrutura teórica dos itens.

Ajustamos então o modelo por máxima verossimilhança, com rotação varimax.

Tabela 6.2: Cargas fatoriais após rotação varimax e comunalidades dos 25 itens. Cargas de módulo inferior a 0.3 foram omitidas.

	N	E	C	A	O	Comunalidade
A1				0.39		0.17
A2				0.6		0.42
A3				0.66		0.53
A4				0.45		0.31
A5		0.35		0.58		0.49
C1			0.53			0.34
C2			0.62			0.43
C3			0.55			0.32
C4			0.65			0.49

Tabela 6.2: Cargas fatoriais após rotação varimax e comunalidades dos 25 itens. Cargas de módulo inferior a 0.3 foram omitidas.

	N	E	C	A	O	Comunalidade
C5			0.57			0.44
E1		0.59				0.37
E2		0.67				0.55
E3		0.49		0.31	0.31	0.44
E4		0.61		0.36		0.53
E5		0.49	0.31			0.41
N1	0.82					0.73
N2	0.79					0.66
N3	0.71					0.52
N4	0.56	-0.37				0.49
N5	0.52					0.34
O1					0.52	0.33
O2					0.45	0.26
O3					0.61	0.48
O4					0.37	0.25
O5					0.51	0.27

A ordem em que os fatores saem da estimação é arbitrária, e os nomes na tabela foram atribuídos depois de olhar para as cargas. Cada bloco de cinco itens se concentra num fator diferente, e a análise recuperou a divisão teórica sem ter acesso a ela.

Alguns itens se ajustam menos à estrutura de cinco fatores. O item A1 tem a menor comunalidade do conjunto, 0.17, de modo que quase 83% da sua variância não é explicada pelos fatores estimados. Essa comunalidade baixa merece atenção, mas, sozinha, não permite concluir que o item mede mal a amabilidade. O item E3 tem cargas em extroversão, amabilidade e abertura, enquanto E5 tem cargas em extroversão e conscienciosidade. Já N4 tem carga substancial em neuroticismo e carga negativa em extroversão, o que combina com o conteúdo do item, que trata de desânimo e abatimento. As comunalidades, em geral entre 0.3 e 0.5, são modestas, o que é normal em itens isolados de questionário e nos lembra que boa parte da variância de cada resposta é específica dela.

Como o ajuste foi por máxima verossimilhança, podemos também aplicar o teste da Equação 6.3. Vale conferir o que ele diz para cinco fatores e para alguns valores maiores de m .

	gl	qui-quadrado	p-valor
m			
5	185	1490.6	1.2e-202
6	165	896.7	5.7e-101
7	146	619.2	1.6e-59

	gl	qui-quadrado	p-valor
m			
8	128	438.2	1.5e-35
9	111	316.1	1.4e-21
10	95	224.8	1.6e-12
11	80	147.8	6.2e-06

Com cinco fatores, a estatística vale 1491 para 185 graus de liberdade, e o teste rejeita a hipótese de ajuste. Se usássemos apenas esse resultado para decidir, aumentaríamos o número de fatores. Mas o teste continua rejeitando com seis, sete e até onze fatores, bem além do que a análise paralela sugere, enquanto a solução com cinco recuperou a estrutura teórica dos itens. Esse resultado ilustra a sensibilidade do teste ao tamanho da amostra. Com mais de dois mil respondentes, mesmo discrepâncias pequenas entre $\mathbf{R}$ e $\hat{\Sigma}$ podem levar à rejeição.

Os cinco traços serão mesmo não correlacionados, como a rotação varimax impôs? O promax responde a isso, partindo da própria solução varimax. Vamos implementá-lo à mão, tanto porque o procedimento é curto quanto porque assim fica explícito de onde vêm a matriz de padrões e a matriz de correlações entre fatores.

Tabela 6.3: Correlações entre os cinco fatores estimadas pela rotação promax.

	N	E	C	A	O
N	1.00	-0.37	-0.25	0.06	0.02
E	-0.37	1.00	0.37	0.25	0.14
C	-0.25	0.37	1.00	0.22	0.24
A	0.06	0.25	0.22	1.00	0.21
O	0.02	0.14	0.24	0.21	1.00

A transformação oblíqua não altera o ajuste. A raiz do erro quadrático médio dos resíduos fora da diagonal é 0.0286 para a solução oblíqua e exatamente a mesma para a varimax, porque a matriz de padrões $\mathbf{L}_{\text{pro}}$ e as cargas varimax $\mathbf{L}_{\text{var}}$ satisfazem $\mathbf{L}_{\text{pro}}\Phi\mathbf{L}_{\text{pro}}^T = \mathbf{L}_{\text{var}}\mathbf{L}_{\text{var}}^T$.

As correlações entre fatores não são desprezíveis e vão na direção que a literatura de personalidade descreve. Neuroticismo correlaciona-se negativamente com extroversão (-0.37) e com conscienciosidade (-0.25), enquanto extroversão, conscienciosidade, amabilidade e abertura formam um bloco de correlações positivas moderadas. A suposição de ortogonalidade, portanto, era uma simplificação, ainda que não das mais graves.

Nos números, a distinção entre as duas matrizes da solução oblíqua fica visível.

Tabela 6.4: Matriz de padrões e matriz de estrutura para quatro itens selecionados. As duas diferem porque os fatores são correlacionados.

	Padrão					Estrutura				
	N	E	C	A	O	N	E	C	A	O
A3	-0.03	0.18	0.00	0.66	0.02	-0.06	0.36	0.22	0.71	0.18
C4	-0.10	-0.02	0.68	-0.06	0.01	-0.27	0.25	0.69	0.08	0.16
E2	-0.03	0.71	-0.02	0.06	0.05	-0.28	0.73	0.27	0.24	0.15
N4	0.40	-0.39	-0.13	0.09	0.09	0.59	-0.55	-0.33	0.01	0.04

Veja o item C4, sobre fazer as coisas pela metade. Seu coeficiente de padrão em neuroticismo é pequeno, mas sua correlação com o fator, na matriz de estrutura, é bem maior em módulo. O item mede conscienciosidade, e esse fator anda junto com neuroticismo. É por esse caminho indireto que o item aparece associado a neuroticismo. Ler a matriz de estrutura como se fossem cargas levaria a atribuir ao item um conteúdo que ele não tem, e por isso os fatores são nomeados pela matriz de padrões.

