Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística
Bacharelado em Estatística

Análise Fatorial do Rendimento Acadêmico nas Disciplinas do Curso de Estatística da UFOP

Thales Tavares Correa

Ouro Preto-MG
Abril de 2025

Thales Tavares Correa

Análise Fatorial do Rendimento Acadêmico nas Disciplinas do Curso de Estatística da UFOP

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador

Prof. Dr. Eduardo Bearzoti

Universidade Federal de Ouro Preto – UFOP
Departamento de Estatística – DEEST

Ouro Preto-MG

02 de abril de 2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Thales Tavares Correa

Análise fatorial do rendimento acadêmico nas disciplinas do Curso de Estatística da UFOP

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 02 de abril de 2025

Membros da banca

Dr. Eduardo Bearzoti - Orientadora (Universidade Federal de Ouro Preto)
Dra. Carolina Silva Pena (Universidade Federal de Ouro Preto)
Dr. Helgem de Souza Ribeiro Martins (Universidade Federal de Ouro Preto)

Professor Dr. Eduardo Bearzoti, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 02/04/2025

Documento assinado eletronicamente por **Eduardo Bearzoti**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 04/04/2025, às 15:25, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carolina Silva Pena**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 04/04/2025, às 15:30, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Helgem de Souza Ribeiro Martins**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 04/04/2025, às 16:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0890417** e o código CRC **3A80CA5F**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

Agradecimentos

A realização deste trabalho e a conclusão da minha graduação não seriam possíveis sem o apoio e a dedicação de muitas pessoas que, de alguma forma, contribuíram para essa jornada.

Em primeiro lugar, expresso minha profunda gratidão ao meu orientador, Prof. Dr. Eduardo Bearzoti, por sua paciência, orientação e incentivo ao longo do desenvolvimento deste trabalho. Seu conhecimento e dedicação foram essenciais para o meu crescimento acadêmico e profissional.

Agradeço também a todos os professores do curso de Estatística da UFOP, que compartilharam seu conhecimento e paixão pela área, contribuindo para minha formação. Cada disciplina, desafio e aprendizado foram fundamentais para minha trajetória.

Sou imensamente grato pelas amizades construídas ao longo da graduação. A troca de experiências, o apoio mútuo e os momentos compartilhados tornaram essa caminhada mais leve e especial.

Por fim, um agradecimento especial à minha mãe, Fátima Tavares, que sempre me incentivou a estudar e acreditou no meu potencial. Seu amor, apoio incondicional e motivação foram fundamentais para que eu chegasse até aqui.

A todos que, de alguma forma, contribuíram para essa conquista, o meu muito obrigado!

Resumo

A analise fatorial consiste em uma técnica utilizada para encontrar estruturas subjacentes em conjuntos de variáveis observáveis, com objetivo de reduzir a dimensionalidade dos dados e identificar fatores que explicam as correlações entre as variáveis. Em relação à técnica de componentes principais, a análise fatorial possui algumas vantagens, como a possibilidade de os fatores serem rotacionados obliquamente, permitindo que correspondam a variáveis latentes correlacionadas. A técnica tem sido utilizada em dados educacionais, potencialmente podendo identificar eixos de aprendizado em um processo pedagógico. Este estudo realizou o ajuste de modelos fatoriais a um banco de dados contendo as notas de aprovação em disciplinas de todos os 118 alunos graduados no curso de Estatística da Universidade Federal de Ouro Preto (UFOP), até o segundo semestre letivo de 2023. Uma análise exploratória (descritiva) inicial revelou maiores notas de aprovação para disciplinas aplicadas e menores notas para disciplinas de Matemática. As disciplinas que apresentaram maior número de correlações significativas com outras disciplinas foram as Estatísticas, Inferências e Métodos Não Paramétricos. Na análise fatorial propriamente dita, inicialmente se verificou ser o número de variáveis excessivo, em relação ao tamanho da amostra. Foram retiradas 5 disciplinas com base na estatística KMO, e em seguida 03 tentativas de ajuste foram consideradas, correspondendo a diferentes modelos fatoriais (5 fatores e 20 variáveis, 5 fatores e 16 variáveis, e 4 fatores e 16 variáveis). A retirada de mais 4 variáveis se deveu aos baixos valores de comunalidade. O modelo com 4 fatores se revelou apropriado em termos de sua interpretação, identificando quatro possíveis eixos de aprendizado, aqui caracterizados como: teórico-aplicado, aplicado, computacional-aplicado, e finalmente um eixo teórico. Embora a presente base de dados não tenha apresentado forte associação entre suas variáveis, a análise fatorial permitiu uma compreensão mais estruturada do aprendizado no curso, ilustrando o potencial da técnica, mesmo em cenários com baixa correlação entre variáveis.

Palavras-chave: análise fatorial, eixos de aprendizado, Estatística Educacional.

Abstract

Factor analysis is a technique used to identify underlying structures in sets of observable variables, aiming to reduce data dimensionality and identify factors that explain the correlations among variables. Compared to principal component analysis, factor analysis has some advantages, such as the possibility of oblique factor rotation, allowing them to correspond to correlated latent variables. This technique has been applied in educational data, potentially identifying learning axes in a pedagogical process. This study fitted factorial models to a database containing the passing grades of all 118 students who graduated from the Statistics program at the Federal University of Ouro Preto (UFOP) up to the second academic semester of 2023. An initial exploratory (descriptive) analysis revealed higher passing grades for applied subjects and lower grades for Mathematics-related subjects. The disciplines that showed the highest number of significant correlations with other subjects were Statistics, Inferences, and Nonparametric Methods. In the factor analysis, it was initially observed that the number of variables was excessive relative to the sample size. Five disciplines were removed based on the KMO statistic, followed by three adjustment attempts, corresponding to different factorial models (5 factors and 20 variables, 5 factors and 16 variables, and 4 factors and 16 variables). The removal of four additional variables was due to low communality values. The model with four factors proved to be the most interpretable, identifying four possible learning axes, characterized here as: theoretical-applied, applied, computational-applied, and a theoretical axis. Although the present dataset did not show strong associations among its variables, factor analysis provided a more structured understanding of learning in the program, illustrating the technique's potential even in scenarios with low variable correlations.

Keywords: Factor analysis, learning axes, Educational Statistics.

Lista de figuras

1 Diagramas de caixa das notas de aprovação de três grupos de disciplinas: grupo 1 (Metodologia Científica), grupo 4 (Estatísticas) e grupo 8 (Laboratório e Monografia). . . . . p. 37

2 Diagramas de caixa das notas de aprovação de dois grupos de disciplinas: grupo 2 (Matemática) e grupo 3 (área computacional). . . . . p. 38

3 Diagramas de caixa das notas de aprovação de dois grupos de disciplinas: grupo 5 (Probabilidades e Inferências) e grupo 6 (Multivariada, Regressão e Processos Estocásticos). . . . . p. 39

4 Diagramas de caixa das notas de aprovação do grupo 7 de disciplinas, de natureza aplicada em áreas específicas (ver codificação das disciplinas no texto). . . . . p. 40

5 Representação gráfica da matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP. Ângulos dos setores circulares são proporcionais aos valores de correlação; setores em azul representam correlações positivas, e setores em vermelho representam correlações negativas. . . . . p. 42

6 Representação gráfica da matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP, contendo apenas correlações significativas a 1% de probabilidade. Os raios das circunferências são proporcionais às magnitudes das correlações; circunferências azuis correspondem a correlações positivas, e a circunferência vermelha representa uma correlação negativa. . . . . p. 43

7 Histograma da distribuição de frequência das 300 correlações de Pearson, envolvendo todas as 25 disciplinas do banco de dados utilizado. . . . . p. 46

8 Gráfico scree apresentando os autovalores da matriz de correlações original, os quais estão associados a uma análise de componentes principais ("PC"), bem como os autovalores de uma análise fatorial ("FA"). . . . . p. 49