6.7 Exercícios

Exercício 6.1. Considere o modelo fatorial ortogonal $\mathbf{x} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\epsilon}$, sob as suposições da Definição 6.1.

- Mostre que $\text{Cov}(\mathbf{x}) = \mathbf{LL}^T + \boldsymbol{\Psi}$.
- Mostre que $\text{Cov}(X_j, F_k) = \ell_{jk}$, isto é, que a carga é a covariância entre a variável e o fator.
- Suponha agora que $\boldsymbol{\Psi}$ não seja diagonal, mas uma matriz de covariâncias qualquer. Que restrições o modelo ainda impõe às covariâncias observadas? O que isso implica para sua interpretação?

Exercício 6.2. Um modelo com $m = 2$ fatores foi ajustado a quatro variáveis padronizadas, produzindo as cargas

$$\hat{\mathbf{L}} = \begin{pmatrix} 0.8 & 0.2 \\ 0.7 & -0.3 \\ 0.3 & 0.8 \\ 0.4 & 0.6 \end{pmatrix}$$

- Calcule as communalidades e as variâncias específicas das quatro variáveis.
- Calcule a correlação entre X_1 e X_3 implicada pelo modelo, e a correlação entre X_3 e X_4 .
- Que proporção da variância total das quatro variáveis é explicada por cada fator?

Exercício 6.3. Seja T a matriz de rotação no plano,

$$T = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

- a) Verifique que T é ortogonal e que, para qualquer θ , as cargas rotacionadas $\hat{L}^* = \hat{L}T$ produzem a mesma matriz $\hat{L}^*(\hat{L}^*)^T$.
- b) Mostre que as comunalidades são invariantes à rotação.
- c) Aplique a rotação com $\theta = 30$ às cargas do Exercício 6.2 e verifique numericamente os dois itens anteriores.
- d) Se a rotação não altera nem o ajuste nem as comunalidades, em que sentido uma solução rotacionada é melhor que outra?

Exercício 6.4. Três variáveis padronizadas têm matriz de correlações

$$R = \begin{pmatrix} 1.0 & 0.8 & 0.3 \\ 0.8 & 1.0 & 0.4 \\ 0.3 & 0.4 & 1.0 \end{pmatrix}$$

- a) Resolva o sistema de um fator e obtenha as três cargas.
- b) Calcule as comunalidades e as variâncias específicas. A solução é admissível?
- c) Use a Equação 6.4 para determinar os graus de liberdade do modelo. O que esse número diz sobre a possibilidade de testar o ajuste?

Exercício 6.5. Considere um modelo de um fator com $p = 3$ variáveis padronizadas, cargas $L = (0.9, 0.6, 0.3)^T$ e as variâncias específicas correspondentes.

- a) Escreva os coeficientes do escore de Bartlett dado pela Equação 6.8. Qual variável recebe o maior peso, e por quê?
- b) Compare os coeficientes do escore de regressão da Equação 6.7 com os de Bartlett. Os dois métodos atribuem os mesmos pesos relativos às variáveis? E produzem escores na mesma escala?
- c) Explique por que os escores de regressão têm variância menor que 1, enquanto os fatores verdadeiros têm variância 1.

Exercício 6.6. Uma equipe de inteligência de mercado aplicou um questionário a 300 consumidores sobre a percepção de uma marca de relógios inteligentes. Os seis itens, medidos de 1 a 7, foram os seguintes:

- X_1 : as medições de batimentos cardíacos são precisas

- X_2 : o GPS integrado registra minhas rotas sem falhas
- X_3 : o design é moderno e elegante
- X_4 : as opções de pulseira são esteticamente atraentes
- X_5 : o relógio resiste a quedas e ao uso intenso em treinos
- X_6 : a bateria dura o tempo prometido pelo fabricante

A matriz de correlações amostral foi

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.72 & 0.15 & 0.10 & 0.65 & 0.60 \\ 0.72 & 1.00 & 0.12 & 0.08 & 0.70 & 0.62 \\ 0.15 & 0.12 & 1.00 & 0.78 & 0.10 & 0.14 \\ 0.10 & 0.08 & 0.78 & 1.00 & 0.05 & 0.11 \\ 0.65 & 0.70 & 0.10 & 0.05 & 1.00 & 0.58 \\ 0.60 & 0.62 & 0.14 & 0.11 & 0.58 & 1.00 \end{pmatrix}$$

- Identifique a estrutura de blocos na matriz. Quantos fatores você espera encontrar, e como os interpretaria?
- A extração por componentes principais com $m = 2$ produziu $\hat{\lambda}_1 = 3.01$ e $\hat{\lambda}_2 = 1.71$. Que proporção da variância total os dois fatores explicam?
- Quantos graus de liberdade tem esse modelo? Seria possível testar seu ajuste?
- O que a rotação varimax faria com essa solução? Você esperaria uma mudança grande nas cargas?

Exercício 6.7 (Composição química de vidros romanos). O conjunto dados/RBGlass1.csv reúne medidas de composição química de fragmentos de vidro romano. Conduza uma análise fatorial completa e justifique a padronização das variáveis, o número de fatores, o método de estimação e o tipo de rotação. Se quiser complementar a análise com diagnósticos de adequação, consulte a Seção 6.8.1. Compare a solução obtida com uma ACP dos mesmos dados e discuta em que as duas diferem, tanto nos resultados quanto naquilo que cada uma se propõe a responder.

6.8 Tópicos avançados

6.8.1 Adequação dos dados

Na análise fatorial, procuramos explicar as correlações entre as variáveis por meio de fatores comuns. Se essas correlações forem muito pequenas, pode ser difícil encontrar fatores que expliquem uma parcela relevante da variância das variáveis. Antes de ajustar o modelo, podemos usar o teste de Bartlett e o índice KMO para avaliar se os dados são adequados à análise.