Lista de tabelas

1 Magnitudes de cargas fatoriais dignas de atenção, conforme o tamanho amostral (fonte: HAIR et al., 2009). . . . . p. 21

2 Adequação da base de dados, conforme a magnitude das medidas de adequação KMO (fonte: FIELD *et al.*, 2012; citados por MATOS e RODRIGUES, 2019). . . . . p. 27

3 Número de bacharéis formados pelo curso de Estatística da UFOP, conforme o ano e o semestre de colação de grau. . . . . p. 31

4 Abreviaturas utilizadas neste trabalho, para referenciar as disciplinas do curso de Estatística da UFOP. . . . . p. 33

5 Medidas descritivas referentes às notas de aprovação das disciplinas do curso de Estatística da UFOP, reunidas em 8 grupos, conforme suas características. . . . . p. 36

6 Matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP. . . . . p. 41

7 Medidas KMO referentes a cada uma das $n = 25$ variáveis, bem como medida KMO global.[1] . . . . . p. 47

8 Medidas KMO referentes ao novo banco de dados, contendo $n = 20$ variáveis.[1] . . . . . p. 48

9 Resultados do ajuste de modelos fatoriais considerando $p = 20$ variáveis e 5 fatores.[1] . . . . . p. 50

10 Cargas fatoriais, comunalidade (h^2) e especificidade (ψ) de cada variável, no ajustamento de um modelo fatorial com $p = 20$ variáveis e 5 fatores F1, F2, ... F5, utilizando máxima verossimilhança e rotação oblíqua.[1] . p. 53

11 Cargas fatoriais, comunalidade (h^2) e especificidade (ψ) de cada variável, no ajustamento de um modelo fatorial com $p = 16$ variáveis e 5 fatores F1, F2, ... F5, utilizando máxima verossimilhança e rotação oblíqua.[1] . p. 55

12 Resultados do ajuste de um modelo fatorial considerando $p = 16$ variáveis e 4 fatores. . . . . p. 55

13 Cargas fatoriais, comunalidade (h^2) e especificidade (ψ) de cada variável, no ajustamento de um modelo fatorial com $p = 16$ variáveis e 4 fatores F1, F2, ... F4, utilizando máxima verossimilhança e rotação oblíqua.[1] . p. 56

14 Coeficientes de correlação de Pearson entre os escores dos fatores de uma análise fatorial com 16 variáveis e 4 fatores. . . . . p. 59

Sumário

1 Introdução p. 10

2 Referencial Teórico p. 12

2.1 Análise Fatorial . . . . . p. 12

2.1.1 Análise de Componentes Principais . . . . . p. 13

2.1.2 O Modelo Fatorial . . . . . p. 17

2.1.3 Estimação . . . . . p. 22

Método dos Componentes Principais . . . . . p. 22

Método dos Fatores Principais . . . . . p. 22

Método da Máxima Verossimilhança . . . . . p. 23

2.1.4 Rotação de Fatores . . . . . p. 23

2.1.5 Estimação de Escores . . . . . p. 24

2.1.6 Etapas de Uma Análise Fatorial . . . . . p. 26

2.2 O Curso de Estatística da UFOP . . . . . p. 29

3 Metodologia p. 32

4 Resultados e Discussão p. 35

4.1 Estatística Descritiva . . . . . p. 35

4.1.1 Correlações . . . . . p. 40

4.2 Análise Fatorial . . . . . p. 45

Adequação do Tamanho da Amostra . . . . . p. 45

Matriz de Correlações e Teste de Bartlet . . . . . p. 45

Medida de Kaiser-Meyer-Olkin (KMO) . . . . . p. 47

Determinação do Número de Fatores . . . . . p. 48

Ajuste de Modelos Fatoriais . . . . . p. 49

5 Considerações Finais p. 60

6 Referências Bibliográficas p. 62

1 Introdução

A análise fatorial é uma técnica multivariada que objetiva uma redução da dimensionalidade de bancos de dados apresentando grande número de variáveis. Esta é uma situação relativamente comum em Ciências Humanas, por exemplo para a análise de questionários, em que as respostas de diferentes questões corresponderiam a diferentes variáveis.

A análise fatorial é uma técnica semelhante à de componentes principais, no sentido de que a informação contida nas variáveis é resumida em um determinado número de *dimensões* (que aqui são denominadas fatores), obtidas a partir da matriz de correlações entre as variáveis. Se estas apresentam associação entre si, será em geral possível reduzir a dimensionalidade, utilizando um número de fatores consideravelmente menor que o número de variáveis.

A análise fatorial pode ser utilizada não apenas para variáveis quantitativas, mas também qualitativas, pelo uso de coeficientes de correlação apropriados, como a correlação bisserial e a correlação policórica. Em relação à análise de componentes principais, a análise fatorial apresenta como vantagem a possibilidade de que os fatores estimados sejam associados entre si. Enquanto que a técnica de componentes principais utiliza dimensões (componentes) ortogonais entre si, na análise fatorial é possível trabalhar com dimensões rotacionadas de maneira oblíqua, e assim os componentes (fatores) gerados potencialmente podem estar correlacionados. Esta é uma possibilidade interessante, uma vez que nem sempre é razoável admitir que as dimensões reflitam aspectos independentes (devido à ortogonalidade), no fenômeno sendo estudado.

A técnica tem sido utilizada em dados educacionais (MATOS e RODRIGUES, 2019), seja pela análise de questionários, como também de indicadores de aprendizado. A análise de desempenho acadêmico dos alunos é uma ferramenta fundamental para instituições de ensino superior. Compreender como os alunos se saem em diferentes disciplinas pode fornecer inferências valiosas para aprimorar currículos e métodos de ensino. Neste sentido, a análise fatorial poderia auxiliar na identificação de eixos fundamentais de aprendizado,

dentro de um processo pedagógico. É razoável admitir que tais eixos, ou dimensões, se apresentem como associados. Por exemplo, um eixo relativo a habilidades matemáticas possivelmente estaria associado a outros eixos, como o de habilidades computacionais. Assim, a análise fatorial tem a potencialidade de uma maior identificação destes "eixos (dimensões) de aprendizado", pois leva em conta a possibilidade de serem eixos correlacionados.

Por exemplo, no processo de formação de bacharéis em Estatística, é natural admitir que existam determinados eixos de aprendizado, refletindo diferentes habilidades, e sua identificação permitiria um maior conhecimento sobre como se dá o processo pedagógico de aprendizado. Sendo assim, este trabalho teve como objetivo principal realizar uma análise fatorial dos dados de notas de aprovação das diferentes disciplinas cursadas pelos alunos formados em Estatística na Universidade Federal de Ouro Preto (UFOP). Concomitantemente, também são utilizadas técnicas de estatística descritiva destas mesmas notas, contribuindo para a caracterização de seu comportamento.

2 Referencial Teórico

2.1 Análise Fatorial

A análise fatorial é uma técnica estatística apresentada e discutida em diversos textos de análise multivariada. Neste trabalho, a bibliografia fundamental para o referenciamento teórico da análise fatorial consistiu das seguintes obras: MATOS e RODRIGUES (2019), HAIR et al. (2009), JOHNSON e WICHERN (2007), e de HÄRDLE e SIMAR (2019). As duas primeiras abordam de maneira didática e prática como a técnica deve ser aplicada e interpretada, enquanto que as outras duas bibliografias apresentam uma representação matemática mais detalhada. A seguir serão apresentados alguns dos principais aspectos teóricos da análise fatorial, mas, para um maior aprofundamento, recomenda-se a leitura das bibliografias supracitadas.