O **teste de esfericidade de Bartlett** avalia a hipótese $H_0 : \boldsymbol{\rho} = \mathbf{I}_p$, ou seja, de que todas as correlações entre variáveis distintas são nulas na população. O teste usa o determinante da matriz de correlações

amostral, que tende a ficar próximo de 1 quando as correlações são pequenas e diminui quando há maior dependência linear entre as variáveis. A estatística do teste é dada por

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

Sob H_0 e supondo normalidade multivariada, essa estatística segue aproximadamente uma distribuição χ^2 com $p(p-1)/2$ graus de liberdade. Valores elevados levam à rejeição da hipótese nula. Mas rejeitar essa hipótese não basta para concluir que a análise fatorial é adequada. Em amostras grandes, o teste pode detectar correlações pequenas, mesmo que elas sejam pouco úteis para o modelo fatorial.

A medida de **Kaiser-Meyer-Olkin** (KMO) compara as correlações simples com as correlações parciais. A correlação parcial mede a associação entre duas variáveis depois de descontar a relação linear de cada uma com as demais. Quando a correlação continua alta após esse ajuste, a associação entre as duas é pouco explicada pelas outras variáveis. Para a análise fatorial, esperamos que as correlações parciais sejam pequenas em relação às simples, o que sugere que as associações podem ser explicadas por fatores comuns. O KMO é calculado por

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

em que $r_{jj'}$ é a correlação simples e $a_{jj'}$ é a correlação parcial entre X_j e $X_{j'}$, ajustada pelas demais variáveis. O índice varia entre 0 e 1. Quanto menores forem as correlações parciais em relação às simples, mais próximo de 1 será o KMO e mais favorável será a indicação para a análise fatorial.

Como orientação prática, valores acima de 0.8 indicam boa adequação, e valores acima de 0.6 costumam ser aceitáveis. Abaixo de 0.5, a análise fatorial geralmente não é recomendada. Também podemos calcular o índice para cada variável, incluindo nos somatórios apenas os pares em que ela aparece. Essa versão ajuda a identificar quais variáveis se ajustam menos à estrutura de fatores comuns e merecem ser reavaliadas antes de prosseguir.

Podemos aplicar esses diagnósticos aos dados do Big Five preparados no Exemplo 6.7.

```
Bartlett: qui-quadrado = 18146, p-valor = 0.0e+00
KMO global: 0.849
Menor KMO por variável: 0.754 (A1)
```

O teste de Bartlett rejeita a hipótese de que todas as correlações entre itens distintos sejam nulas, mas esse resultado não basta para avaliar a adequação à análise fatorial, sobretudo com uma amostra deste tamanho. O KMO global de 0.849 indica boa adequação, e mesmo o menor KMO individual fica bem acima de 0.5. Esses resultados apoiam o uso da análise fatorial para os dados.

6.8.2 Análise fatorial confirmatória

Tudo o que fizemos até aqui é **análise fatorial exploratória**. Deixamos os dados dizerem quantos fatores existem e quais variáveis pesam em cada um, e só depois, olhando as cargas, atribuímos nomes. A rotação é parte dessa postura, já que serve para encontrar a orientação mais legível entre infinitas equivalentes.

A **análise fatorial confirmatória** inverte a ordem. A estrutura é especificada antes de ver os dados, a partir da teoria, tipicamente fixando em zero as cargas dos itens nos fatores a que eles não pertencem. No exemplo do Big Five, isso significaria declarar que os cinco itens de neuroticismo carregam apenas no fator de neuroticismo, e assim por diante. Com cargas fixadas, a indeterminação rotacional desaparece, não há o que rotacionar, e o modelo fica bem mais fácil de testar. Ou ele reproduz a matriz de correlações observada, ou não.

O ajuste é avaliado pelo teste da Equação 6.3 e por índices construídos para contornar a sensibilidade daquele teste ao tamanho da amostra. A análise fatorial confirmatória é o caso mais simples dos **modelos de equações estruturais**, que permitem ainda especificar relações entre os próprios fatores latentes.

6.8.3 Dados ordinais e correlações policóricas

O exemplo deste capítulo tratou respostas em escala de 1 a 5 como se fossem contínuas, e calculou correlações de Pearson entre elas, como é usual. Uma escala Likert, porém, não é contínua nem tem espaçamento garantido entre categorias, e a correlação de Pearson entre variáveis ordinais é sistematicamente atenuada em relação à correlação entre as variáveis contínuas que as originaram, especialmente quando as respostas se concentram nos extremos.

A alternativa é a **correlação policórica**, que supõe que por trás de cada resposta ordinal existe uma variável contínua normal, cortada em faixas por limiares desconhecidos, e estima a correlação entre essas variáveis latentes por máxima verossimilhança. A análise fatorial é então conduzida sobre a matriz de correlações policóricas. O efeito prático costuma ser um aumento das comunalidades e, com alguma frequência, uma estrutura mais limpa. Em troca, a estimação é mais pesada e exige amostras maiores para ser estável.

Com poucas categorias, digamos duas ou três, a diferença é grande e compensa usar a correlação policórica. Com cinco ou mais categorias e respostas razoavelmente distribuídas, como no exemplo do Big Five, a atenuação é modesta e a análise sobre correlações de Pearson tende a levar às mesmas conclusões.

6.8.4 A análise fatorial como modelo probabilístico

A ACP probabilística, mencionada na Seção 4.11.5, supõe que a variância descartada se distribui igualmente por todas as direções não retidas, com $\Sigma = E_q(\Lambda_q - \sigma^2 I_q)E_q^T + \sigma^2 I_p$, isto é, um único

nível de ruído σ^2 comum a todas as variáveis. A análise fatorial surge quando permitimos que cada variável tenha o seu próprio nível de ruído, trocando $\sigma^2 \mathbf{I}_p$ pela matriz diagonal Ψ .

Com um único nível de ruído para todas as variáveis, a ACP probabilística, tal como a ACP usual, depende da escala em que cada variável foi medida, e por isso precisamos padronizar antes de aplicá-la. Já Ψ pode absorver diferenças de escala variável a variável, e é daí que vem a invariância de escala da solução de máxima verossimilhança da análise fatorial, vista na seção de estimação.

7 Análise de correlação canônica

A ACP (Capítulo 4) olha para um único conjunto de variáveis e procura as direções em que ele mais varia. Muitas perguntas, porém, envolvem dois conjuntos. Um médico mede a pressão, o colesterol e a glicemia de seus pacientes e quer saber como esses exames se relacionam com horas de sono, consumo de álcool e atividade física. Um biólogo mede o bico e o corpo de pinguins e quer saber como o formato de um acompanha o tamanho do outro.

Com uma variável de cada lado, bastaria calcular uma correlação. Com várias variáveis de cada lado, são muitas correlações cruzadas, e fica difícil enxergar o padrão numa tabela desse tamanho. A **análise de correlação canônica** (ACC), proposta por Harold Hotelling nos anos 1930, resume essa tabela procurando uma combinação linear de cada conjunto de modo que as duas combinações sejam o mais correlacionadas possível. Em seguida procura um segundo par, não correlacionado com o primeiro, e assim por diante. Os pares encontrados são as **variáveis canônicas**, e as correlações dentro de cada par, as **correlações canônicas**.

Os dois conjuntos entram na análise do mesmo jeito. Nenhum deles faz o papel de resposta, como numa regressão, embora nada impeça que, na hora de interpretar, um deles seja lido como explicação do outro.

7.1 Correlações canônicas

Vamos dividir o vetor aleatório $\mathbf{x}$ em dois blocos, o primeiro com p variáveis e o segundo com q , chamando de primeiro o bloco menor, de modo que $p \leq q$. A matriz de covariâncias se divide da mesma forma,

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Os blocos da diagonal são as matrizes de covariâncias de cada conjunto, e Σ_{12} reúne as covariâncias entre variáveis de conjuntos diferentes. Vamos supor Σ positiva-definida, o que garante que os blocos da diagonal sejam inversíveis. Para uma combinação linear de cada bloco,

$$U = \mathbf{a}^T \mathbf{x}^{(1)}, \quad V = \mathbf{b}^T \mathbf{x}^{(2)}$$

as variâncias e a covariância são

$$\text{Var}(U) = \mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a}, \quad \text{Var}(V) = \mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b}, \quad \text{Cov}(U, V) = \mathbf{a}^T \boldsymbol{\Sigma}_{12} \mathbf{b}$$

Multiplicar $\mathbf{a}$ ou $\mathbf{b}$ por uma constante positiva não muda a correlação entre as duas combinações, então podemos fixar a escala exigindo variância 1 para ambas. Com essa escolha, a correlação é a própria covariância.

Definição 7.1 (Correlações canônicas). As **variáveis canônicas** são pares de combinações lineares $(U_1, V_1), \dots, (U_p, V_p)$, todas com variância 1, em que U_k vem do primeiro bloco e V_k do segundo. O primeiro par é o de maior correlação possível, e cada par seguinte é o de maior correlação entre as combinações que não se correlacionam com as variáveis canônicas anteriores do mesmo bloco. São p pares porque o primeiro bloco só comporta p combinações não correlacionadas entre si. As correlações dentro de cada par, $\rho_1 \geq \rho_2 \geq \dots \geq \rho_p$, são as **correlações canônicas**.

Para encontrar o primeiro par, maximizamos a covariância entre as duas combinações com a restrição de variância 1,

$$\begin{aligned} & \max_{\mathbf{a}, \mathbf{b}} \quad \mathbf{a}^T \boldsymbol{\Sigma}_{12} \mathbf{b} \\ & \text{sujeito a} \quad \mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a} = 1, \quad \mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b} = 1 \end{aligned}$$

Como são duas restrições, usamos dois multiplicadores de Lagrange, e a função a ser maximizada é

$$L(\mathbf{a}, \mathbf{b}, \lambda, \mu) = \mathbf{a}^T \boldsymbol{\Sigma}_{12} \mathbf{b} - \frac{\lambda}{2} (\mathbf{a}^T \boldsymbol{\Sigma}_{11} \mathbf{a} - 1) - \frac{\mu}{2} (\mathbf{b}^T \boldsymbol{\Sigma}_{22} \mathbf{b} - 1)$$

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

$$\boldsymbol{\Sigma}_{12} \mathbf{b} = \lambda \boldsymbol{\Sigma}_{11} \mathbf{a}, \quad \boldsymbol{\Sigma}_{21} \mathbf{a} = \mu \boldsymbol{\Sigma}_{22} \mathbf{b} \quad (7.1)$$

Multiplicando a primeira equação à esquerda por $\mathbf{a}^T$ e a segunda por $\mathbf{b}^T$ e usando as restrições, os dois multiplicadores são iguais à própria correlação entre as combinações,

$$\lambda = \mu = \mathbf{a}^T \boldsymbol{\Sigma}_{12} \mathbf{b} = \rho$$

Isolando $\mathbf{b}$ na segunda equação e substituindo na primeira, chegamos a

$$\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \mathbf{a} = \rho^2 \mathbf{a} \quad (7.2)$$

Assim como na ACP, a solução é uma equação de autovalores. O vetor $\mathbf{a}$ é um autovetor da matriz $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, e o quadrado da correlação é o autovalor correspondente. Para maximizar a correlação, escolhamos o maior autovalor, ρ_1^2 , e um autovetor $\mathbf{a}_1$ associado, reescalado para que U_1 tenha variância 1. O coeficiente do outro bloco sai da segunda equação de Equação 7.1,

$$\mathbf{b}_1 = \frac{1}{\rho_1} \Sigma_{22}^{-1} \Sigma_{21} \mathbf{a}_1$$

Os pares seguintes saem da mesma equação. Repetindo a maximização com as restrições de não correlação, o k -ésimo par corresponde ao k -ésimo maior autovalor, e a k -ésima correlação canônica é a raiz dele. As variáveis canônicas obtidas assim também não se correlacionam com as parceiras dos outros pares, isto é, U_k e V_l são não correlacionadas sempre que $k \neq l$. Cada variável canônica só se correlaciona com a sua parceira, e a associação entre os blocos fica organizada em pares que não se misturam.