A análise fatorial é uma técnica estatística utilizada para identificar a estrutura subjacente de um conjunto de variáveis observáveis, com o objetivo de reduzir a dimensionalidade dos dados e identificar fatores latentes que expliquem as correlações entre as variáveis. Este método é particularmente útil em áreas como as Ciências Humanas e Sociais, onde muitas vezes é necessário mensurar fenômenos que não são diretamente observáveis, conhecidos como variáveis latentes ou construtos. Variáveis latentes são aquelas que não podem ser medidas diretamente, como inteligência ou motivação. Variáveis observáveis são medidas diretamente e utilizadas para inferir acerca das variáveis latentes, por meio de um modelo matemático. Por exemplo, o nível socioeconômico pode ser inferido a partir de variáveis como ocupação e escolaridade dos pais, bens domésticos, entre outros (MATOS e RODRIGUES, 2019).

A análise fatorial tem assim como objetivo a redução da dimensionalidade dos dados, identificando um número menor (em relação ao número de variáveis) de fatores que explicam a maior parte da variabilidade das variáveis observáveis, identificar a estrutura subjacente das correlações entre estas, e ajudar na interpretação dos dados, ao agrupar variáveis correlacionadas em fatores.

Uma técnica multivariada que também busca reduzir a dimensionalidade dos dados é a análise de componentes principais. Por se tratar de uma técnica que por vezes é utilizada em associação com a análise fatorial (por exemplo para auxiliar na determinação no número de fatores), é brevemente apresentada a seguir.

2.1.1 Análise de Componentes Principais

A análise de componentes principais é semelhante à análise fatorial, no sentido de que, em ambas, as variáveis originais são vistas como combinações lineares de componentes (ou fatores), com o objetivo de reduzir a dimensionalidade dos dados.

O ponto de partida da análise de componentes principais é um conjunto de dados multivariado, referente a p diferentes variáveis, avaliadas em n unidades amostrais. Neste conjunto, represente-se como x_{ij} o valor da i-ésima variável observado na j-ésima unidade amostral.

De acordo com esta técnica, são definidos um número de componentes principais igual ao número p de variáveis no conjunto, sendo cada componente uma combinação linear das variáveis originais. Por exemplo, o primeiro dos componentes principais, para a j-ésima unidade amostral, poderia ser definido como:

$$g_{1j} = c_{11}x_{1j} + c_{12}x_{2j} + \ldots + c_{1p}x_{pj}, \quad j = 1, 2, \ldots n \tag{2.1}$$

Considerando todas as n unidades amostrais, os valores definidos em 2.1 podem ser dispostos em um vetor $\mathbf{g}_1'$, referente ao primeiro componente. Aqui, $'$ indica transposição (estamos considerando um vetor linha[1], de dimensão $1 \times n$). Os componentes principais são definidos de maneira que seus p vetores $\mathbf{g}_i'$ sejam ortogonais entre si, com o objetivo de captar diferentes aspectos da variação contida nos dados.

É frequente que as variáveis X_i apresentem diferentes escalas, ou unidades. Assim, é conveniente, para uma melhor interpretação dos componentes, que as variáveis estejam padronizadas. Se o valor original de cada variável é representado por X_{ij}, então os valores padronizados são dados por:

$$x_{ij} = \frac{X_{ij} - \bar{X}_i}{\sqrt{\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2}} \tag{2.2}$$

sendo $\bar{X}_i$ a média da variável i. Estes valores padronizados x_{ij} são os que estão sendo con-

[1] neste texto, está-se trabalhando com um padrão de p linhas e n colunas.

siderados em 2.1. A padronização 2.2 faz com que a soma dos quadrados das observações referentes a cada variável seja igual a 1. Ou seja, faz com que cada vetor $\mathbf{x}_i'$ $(i = 1, 2, \ldots p)$ tenha norma igual a 1.

O conjunto dos dados padronizados pode então ser expresso por uma matriz, contendo as linhas $\mathbf{x}_i'$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \tag{2.3}$$

A partir da matriz $\mathbf{X}$, verifica-se que a matriz de correlações entre as p variáveis X_i é facilmente obtida por:

$$\mathbf{R} = \mathbf{X}\mathbf{X}' \tag{2.4}$$

Por sua vez, o vetor linha $\mathbf{g}_1'$ pode ser obtido por:

$$\mathbf{g}_1' = \mathbf{c}_1'\mathbf{X} \tag{2.5}$$

em que:

$$\mathbf{c}_1' = [c_{11}\ c_{12} \ldots c_{1p}] \tag{2.6}$$

contém os coeficientes do primeiro componente principal, apresentados em 2.1. O vetor $\mathbf{c}_1$ (assim como os vetores dos demais componentes principais) é definido de maneira a ter norma igual a 1, ou seja, de maneira que $\mathbf{c}_1'\mathbf{c}_1 = 1$.

Pela definição em 2.5, a variância amostral entre os elementos de $\mathbf{g}_1$ é dada por:

$$V(g_{1j}) = \frac{1}{n-1}\mathbf{c}_1'\mathbf{X}\mathbf{X}'\mathbf{c}_1 = \frac{1}{n-1}\mathbf{c}_1'\mathbf{R}\mathbf{c}_1$$

A obtenção do primeiro componente principal consiste em obter o vetor $\mathbf{c}_1$, de maneira a maximizar a variância $V(g_{1j})$, o que equivale a maximizar $\mathbf{c}_1'\mathbf{R}\mathbf{c}_1$, e sujeito à restrição $\mathbf{c}_1'\mathbf{c}_1 = 1$. Esta restrição pode ser imposta através de um multiplicador de Lagrange λ.

Assim, pode-se demonstrar que essa maximização, sujeita à restrição $\mathbf{c}_1'\mathbf{c}_1 = 1$, é obtida pelas soluções do seguinte sistema homogêneo:

$$(\mathbf{R} - \lambda\mathbf{I})\mathbf{c}_1 = \mathbf{0}$$

sendo $\mathbf{I}$ a matriz identidade $p \times p$, e $\mathbf{0}$ um vetor de 0's. Desconsiderando a solução trivial

$\mathbf{c}_1 = \mathbf{0}$, tem-se que as demais soluções são obtidas pela chamada *equação característica*:

$$\det(\mathbf{R} - \lambda\mathbf{I}) = 0$$

As soluções λ e $\mathbf{c}_1$ são denominadas, respectivamente, de *autovalores* e *autovetores* de $\mathbf{R}$. Haverá p autovalores correspondentes a p soluções distintas (ou, mais rigorosamente, p grupos distintos de soluções), sendo que, para cada autovalor, haverá um correspondente autovetor. Uma vez que a matriz $\mathbf{R}$ é em geral positiva definida, os autovalores serão positivos, podendo haver alguns iguais a zero, caso $\mathbf{R}$ seja positiva semi-definida. Assim, podemos ordenar o conjunto dos autovalores, referenciando-os como $\lambda_1 > \lambda_2 > \ldots > \lambda_p$.

Neste processo de maximização, também pode-se demonstrar que a variância do primeiro componente fica dada por:

$$V(g_{1j}) = \frac{\lambda}{n-1}$$

de maneira que, para maximizar $V(g_{1j})$ devemos tomar $\lambda = \lambda_1$, tomando o seu respectivo autovetor como sendo o vetor $\mathbf{c}_1$.

Uma vez obtido o primeiro componente principal, o segundo componente é obtido de maneira semelhante, maximizando a sua variância (desconsiderando a solução anterior, do primeiro componente), com as restrições $\mathbf{c}_2'\mathbf{c}_2 = 1$ e $\mathbf{c}_2'\mathbf{c}_1 = 0$ (esta última, referente à ortogonalidade com o componente 1). Assim, este processo também resulta na equação característica, de maneira que a solução que maximiza a variância $V(g_{2j})$ (descontando a solução do primeiro componente) consiste em tomar $\lambda = \lambda_2$, tomando o seu respectivo autovetor como sendo $\mathbf{c}_2$.

Os demais componentes são obtidos de maneira semelhante, conforme a ordem dos autovalores, de maneira que a sequência de componentes irá captando frações decrescentes de variância. Idealmente, um número reduzido de componentes irá captar uma proporção considerável da variância, promovendo assim uma redução de dimensionalidade.