Nota

A matriz da Equação 7.2 não é simétrica, e por isso os programas costumam seguir outro caminho, que chega à mesma solução. Eles branqueiam cada bloco com a inversa da raiz quadrada da Definição 2.13 e fazem a decomposição em valores singulares (Definição 2.14) da matriz

$$\mathbf{K} = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$$

Os valores singulares de $\mathbf{K}$ são as correlações canônicas, e os vetores singulares, multiplicados por $\Sigma_{11}^{-1/2}$ e $\Sigma_{22}^{-1/2}$, dão os coeficientes. É esse o caminho do código no Exemplo 7.4.

As correlações canônicas também não mudam quando reescalamos as variáveis, nem, de forma mais geral, quando cada bloco passa por uma transformação linear inversível (Exercício 7.2). Por isso é comum trabalhar com variáveis padronizadas, usando a matriz de correlações ρ no lugar de Σ . Os coeficientes mudam de escala, mas as correlações e as próprias variáveis canônicas continuam as mesmas.

Exemplo 7.1 (Medidas da cabeça de irmãos). No começo do século XX, antropólogos mediram o comprimento e a largura da cabeça do primeiro e do segundo filho de várias famílias para estudar o quanto irmãos se parecem. Vamos usar uma versão simplificada desse estudo, com as quatro medidas padronizadas. O primeiro bloco traz o comprimento e a largura da cabeça do primeiro filho, e o segundo, as mesmas medidas do segundo filho, nessa ordem. A matriz de correlações é

$$\mathbf{R} = \left(\begin{array}{cc|cc} 1.00 & 0.60 & 0.65 & 0.55 \\ 0.60 & 1.00 & 0.55 & 0.65 \\ \hline 0.65 & 0.55 & 1.00 & 0.60 \\ 0.55 & 0.65 & 0.60 & 1.00 \end{array} \right)$$

Dentro de cada filho, comprimento e largura têm correlação 0.60. Entre irmãos, a mesma medida tem correlação 0.65, e medidas trocadas, 0.55. Por enquanto, vamos tratar essas correlações como conhecidas.

Os três blocos da matriz têm a mesma forma, com os dois elementos da diagonal iguais e os dois de fora também, e por isso compartilham os autovetores

$$e_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad e_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Em cada bloco, o autovalor ligado a e_1 é a soma dos dois elementos distintos, e o ligado a e_2 é a diferença. Os blocos da diagonal têm autovalores 1.6 e 0.4, e o bloco cruzado, 1.2 e 0.1. Como todos se decompõem na mesma base, a matriz da Equação 7.2 também tem autovetores e_1 e e_2 , e seus autovalores saem multiplicando e dividindo os autovalores dos blocos,

$$\rho_1^2 = \frac{1.2 \times 1.2}{1.6 \times 1.6} = 0.5625, \quad \rho_2^2 = \frac{0.1 \times 0.1}{0.4 \times 0.4} = 0.0625$$

As correlações canônicas são 0.75 e 0.25. Os coeficientes são os autovetores reescalados para dar variância 1, e como as combinações definidas por e_1 e e_2 têm variâncias 1.6 e 0.4,

$$a_1 = \frac{e_1}{\sqrt{1.6}} = \begin{pmatrix} 0.559 \\ 0.559 \end{pmatrix}, \quad a_2 = \frac{e_2}{\sqrt{0.4}} = \begin{pmatrix} 1.118 \\ -1.118 \end{pmatrix}$$

Pela Equação 7.1, os coeficientes do segundo filho são os mesmos, como se espera da simetria entre os irmãos. O primeiro par soma as duas medidas e mede o tamanho da cabeça de cada filho. O segundo subtrai uma da outra e mede o formato, isto é, se a cabeça é mais alongada ou mais larga. O tamanho da cabeça dos irmãos é bem parecido, com correlação 0.75, e o formato, bem menos, com 0.25.

A primeira correlação dá para conferir sem calcular autovalor nenhum. A soma das duas medidas de cada filho tem variância $2(1 + 0.60) = 3.2$, e a covariância entre as somas dos dois irmãos é a soma dos quatro elementos do bloco cruzado, 2.4. A correlação é $2.4/3.2 = 0.75$.

7.2 Interpretação

Os coeficientes definem as variáveis canônicas, mas costumam ser ruins de ler. Eles dependem da escala das variáveis e, quando as variáveis de um bloco são correlacionadas entre si, dividem o efeito entre elas de forma instável, como os coeficientes de uma regressão com multicolinearidade. Um coeficiente pode até ter o sinal oposto ao da relação entre a variável e a variável canônica.

A leitura costuma ser feita pelas **cargas canônicas**, as correlações entre cada variável e a variável canônica do seu próprio bloco. Como a covariância entre o primeiro bloco e U_k é $\Sigma_{11}a_k$ e U_k tem variância 1, com variáveis padronizadas os vetores de cargas do k -ésimo par são

$$\ell_k^{(1)} = \rho_{11} \mathbf{a}_k, \quad \ell_k^{(2)} = \rho_{22} \mathbf{b}_k$$

As **cargas cruzadas** são as correlações de cada variável com a variável canônica do outro bloco, e não pedem conta nova. A primeira equação de Equação 7.1 diz que $\Sigma_{12} \mathbf{b}_k = \rho_k \Sigma_{11} \mathbf{a}_k$, e portanto a carga cruzada é a carga do próprio bloco multiplicada pela correlação canônica,

$$\text{Corr}(\mathbf{x}^{(1)}, V_k) = \rho_k \ell_k^{(1)}, \quad \text{Corr}(\mathbf{x}^{(2)}, U_k) = \rho_k \ell_k^{(2)}$$