Considerando todo o conjunto de p componentes principais, estes podem ser agrupados em uma matriz:

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \vdots & \vdots & & \vdots \\ g_{p1} & g_{p2} & \cdots & g_{pn} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1' \\ \mathbf{g}_2' \\ \vdots \\ \mathbf{g}_p' \end{bmatrix}$$

Da mesma maneira, os vetores contendo os coeficientes de todos os componentes principais

podem ser reunidos em uma matriz:

$$\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \dots \mathbf{c}_p] \tag{2.7}$$

Esta matriz 2.7 tem a propriedade de ser *ortogonal*, ou seja, sua inversa é igual à sua transposta.

Pela própria definição dos componentes, decorre que:

$$\mathbf{G} = \mathbf{C}'\mathbf{X} \tag{2.8}$$

Assim, em virtude da ortogonalidade de $\mathbf{C}$, decorre que:

$$\mathbf{X} = \mathbf{C}\mathbf{G} \tag{2.9}$$

ressaltando o aspecto interessante de que não só os componentes são combinações lineares das variáveis, mas também as variáveis são combinações lineares dos componentes, e de maneira que ambas estas combinações utilizam coeficientes que são elementos da mesma matriz $\mathbf{C}$.

Os componentes assim definidos em 2.8, conforme visto, possuem variância $\frac{\lambda_i}{n-1}$. É também possível defini-los de maneira que tenham a mesma variância, padronizando a sua escala, utilizando a seguinte matriz alternativa:

$$\mathbf{G}^* = \boldsymbol{\lambda}^{-1/2}\mathbf{G}$$

sendo $\boldsymbol{\lambda}$ uma matriz diagonal contendo os autovalores $\lambda_1, \ \lambda_2 \dots \lambda_p$. Assim, todos os componentes terão a mesma variância, igual a $\frac{1}{n-1}$.

Também é possível, e talvez vantajoso, definir os componentes de maneira que tenham todos variância igual a 1 (variância unitária). Uma vantagem dessa definição é a de que seria possível identificar unidades amostrais com valores relativamente altos, ou relativamente baixos, para um ou mais componentes (por exemplo, acima de 3 ou abaixo de -3). Para tanto, basta utilizar a matriz alternativa:

$$\mathbf{G}^{**} = \sqrt{n-1}\ \boldsymbol{\lambda}^{-1/2}\mathbf{G}$$

De qualquer forma, seja utilizando $\mathbf{G}$, $\mathbf{G}^*$ ou $\mathbf{G}^{**}$, tem-se, em cada linha dessas matrizes, os valores de cada componente, para cada unidade amostral. Apenas a título de esclarecimento, reforça-se que um dado componente principal i consiste de um vetor $\mathbf{g}_i$, e não do seu vetor $\mathbf{c}_i$, sendo esta uma confusão comum.

2.1.2 O Modelo Fatorial

Assim como ocorre na análise de componentes principais, na análise fatorial são definidos *fatores*, de tal maneira que as variáveis originais consistam de combinações lineares suas (de maneira semelhante a 2.9). Os fatores, conforme já apontado, representariam variáveis latentes (não-observáveis), que ajudariam a descrever o fenômeno estudado, consistindo de dimensões de interpretação aplicada, como dimensões de aprendizagem, em uma avaliação multivariada de aprendizado.

A grande diferença em relação à análise de componentes principais é a de que na análise fatorial assume-se um modelo (chamado modelo fatorial), estabelecendo que a maior parte das correlações entre as variáveis X_i possam ser explicadas por um número limitado de fatores. Enquanto que na análise de componentes principais são definidos p componentes, na análise fatorial são definidos k fatores (denominados *fatores comuns*), sendo $k < p$. Idealmente, k seria bem menor que p, resultando em uma diminuição na dimensionalidade.

No modelo fatorial, além dos k fatores comuns, é proposto também, para cada variável, um *fator específico*, representando aspectos particulares de cada variável, não associados com as demais variáveis. Desta forma, o modelo fatorial é dado por:

$$x_{ij} = l_{i1}f_{1j} + l_{i2}f_{2j} + \ldots + l_{ik}f_{kj} + \epsilon_{ij}, \quad i = 1, 2, \ldots p, \text{ e } j = 1, 2, \ldots n \tag{2.10}$$

ou

$$x_{ij} = \sum_{m=1}^{k} l_{im}f_{mj} + \epsilon_{ij}$$

Aqui, f_{mj} é o valor do m-ésimo fator comum para a j-ésima unidade amostral, l_{im} são coeficientes a serem estimados, e ϵ_{ij} corresponde ao valor do fator específico i para a unidade amostral j.

Conforme a notação ϵ_{ij} sugere, os fatores específicos poderiam ser entendidos como resíduos do modelo fatorial. Além disso, os coeficientes l_{im} são denominadas *cargas fatoriais*. Ou seja, l_{im} é interpretado como a *carga* da i-ésima variável sobre o m-ésimo fator.

Diga-se de passagem, aqui cabe uma observação, em relação à técnica de componentes principais. Pode-se dizer que as cargas fatoriais guardam uma certa correspondência com os coeficientes apresentados na matriz C, em 2.9. Ou seja, algebricamente, a relação

matricial 2.9 poderia ser escrita como:

$$x_{ij} = c_{1i}g_{1j} + c_{2i}g_{2j} + \ldots + c_{pi}g_{pj}, \quad i = 1, 2, \ldots p; \quad j = 1, 2, \ldots n \tag{2.11}$$

Assim, há uma correspondência entre esta relação e 2.10, no sentido que, em ambas, x_{ij} é expressa como uma combinação linear de componentes (ou fatores). Esta semelhança faz com que, na análise de componentes principais, os coeficientes presentes na matriz C também sejam frequentemente chamados de *cargas*. Por exemplo, na função *princomp()* da linguagem **R** (R CORE TEAM, 2024), tais coeficientes são chamados de *loadings* (cargas).

Rigorosamente falando, porém, "carga" corresponde a um termo emprestado da análise fatorial. Por exemplo, JOHNSON e WICHERN (2007), no capítulo referente à análise de componentes principais, evitam utilizar o termo *carga*, para se referir aos elementos de C.

De qualquer forma, se emprestamos o termo "carga" na análise de componentes principais, poderíamos dizer (conforme 2.11) que um coeficiente c_{im} $(i, m = 1, 2, \ldots p)$ é a carga da i-ésima variável sobre o m-ésimo componente. Ou seja, são coeficientes que expressam as variáveis como combinações lineares dos componentes. E, conforme apontado anteriormente (como visto em 2.1), estes mesmos coeficientes são utilizados para expressar cada componente como uma combinação linear das variáveis. Porém, isto é uma exclusividade da análise de componentes principais, devido à ortogonalidade da matriz C.

Na análise fatorial, as cargas representam coeficientes que expressam cada variável como uma combinação linear dos fatores, mas a recíproca não é necessariamente verdadeira. Na análise fatorial, em geral *não* podemos utilizar as cargas fatoriais para expressar cada fator como uma combinação linear das variáveis, sendo esta uma confusão comum. Isto poderia acontecer com a utilização do método de estimação dos componentes principais (ver adiante), mas este é considerado um método simples e com restrições, sendo mais utilizados outros métodos, como o dos fatores principais ou da máxima verossimilhança (ver adiante).

Uma outra grande diferença entre as duas abordagens é a de que, enquanto que a análise de componentes principais pode ser vista como um método puramente geométrico, na análise fatorial assume-se explicitamente que as variáveis observadas sejam variáveis aleatórias. Além disso, admite-se ainda que tanto os fatores comuns como os fatores específicos também sejam variáveis aleatórias, embora não diretamente observáveis.

Desta maneira, pode-se definir um vetor aleatório $\mathbf{x}$, de dimensões $p \times 1$, correspon-

dentes às p variáveis. Ou seja, cada coluna da matriz $\mathbf{X}$ definida em 2.3, que se refere a cada unidade amostral, pode ser entendida como uma manifestação (ou realização) deste vetor aleatório $\mathbf{x}$. Da mesma maneira como considerado anteriormente, está-se admitindo que o vetor aleatório $\mathbf{x}$ seja um vetor padronizado, ou que pelo menos esteja centralizado em relação à média, de maneira que $E(\mathbf{x}) = \mathbf{0}$. Assim, matricialmente, o modelo 2.10 pode ser expresso como:

$$\mathbf{x} = \mathbf{L}\mathbf{f} + \boldsymbol{\epsilon} \tag{2.12}$$

sendo $\mathbf{L}$ a matriz $p \times k$ de cargas fatoriais, $\mathbf{f}$ o vetor aleatório $k \times 1$ de fatores comuns, e $\boldsymbol{\epsilon}$ o vetor $p \times 1$ de fatores específicos. Em relação aos fatores específicos, admite-se que:

$$E(\boldsymbol{\epsilon}) = \mathbf{0}$$

$$\mathrm{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

O vetor de fatores comuns $\mathbf{f}$ é admitido como tendo esperança igual a zero ($E(\mathbf{f}) = \mathbf{0}$) e variâncias unitárias. Porém, em relação às covariâncias entre os fatores comuns, existem duas variações quanto ao modelo fatorial. No chamado *modelo fatorial ortogonal*, admite-se que os fatores sejam ortogonais entre si, de maneira que:

$$\mathrm{Cov}(\mathbf{f}) = E(\mathbf{f}\mathbf{f}') = \mathbf{I}$$

Já no chamado *modelo fatorial oblíquo*, admite-se a possibilidade de que esta matriz não seja diagonal, ou seja, que os fatores comuns apresentem associações entre si. Isto corresponde a uma grande flexibilização do modelo fatorial, de grande potencial aplicado, pois em muitas situações é razoável admitir que as variáveis latentes estejam correlacionadas. Por exemplo, se as variáveis latentes correspondem a dimensões de aprendizado em um curso de Estatística, seria razoável admitir que uma dimensão referente (digamos) a habilidades matemáticas esteja correlacionada a outra dimensão referente a habilidades computacionais. Fatores ortogonais podem ser convertidos em fatores correlacionados através de um procedimento de rotação oblíqua.

Finalmente, admite-se ainda que os fatores comuns sejam independentes dos fatores específicos, de maneira que:

$$\mathrm{Cov}(\boldsymbol{\epsilon}, \mathbf{f}) = E(\boldsymbol{\epsilon}\mathbf{f}') = \mathbf{0}$$

sendo $\mathbf{0}$, aqui, uma matriz $p \times k$ de zeros.

Em posse dessas pressuposições, pode-se obter a matriz de covariâncias de $\mathbf{x}$, bem como a matriz de covariâncias entre $\mathbf{x}$ e $\mathbf{f}$:

$$\mathrm{Cov}(\mathbf{x}) = E(\mathbf{x}\mathbf{x}') = \mathbf{L}E(\mathbf{f}\mathbf{f}')\mathbf{L}' + \mathbf{\Psi} \tag{2.13}$$

$$\mathrm{Cov}(\mathbf{x}, \mathbf{f}) = E(\mathbf{x}\mathbf{f}') = \mathbf{L}E(\mathbf{f}\mathbf{f}') \tag{2.14}$$

Em particular, no modelo fatorial ortogonal tem-se $E(\mathbf{f}\mathbf{f}') = \mathbf{I}$, e as relações 2.13 e 2.14 simplificam para:

$$\mathrm{Cov}(\mathbf{x}) = \mathbf{L}\mathbf{L}' + \mathbf{\Psi} \tag{2.15}$$

$$\mathrm{Cov}(\mathbf{x}, \mathbf{f}) = \mathbf{L} \tag{2.16}$$

A relação 2.16 é muito interessante, mostrando que cada carga fatorial l_{im} corresponde à covariância entre a variável i e o fator m. Caso o vetor $\mathbf{x}$, além de centralizado, esteja também padronizado (o que geralmente é o caso), isto facilitará muito a interpretação de cada carga fatorial, correspondendo à correlação entre a variável i e o fator m. Por outro lado, caso esteja-se considerando o modelo fatorial oblíquo, a interpretação das cargas fatoriais poderia ser feita de maneira semelhante à interpretação de coeficientes de uma regressão múltipla. Consequentemente, aqui, pode acontecer de as cargas fatoriais serem maiores que 1 ou menores que -1.

De qualquer maneira, se uma determinada carga fatorial tiver valor positivo, isto indica que a variável e o fator latente estão positivamente correlacionados. Valores negativos, por outro lado, indicam que a variável tende a variar no sentido oposto ao do fator.

Como exemplo, considere-se um conjunto de dados, referente ao conjunto de notas de aprovação em disciplinas de um curso de Estatística. Caso haja um fator que representa uma dimensão latente para "habilidades matemáticas", isto pode se manifestar através de cargas fatoriais positivas e elevadas para alguma (ou todas) as variáveis: "nota em Cálculo", "nota em Geometria Analítica" e/ou "nota em Álgebra Linear".

Na realidade, na prática seguimos o caminho contrário: a verificação de quais variáveis apresentam cargas elevadas para um dado fator é o que nos ajuda a identificar a natureza, ou a que se refere, o fator em questão.

As cargas fatoriais também permitem calcular a proporção da variância de uma variável que é explicada por cada fator comum, correspondendo à carga fatorial ao quadrado. No modelo fatorial ortogonal, pode-se demonstrar que a variância de uma variável observada X_i é dada por:

$$\mathrm{V}(X_i) = l_{i1}^2 + l_{i2}^2 + \ldots + l_{ik}^2 + \psi_i \tag{2.17}$$

A soma

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{ik}^2 \tag{2.18}$$

é denominada *comunalidade*, enquanto que o termo ψ_i é denominado *variância específica*. No processo de estimação, eventualmente acontece de a variância específica apresentar uma estimativa negativa. Isto pode ser um indicativo de que a variável em questão deva ser retirada, ou de que um número excessivo de fatores comuns esteja sendo utilizado.

No modelo fatorial oblíquo, a comunalidade não se limitará a uma soma de quadrados, pois envolverá covariâncias entre os fatores comuns. Apesar deste aspecto, é interessante apontar que o valor total da comunalidade h_i^2 de uma variável i não se altera com a utilização de rotação, seja esta ortogonal ou oblíqua.

Embora seja esperado que as variáveis apresentem cargas em todos os fatores, buscando um padrão para a associação entre elas, idealmente cada variável teria uma carga fatorial alta no fator (ou dimensão latente) ao qual pertence, e carga baixa nos demais fatores. Caso isso não ocorra, mesmo rotacionando os fatores (ver adiante), isto pode ser um indicativo de má especificação do modelo fatorial, por exemplo quanto ao número de fatores utilizados.

Para se ter uma ideia se uma determinada carga fatorial seria alta ou baixa (considerando variáveis padronizadas), HAIR et al. (2009) propõem um critério construído a partir de uma análise de poder, com recomendações conforme o tamanho amostral. Tais recomendações estão reproduzidas na Tabela 1.

Tabela 1: Magnitudes de cargas fatoriais dignas de atenção, conforme o tamanho amostral (fonte: HAIR et al., 2009).

Carga Fatorial	Tamanho da Amostra
0,30	350
0,35	250
0,40	200
0,45	150
0,50	120
0,55	100
0,60	85
0,65	70
0,70	60
0,75	50

2.1.3 Estimação

Na análise fatorial, o processo de estimação consiste na obtenção de estimativas para as cargas fatoriais (elementos da matriz $\mathbf{L}$) e para os fatores específicos ψ_i (elementos da matriz diagonal $\mathbf{\Psi}$).