Uma correlação canônica alta diz que as duas variáveis do par andam juntas, mas não diz quanto da variação das variáveis originais passa por elas. Se U_k resume mal o seu bloco, uma correlação alta com V_k interessa pouco. A **variância extraída** por U_k é a média das cargas ao quadrado e mede a fração da variância total do primeiro bloco, já padronizado, que U_k reproduz. O **índice de redundância** faz a mesma conta com as cargas cruzadas e mede a fração explicada pela variável canônica do outro bloco,

$$\text{VE}_k^{(1)} = \frac{1}{p} \sum_{j=1}^p \left(\ell_{jk}^{(1)} \right)^2, \quad \text{Red}_k^{(1)} = \frac{1}{p} \sum_{j=1}^p \left(\rho_k \ell_{jk}^{(1)} \right)^2 = \rho_k^2 \text{VE}_k^{(1)}$$

Para o segundo bloco, as fórmulas são as mesmas, com as cargas do segundo bloco e q no lugar de p . No bloco menor, as variáveis canônicas reconstituem o bloco inteiro, e as variâncias extraídas somam 1. As redundâncias, ao contrário das correlações canônicas, não são simétricas, e um bloco pode explicar boa parte do outro sem que o contrário aconteça.

Exemplo 7.2. Voltando aos irmãos do Exemplo 7.1, as cargas do primeiro filho nos dois pares são

$$\ell_1^{(1)} = \mathbf{R}_{11} \mathbf{a}_1 = \begin{pmatrix} 0.894 \\ 0.894 \end{pmatrix}, \quad \ell_2^{(1)} = \mathbf{R}_{11} \mathbf{a}_2 = \begin{pmatrix} 0.447 \\ -0.447 \end{pmatrix}$$

Comprimento e largura carregam igualmente no primeiro par, o que confirma a leitura de tamanho, e com sinais opostos no segundo, a leitura de formato. Multiplicando pelas correlações canônicas, as cargas cruzadas ficam em 0.671 e 0.112, em módulo.

As variâncias extraídas são $0.894^2 = 0.8$ no primeiro par e $0.447^2 = 0.2$ no segundo, e somam 1. As redundâncias são

$$\text{Red}_1^{(1)} = 0.75^2 \times 0.8 = 0.45, \quad \text{Red}_2^{(1)} = 0.25^2 \times 0.2 = 0.0125$$

O tamanho da cabeça do segundo filho explica 45% da variância das medidas do primeiro, e o formato quase nada. Como os dois blocos têm a mesma estrutura, os números do segundo filho são iguais.

7.3 Inferência

Na prática, trocamos Σ pela matriz de covariâncias amostral, ou pela matriz de correlações, e resolvemos a Equação 7.2 com as matrizes amostrais. As correlações canônicas amostrais tendem a exagerar a associação, porque a maximização também aproveita o ruído da amostra. O exagero cresce com o número de variáveis e diminui com o tamanho da amostra, e por isso vale testar quais correlações canônicas diferem de zero antes de interpretá-las.

Supondo normalidade multivariada (Capítulo 3), a independência entre os blocos equivale a todas as correlações canônicas serem nulas e pode ser testada pela razão de verossimilhanças. De forma mais geral, para saber se sobra associação depois dos k primeiros pares, testamos se as correlações canônicas restantes são nulas com o Λ de Wilks e a estatística corrigida de Bartlett,

$$\Lambda_k = \prod_{j=k+1}^p (1 - \hat{\rho}_j^2), \quad \chi^2 = - \left[n - 1 - \frac{p + q + 1}{2} \right] \ln \Lambda_k \quad (7.3)$$

que segue aproximadamente uma distribuição χ^2 com $(p - k)(q - k)$ graus de liberdade. O caso $k = 0$ é o teste global de independência. Os testes são feitos em sequência, e retemos os pares até o primeiro teste que não rejeita. Com amostras grandes, correlações pequenas demais para interessar acabam significativas, e o teste deve ser lido junto com as redundâncias.

Se as variáveis de um bloco forem quase colineares, a matriz de covariâncias desse bloco fica perto de singular, a inversa da raiz quadrada amplifica o ruído e os coeficientes variam muito de uma amostra para outra. O mesmo acontece quando o número de variáveis se aproxima do tamanho da amostra. As cargas sofrem bem menos com isso, mais um motivo para interpretá-las no lugar dos coeficientes.

Exemplo 7.3. Suponha que as correlações do Exemplo 7.1 tenham vindo de uma amostra de 25 famílias. O fator de correção da Equação 7.3 vale $25 - 1 - 5/2 = 21.5$. No teste global,

$$\Lambda_0 = (1 - 0.75^2)(1 - 0.25^2) = 0.410, \quad \chi^2 = -21.5 \ln 0.410 = 19.16$$

com 4 graus de liberdade e p-valor de 0.0007. Sem o primeiro par,

$$\Lambda_1 = 1 - 0.25^2 = 0.9375, \quad \chi^2 = -21.5 \ln 0.9375 = 1.39$$

com 1 grau de liberdade e p-valor de 0.24. Com 25 famílias, só a semelhança de tamanho entre irmãos se distingue do ruído.

7.4 Exemplo prático

Exemplo 7.4 (Bico e corpo dos pinguins). No Exemplo 4.6, a ACP tratou as quatro medidas dos pinguins como um bloco só. Agora vamos separá-las em dois blocos, as medidas do bico (comprimento e profundidade) e as do corpo (comprimento da nadadeira e massa corporal), e perguntar como o formato do bico acompanha o tamanho do corpo. Os dados e o tratamento dos valores ausentes são os mesmos da ACP.

Tabela 7.1: Matriz de correlações das medidas dos pinguins. As duas primeiras variáveis formam o bloco do bico, e as duas últimas, o bloco do corpo.

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

As correlações cruzadas, no canto superior direito da Tabela 7.1, já contam boa parte da história. O comprimento do bico cresce com o tamanho do corpo, e a profundidade diminui. Dentro do bloco do corpo, nadadeira e massa têm correlação 0.87, de modo que esse bloco é quase unidimensional.

Para calcular as correlações canônicas, fazemos a SVD da matriz $\mathbf{K}$ com a matriz de correlações e testamos os pares com a Equação 7.3.