A estimação se baseia na relação 2.15, substituindo a matriz paramétrica $\text{Cov}(\mathbf{x})$ pela matriz amostral $\mathbf{XX'}$ (ver relação 2.4), que corresponderá a uma matriz de covariâncias ou a uma matriz de correlações, caso as variáveis tenham sido padronizadas.

Não é do escopo deste material apresentar um detalhamento matemático aprofundado dos métodos de estimação utilizados em análise fatorial, mas sim apresentar os princípios básicos dos principais métodos, o que é feito a seguir.

Método dos Componentes Principais

Uma das soluções mais simples para a estimação de $\mathbf{L}$ e $\mathbf{\Psi}$ consiste em realizar uma análise de componentes principais a partir da matriz $\mathbf{XX'}$, tomando-se as cargas referentes apenas aos k primeiros autovalores λ_i, cargas essas presentes na matriz $\mathbf{C}$ (definida em 2.7).

Ou seja, o método admite que de fato os k primeiros componentes explicariam grande parte da variação, bem como a maior parte (idealmente, a totalidade) das covariâncias presentes em $\mathbf{XX'}$ (pois a variação remanescente, devida aos fatores específicos, segundo o modelo fatorial é dada pela matriz $\mathbf{\Psi}$, que é uma matriz diagonal). Trata-se de pressuposições restritivas, que possivelmente seriam difíceis de serem satisfeitas na prática. Assim, este método é apresentado aqui por razões didáticas, não sendo em geral utilizado em situações reais, nem disponível nas ferramentas computacionais de análise.

Método dos Fatores Principais

O método dos fatores principais consiste em um método numérico iterativo, visando melhorar as estimativas em cada iteração. Conforme visto na relação 2.17, a variância de uma variável X_i é dada pela soma da comunalidade com a variância específica, ou seja, $h_i^2 + \psi_i$.

Este método consiste em uma sequência de passos recursivos. A partir de estimativas iniciais para as cargas fatoriais e os fatores específicos, retorna-se à matriz $\mathbf{XX'}$, substituindo os elementos da diagonal pelas estimativas de comunalidades h_i^2. Isto conduz

a novas estimativas de $\mathbf{L}$ e $\mathbf{\Psi}$. O procedimento prossegue iterativamente, até que haja uma estabilização das estimativas. Este método foi talvez o mais utilizado nas primeiras décadas a partir do surgimento da análise fatorial, até o advento do método da máxima verossimilhança e sua disponibilidade em ferramentas computacionais.

Método da Máxima Verossimilhança

Admitindo que o conjunto das p variáveis X_i sejam quantitativas contínuas, o método da máxima verossimilhança em geral se baseia na distribuição normal p-variada, especificando que a matriz Cov($\mathbf{x}$) seja dada conforme o modelo fatorial, ou seja, conforme a relação 2.15. Este método também demanda procedimentos numéricos, mas está implementado na maioria dos *softwares* de análise.

Pode-se dizer que o método da máxima verossimilhança poderia ser considerado um padrão de análise, dada sua ampla popularidade.

2.1.4 Rotação de Fatores

Conforme apontado anteriormente, na análise fatorial a situação ideal (conforme a proposta do modelo fatorial) consiste em se ter cada variável com carga fatorial alta em determinado fator, e relativamente baixa nos demais fatores. Quando isso não acontece, uma possível solução consiste em se rotacionar os vetores dos fatores (obtidos com o ajuste de um primeiro modelo ortogonal, inicialmente). Em um processo de rotação, os valores das cargas fatoriais se alteram, embora, conforme apontado anteriormente, o valor da comunalidade h_i^2 permaneça inalterado.

Na análise de componentes principais, embora seja matematicamente possível rotacionar os eixos dos componentes, este procedimento não guardaria muito sentido, uma vez que comprometeria a própria definição de componentes principais, como sendo combinações que contabilizam proporções decrescentes da variação.

Mas na análise fatorial a rotação dos fatores permite uma grande flexibilização do modelo. Pode-se implementar algoritmos que busquem ângulos de rotação que maximizem a condição acima, ou seja, a de que cada variável tenda a ter uma carga fatorial elevada em um único fator.

A rotação pode ser ortogonal, quando os fatores ortogonais são rotacionados em conjunto, ou oblíqua, na qual os fatores deixarão de apresentar ângulo reto entre si.

Uma rotação ortogonal pode ser facilmente realizada, multiplicando-se a matriz **L** por alguma matriz ortogonal, conforme o ângulo da rotação. Segundo HAIR et al. (2009), existem dois principais algoritmos de obtenção de melhores rotações ortogonais, os chamados métodos *quadrimax* e *varimax*. No método *quadrimax*, busca-se uma rotação ortogonal que tende a maximizar as cargas fatoriais de cada variável em um único fator, se possível. Já no método *varimax* procura-se maximizar as cargas fatoriais de cada fator em um número pequeno de variáveis. Uma vez que o objetivo da rotação é o da melhor caracterização do significado latente de cada fator, o método *varimax* geralmente tem sido preferido.

Na rotação oblíqua, os fatores passarão a apresentar uma associação entre si, o que consiste numa grande flexibilização quanto ao leque de aplicações, permitindo que no fenômeno sendo estudado as variáveis latentes sejam correlacionadas, o que faz sentido em uma ampla gama de situações. Mesmo com uma rotação oblíqua, o valor da comunalidade não é afetado.

Conforme MATOS e RODRIGUES (2019) apontam, os dois métodos de rotação oblíqua mais importantes são os métodos *promax* e *oblimin*. O primeiro método consiste de um algoritmo mais rápido, indicado quando se trabalha com conjuntos de dados muito grandes. Não sendo este o caso, contudo, em geral o método *oblimin* corresponderá a um procedimento mais indicado.

2.1.5 Estimação de Escores

Uma vez ajustado um modelo fatorial, com ou sem rotacionamento (o qual em geral é feito), pode haver interesse na estimação (ou, mais apropriadamente, na predição) do valor de cada um dos fatores comuns para cada unidade amostral, lembrando que os fatores são variáveis aleatórias não observáveis.

Estes valores preditos são denominados *escores fatoriais*. Indicam a posição de cada indivíduo em relação ao fator latente, com base em suas respostas ou valores nas variáveis observadas. Um escore fatorial alto para um indivíduo em um determinado fator sugere que o indivíduo exibe uma alta quantidade do traço ou característica representado por este fator. Escores baixos indicam o oposto. Em geral, os escores são estimados em sua forma padronizada, com média 0 e variância unitária.

Em geral, da mesma maneira que o termo *carga*, o termo *escore* também é emprestado da análise fatorial para a análise de componentes principais. No caso desta última técnica,

os escores seriam obtidos diretamente da matriz $\mathbf{G}$, definida em 2.8.

Existem basicamente dois métodos de estimação de escores fatoriais, o método dos quadrados mínimos ponderado, e o método da regressão.

O método dos quadrados mínimos ponderado se baseia no modelo fatorial apresentado em 2.12, ou seja:

$$\mathbf{x} = \mathbf{L}\mathbf{f} + \boldsymbol{\epsilon}$$

adaptando-o, utilizando a matriz observada $\mathbf{X}$, e substituindo $\mathbf{L}$ e $\boldsymbol{\Psi}$ pelas estimativas $\hat{\mathbf{L}}$ e $\hat{\boldsymbol{\Psi}}$ obtidas com o ajuste do modelo fatorial:

$$\mathbf{X} = \hat{\mathbf{L}}\mathbf{F} + \mathbf{E} \tag{2.19}$$

sendo $\mathbf{F}$ a matriz $k \times n$ dos escores a serem estimados, e $\mathbf{E}$ a matriz $p \times n$ dos fatores específicos, com matriz de covariâncias admitida como sendo igual a $\hat{\boldsymbol{\Psi}}$. Como o modelo fatorial prevê que os elementos desta matriz diagonal sejam diferentes, trata-se de uma situação análoga à de uma condição de variâncias heterogêneas, e daí a sugestão do uso do método de quadrados mínimos ponderado.

A utilização deste método conduz à estimação da matriz $\mathbf{F}$. Suas colunas $\hat{\mathbf{f}}_j$ conterão as estimativas de escore para o j-ésimo elemento da amostra, sendo estimadas, conforme este método, por:

$$\hat{\mathbf{f}}_j = \left(\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}\right)^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\mathbf{x}_j \tag{2.20}$$

O chamado método da regressão é semelhante ao método dos quadrados mínimos ponderado, no sentido de também se basear em 2.19. Porém, o método da regressão leva em conta que os vetores $\mathbf{f}_j$ são vetores aleatórios, trabalhando com a distribuição conjunta destes com as variáveis observadas, utilizando momentos de distribuições condicionais.

Se representarmos o vetor de estimativas obtido pelo método dos quadrados mínimos ponderado por $\hat{\mathbf{f}}_j^{QM}$, e o vetor de estimativas obtido pelo método da regressão por $\hat{\mathbf{f}}_j^{R}$, tem-se (JOHNSON e WICHERN, 2007) que as soluções dos dois métodos estão relacionadas, mediante:

$$\hat{\mathbf{f}}_j^{QM} = \left[\mathbf{I} + \left(\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}\right)^{-1}\right]\hat{\mathbf{f}}_j^{R}$$

sendo que os autores apontam que as estimativas de ambos estes métodos podem ser muito semelhantes.

2.1.6 Etapas de Uma Análise Fatorial

MATOS e RODRIGUES (2019) descrevem, de maneira didática, as principais etapas de uma análise fatorial. Abaixo, é apresentada uma relação destas etapas, em uma compilação livremente adaptada.

1. **Adequação do Tamanho da Amostra**. O ajuste de um modelo fatorial adequado em grande medida depende do tamanho da amostra utilizado, pois, quanto maior o número de variáveis, maior o número de parâmetros a serem estimados. HAIR et al. (2009) apresentam alguns critérios empíricos para verificar se o tamanho da amostra é adequado, apontando ser difícil um bom ajuste com menos de 50 observações, sendo recomendado pelo menos 100. Os autores ainda recomendam que o número de observações seja pelo menos cinco vezes o número de variáveis consideradas, sendo preferíveis proporções de dez para um, ou maiores. Trata-se de critérios empíricos, pois em geral o tamanho de amostra adequado dependerá da maior ou menor associação entre as variáveis, da magnitude das cargas fatoriais, e consequentemente da ocorrência de maior ou menor comunalidade (maiores comunalidades demandariam tamanhos de amostra menores). De qualquer forma, são critérios que oferecem uma referência para a verificação da adequação do tamanho da amostra.

2. **Obtenção da Matriz de Correlações**. O ponto de partida para a análise fatorial (bem como o da análise de componentes principais) é a matriz de correlações entre as variáveis consideradas no estudo. Em se tratando de variáveis contínuas (como no presente trabalho), tais correlações corresponderão a correlações de Pearson. No entanto, é interessante destacar, a análise fatorial também pode ser empregada quando parte, ou a totalidade das variáveis consideradas não são quantitativas. Neste caso, basta trabalhar com coeficientes de correlação apropriados em cada caso:

    - Correlação bisserial: medida de associação entre uma variável quantitativa e uma variável binária.
    - Correlação polisserial: medida de associação entre uma variável quantitativa e uma variável categórica ordinal com três ou mais categorias.
    - Correlação tetracórica: medida de associação entre duas variáveis binárias.
    - Correlação policórica: medida de associação entre duas variáveis categóricas ordinais com três ou mais categorias.

3. **Teste de Bartlett**. Conforme discorrido anteriormente, a análise fatorial somente é adequada se as variáveis em questão apresentarem associação (correlações significativas) entre si. Dessa forma, o Teste de Bartlet consiste no julgamento de uma hipótese de nulidade estabelecendo que a matriz de correlações seria a matriz identidade (ausência de associações). Rejeitada esta hipótese, isto seria um indicativo de que a matriz de correlação é adequada, para justificar a análise fatorial. Porém, para um bom ajustamento de um modelo fatorial, FIELD *et al.* (2012), citados por MATOS e RODRIGUES (2019), recomendam que a maioria dos elementos da matriz de correlações (fora da diagonal) tenham magnitude acima de 0,3, em valor absoluto.

4. **Medida de Kaiser-Meyer-Olkin (KMO)**. Antes de se realizar uma análise fatorial, o cálculo das medidas KMO é costumeiramente feito, correspondendo a medidas que auxiliam a verificação da adequação da base de dados para a análise fatorial. São medidas de adequação que variam entre 0 e 1, representando a proporção da variância das variáveis que poderia ser explicada pelos fatores. É possível obter tanto uma medida KMO geral, para o conjunto dos fatores, bem como uma medida para cada fator individualmente. FIELD *et al.* (2012), citados por MATOS e RODRIGUES (2019), oferecem critérios para verificar a adequação da base de dados, conforme a magnitude das medidas KMO. Estes critérios estão reproduzidos na Tabela 2.

Tabela 2: Adequação da base de dados, conforme a magnitude das medidas de adequação KMO (fonte: FIELD *et al.*, 2012; citados por MATOS e RODRIGUES, 2019).

Medida KMO	Adequação da Amostra
$< 0{,}5$	Inaceitável
0,5 a 0,7	Medíocre
0,7 a 0,8	Boa
0,8 a 0,9	Ótima
$> 0{,}9$	Excelente

5. **Determinação do Número de Fatores**. A determinação do número de fatores é um dos aspectos mais importantes da análise fatorial, pois corresponde à escolha de um modelo fatorial apropriado. Frequentemente faz-se necessário trabalhar com modelos de diferentes números de fatores, para verificar qual seria mais adequado. A ideia é a de reduzir a dimensionalidade dos dados, mas sem perder informação relevante.

A literatura sugere três principais critérios para subsidiar a escolha do número de fatores: o critério de Kaiser, a porcentagem de variância explicada, e o diagrama de inclinação (em inglês, *scree plot*). Todos estes três critérios se baseiam na magnitude dos autovalores da matriz de correlações original (componentes principais) ou modificada (análise fatorial). No caso da análise fatorial, são reportados os autovalores da matriz de correlações modificada conforme o método de estimação dos fatores principais, considerando p fatores. Conforme apontado anteriormente, a modificação na matriz de correlações consiste na substituição dos elementos de sua diagonal pelas comunalidades estimadas.

O critério de Kaiser propõe considerar um número de fatores correspondente ao número de autovalores maiores que 1. O critério de Kaiser pode ser um pouco restritivo quanto ao número de fatores, e assim é sugerido o segundo critério, de maneira a reter no modelo fatorial um número de fatores igual ao número de autovalores que tenham, em conjunto, uma proporção acumulada alta de variância explicada, sendo comum admitir uma proporção de pelo menos 60%.

Já o diagrama de inclinação (*scree plot*) é um critério muito utilizado tanto na análise fatorial, como na própria análise de componentes principais. Este diagrama consiste em um gráfico dispondo a identificação dos autovalores (1, 2, ...) no eixo das abscissas, ordenados conforme as suas magnitudes, que por sua vez são dispostas no eixo das ordenadas. Por fim, os pares ordenados deste gráfico são ligados por segmentos de reta, formando uma espécie de "curva".

A ideia seria identificar o número de fatores como aquele correspondente ao ponto de máxima curvatura desta curva do gráfico, ou ainda àquele ponto onde a inclinação da curva tenderia a se estabilizar. Não seria considerado um número de fatores acima desse ponto, pois explicariam proporções de variância relativamente baixas. Em relação a este ponto de referência desta "curva", ressalta-se que se trata de um ponto de máxima curvatura, e não de um ponto de inflexão, sendo esta também uma confusão comum.