Tabela 7.2: Correlações canônicas e testes sequenciais de Bartlett. A linha de cada par testa se ele e os seguintes têm correlação canônica nula.

	Correlação canônica	Λ de Wilks	Estatística	gl	p-valor
Par 1	0.791	0.37	336.355	4	1.5e-71
Par 2	0.100	0.99	3.409	1	0.065

A primeira correlação canônica vale 0.79 e a segunda, 0.10. O teste global rejeita com folga a independência entre os blocos, e o teste do segundo par tem p-valor 0.065, que não rejeita a 5%. Mesmo que rejeitasse, uma correlação de 0.10 descreve uma associação pequena demais para interessar, e a análise fica com um par só.

Tabela 7.3: Cargas canônicas (correlação com a variável canônica do próprio bloco) e cargas cruzadas (correlação com a do outro bloco) no primeiro par.

	Carga	Carga cruzada
Comp. do bico	0.828	0.656
Prof. do bico	-0.739	-0.585
Comp. da nadadeira	1.000	0.791
Massa corporal	0.865	0.684

No bloco do corpo, as duas cargas são altas e positivas, e V_1 é simplesmente o tamanho do pinguim. No bloco do bico, o comprimento carrega positivamente e a profundidade negativamente, e U_1 opõe bicos longos e rasos a bicos curtos e profundos. A correlação de 0.79 diz que pinguins maiores tendem a ter bicos longos e rasos.

Pelas redundâncias, o tamanho do corpo, V_1 , explica 39% da variância das medidas do bico, e o formato do bico, U_1 , explica 55% da variância das medidas do corpo. A assimetria vem da variância extraída, já que V_1 reproduz 87% do bloco do corpo, quase unidimensional, enquanto U_1 reproduz 62% do bloco do bico.

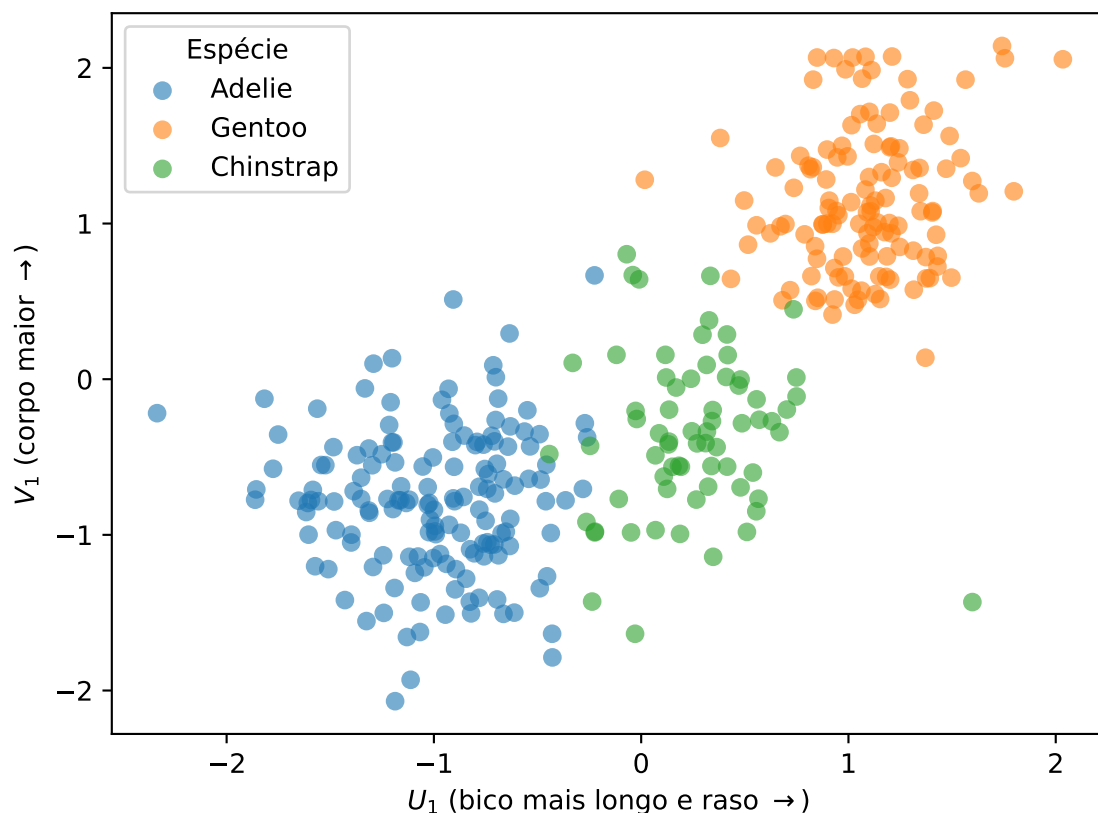


Figura 7.1: Escores do primeiro par de variáveis canônicas, com os pinguins coloridos por espécie.

A Figura 7.1 mostra de onde vem boa parte dessa correlação. Os Gentoo ocupam o canto superior direito, com corpo grande e bico longo e raso, e os Adelie ficam no canto oposto. Os Chinstrap, de bico longo e profundo, ficam no meio do eixo de U_1 , com o corpo pequeno, mais perto dos Adelie em V_1 . As espécies se separam ao longo da diagonal, e resta saber se a relação se mantém dentro de cada uma.

Tabela 7.4: Primeira correlação canônica e cargas do primeiro par, calculadas separadamente em cada espécie.

	Adelie	Chinstrap	Gentoo
Correlação canônica	0.68	0.67	0.83
Comp. do bico	0.82	0.81	0.87
Prof. do bico	0.85	0.97	0.94
Comp. da nadadeira	0.56	0.88	0.92
Massa corporal	0.99	0.93	0.93

Em todas as espécies a primeira correlação canônica continua alta, mas a profundidade do bico passa a carregar positivamente, junto com o comprimento. Dentro de uma espécie, pinguins maiores têm bicos maiores em todas as direções, como se espera de um animal que cresce por inteiro. O contraste da análise conjunta aparece porque os Gentoo, a espécie de corpo maior, têm os bicos mais rasos.

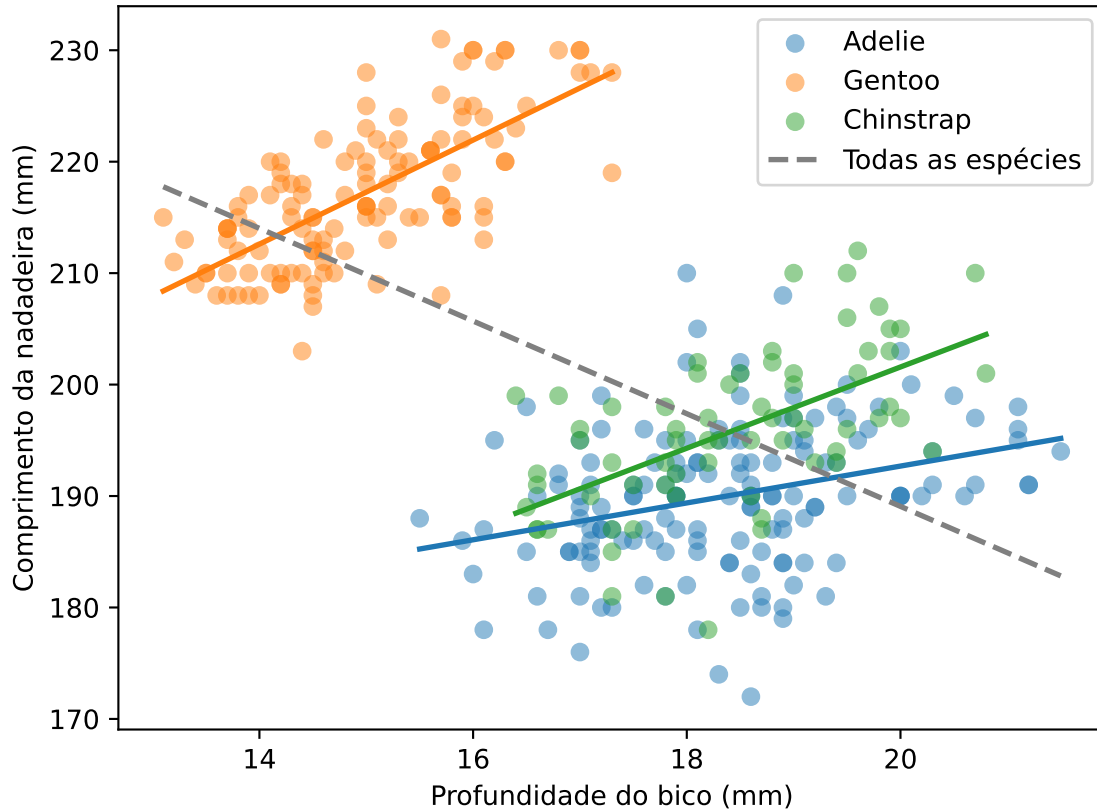


Figura 7.2: Profundidade do bico e comprimento da nadadeira. As retas coloridas são ajustadas dentro de cada espécie, e a tracejada, com todas as espécies juntas.

A Figura 7.2 mostra o mesmo efeito numa só dupla de variáveis. A correlação entre a profundidade do bico e o comprimento da nadadeira é negativa com as espécies juntas e positiva dentro de cada uma, um caso do paradoxo de Simpson. Sempre que a amostra mistura grupos, as variáveis canônicas podem descrever as diferenças entre os grupos, e não a relação dentro deles. Se a pergunta é biológica, sobre como o bico acompanha o corpo num mesmo animal, a análise por espécie responde melhor do que a conjunta.

7.5 Exercícios

Exercício 7.1. Mostre que, quando o primeiro bloco tem uma única variável X_1 , a única correlação canônica é a correlação múltipla entre X_1 e o segundo bloco, isto é,

$$\rho_1^2 = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}$$

que é o coeficiente de determinação da regressão linear de X_1 nas variáveis do segundo bloco. Por que a matriz K tem um único valor singular nesse caso?

Exercício 7.2. Sejam C_1 e C_2 matrizes inversíveis e considere os novos blocos $C_1\mathbf{x}^{(1)}$ e $C_2\mathbf{x}^{(2)}$.

- a) Mostre que as correlações canônicas não mudam. Toda combinação linear dos novos blocos é também uma combinação linear dos antigos.
- b) Determine os coeficientes nos novos blocos que reproduzem as variáveis canônicas originais.
- c) Conclua que fazer a ACC com a matriz de covariâncias ou com a matriz de correlações leva às mesmas correlações canônicas.

Exercício 7.3. Numa ACC com três variáveis padronizadas no primeiro bloco e quatro no segundo, a primeira correlação canônica é 0.9, e as cargas das variáveis do primeiro bloco em U_1 são 0.35, 0.20 e -0.25 .

- a) Calcule as cargas cruzadas, a variância extraída e o índice de redundância do primeiro bloco no primeiro par.
- b) Um colega conclui, pela correlação canônica, que o segundo bloco explica bem o primeiro. Você concorda?
- c) Como duas combinações lineares podem ter correlação tão alta e, ao mesmo tempo, dizer tão pouco sobre as variáveis originais?

Exercício 7.4 (Exercícios físicos e medidas corporais). O conjunto Linnerud, carregado no Python pela função `load_linnerud` do módulo `sklearn.datasets`, traz dados de 20 homens de meia-idade atendidos numa academia. Para cada um, foram registrados três exercícios, o número de barras (Chins), de abdominais (Situps) e de saltos (Jumps), e três medidas físicas, o peso em libras (weight), a cintura em polegadas (waist) e a pulsação em repouso (Pulse).

- a) Calcule a matriz de correlações e comente as correlações entre os dois blocos.
- b) Obtenha as correlações canônicas e use os testes sequenciais de Bartlett para decidir quantos pares reter.
- c) Interprete o primeiro par pelas cargas canônicas e calcule as redundâncias dos dois blocos.

d) Compare a magnitude da primeira correlação canônica com o resultado do teste global. O que o tamanho da amostra tem a ver com isso?

Leituras recomendadas

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