6. **Extração dos Fatores**. Definido o número de fatores, estes são por fim extraídos, utilizando um dos métodos de estimação descritos anteriormente (geralmente fatores principais ou máxima verossimilhança), sendo que há uma certa predominância na escolha pelo método da máxima verossimilhança.

7. **Rotação dos Fatores**. A rotação de fatores quase sempre é feita na análise fatorial, pois (eventualmente) possibilita tornar a interpretação dos fatores mais simples.

Conforme descrito anteriormente, partindo de um conjunto inicial de fatores ortogonais, a rotação pode ser ortogonal ou oblíqua. Na rotação ortogonal os fatores permanecem independentes entre si, enquanto que na rotação oblíqua admite-se a possibilidade de que os fatores, ou variáveis latentes, guardem diferentes graus de associação entre si, o que pode fazer sentido em uma ampla gama de situações.

O método de rotação ortogonal mais empregado é o *varimax*, enquanto que o método de rotação oblíqua mais utilizado é o *oblimin*.

Conforme apontado anteriormente, após a rotação a situação ideal seria a de que cada variável tivesse uma carga elevada em um dos fatores, e relativamente baixa, nos demais. Isto contribuiria para facilitar a interpretação dos fatores. Neste sentido, é interessante levar em consideração o chamado Índice de Complexidade de Hofmann (HOFMANN, 1978). Após o ajuste de um modelo fatorial, este índice é calculado para cada variável i, sendo dado por:

$$\frac{\left(\sum_{m=1}^{k} l_{im}^2\right)^2}{\sum_{m=1}^{k} l_{im}^4} \tag{2.21}$$

sendo l_{im} (conforme definido em 2.10) a carga fatorial da i-ésima variável sobre o m-ésimo fator. O valor do índice 2.21 fornece uma estimativa acerca do número de fatores aos quais a variável i estaria associada. Na situação ideal, tais valores seriam próximos de 1.

8. **Interpretação dos Fatores**. Com base nas variáveis que têm cargas fatoriais elevadas, procura-se identificar a natureza de cada fator. Esta interpretação para os diferentes fatores, idealmente, ajudará na descrição da natureza da associação entre as variáveis.

9. **Estimação dos Escores**. Uma vez ajustado o modelo fatorial, caso haja interesse, poderão ser obtidas as estimativas dos escores, ou seja, dos valores de cada um dos fatores para cada elemento da amostra.

2.2 O Curso de Estatística da UFOP

O curso de Bacharelado em Estatística da UFOP teve sua criação aprovada pelo seu Conselho Universitário em dezembro de 2007, como parte das ações do Programa REUNI (Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais),

lançado naquele mesmo ano, em âmbito nacional. O curso oferta anualmente 40 vagas, em período noturno, com ingresso no segundo semestre letivo de cada ano, e com duração prevista de quatro anos. A primeira turma iniciou suas atividades em agosto de 2008.

O Sistema de Seleção Unificada (Sisu) é a principal forma de ingresso no curso de Estatística, oferecendo vagas a participantes do Exame Nacional do Ensino Médio (Enem). A política inclusiva da UFOP prevê cotas (reserva de vagas) a portadores de deficiência, candidatos de baixo nível socioeconômico que tenham estudado em escolas públicas, bem como candidatos que se autodeclaram pretos e pardos.

Segundo o Projeto Pedagógico do Curso (PPC) atualmente em vigor (UNIVERSIDADE FEDERAL DE OURO PRETO, 2017), o curso de Bacharelado em Estatística da UFOP tem como objetivo:

> *(...) propiciar uma formação sólida e atual aos seus discentes, oferecendo disciplinas desde as áreas de fundamentos, tais como Matemática, Computação, Teoria de Probabilidades e Inferência Estatística, até as disciplinas mais profissionalizantes como: Técnicas de Amostragem, Análise de Séries Temporais, Controle Estatístico de Qualidade, Análise de Sobrevivência e Confiabilidade, Bioestatística, Estatística Espacial, entre outras.*

O curso passou pelo processo de Reconhecimento de Curso pelo MEC em 2013, tendo recebido conceito final 4 (em uma escala que vai até 5), caracterizando um "perfil muito bom de qualidade". Em 2019 o curso foi novamente avaliado pelo MEC, no primeiro processo de Renovação de Reconhecimento de Curso, tendo o curso sido renovado com conceito final 5.

Até o segundo semestre letivo de 2023, no curso de Estatística da UFOP haviam se graduado 118 bacharéis, conforme detalhamento apresentado na Tabela 3.

O curso sofreu duas alterações curriculares, em 2010 e 2012, e uma reforma curricular, em 2017. Nas alterações curriculares foram criadas algumas disciplinas adicionais, enquanto que na reforma curricular de 2017 houve uma ampla restruturação da grade, caracterizando um novo currículo. Para uma melhor compreensão sobre como o conjunto de dados utilizado no presente trabalho foi construído, é importante destacar algumas mudanças na grade que ocorreram com a reforma curricular de 2017.

Uma das principais mudanças da reforma foi a eliminação de disciplinas de 90 horas semestrais ofertadas pelo Departamento de Estatística, carga horária essa que foi tida como

Tabela 3: Número de bacharéis formados pelo curso de Estatística da UFOP, conforme o ano e o semestre de colação de grau.

	Semestre		
Ano de Colação	Primeiro	Segundo	TOTAIS
2012	0	2	2
2013	5	3	8
2014	11	7	18
2015	3	3	6
2016	9	1	10
2017	5	1	6
2018	8	1	9
2019	8	0	8
2020	8	5	13
2021	2	8	10
2022	8	7	15
2023	3	10	13
TOTAIS	70	48	118

elevada e contraprodutiva, pedagogicamente. No primeiro currículo, havia duas disciplinas – Estatística I e Estatística II, ambas de 90 horas, que tiveram seus conteúdos programáticos redistribuídos em três novas disciplinas de 60 horas, no currículo 2 (respectivamente, Estatística Descritiva, Estatística I e Estatística II). Um processo semelhante ocorreu com as disciplinas Probabilidade I e II, que também tiveram seus conteúdos redistribuídos em três novas disciplinas de 60 horas, as Probabilidades I, II e III. Outra mudança relevante nesta reforma curricular de 2017 foi a de que algumas disciplinas eletivas passaram a ser obrigatórias, como Inferência Bayesiana, e Modelos Lineares Generalizados. Além disso, o conteúdo de Estatística Multivariada passou a ser abordado em uma única disciplina (no primeiro currículo eram duas disciplinas). Finalmente, também vale destacar que a disciplina de Pacotes Estatísticos II, com ênfase na linguagem **R**, passou de 30 horas para 60 horas semestrais.

O rendimento acadêmico destes bacharéis formados até o segundo semestre de 2023, nas diferentes disciplinas do curso, constituiu a base de dados para a realização do presente estudo, conforme detalhado no próximo Capítulo.

3 Metodologia

Este estudo utilizou um banco de dados contendo as notas de aprovação em disciplinas de todos os 118 alunos graduados no curso de Estatística da Universidade Federal de Ouro Preto (UFOP), até o segundo semestre letivo de 2023.

Conforme apontado no Capítulo anterior, desde o início do curso de Estatística em 2008 ocorreram mudanças significativas na grade curricular.

Assim, para a construção de um banco de dados unificado, abrangendo alunos de diferentes grades, alguns critérios tiveram que ser adotados. Apenas disciplinas que estavam presentes nas duas grades curriculares foram incluídas na análise. Disciplinas que faziam parte do conteúdo obrigatório apenas de uma das grades curriculares, como Inferência Bayesiana, Modelos Lineares Generalizados (GLM) e Estatística Multivariada II foram excluídas do banco.

Em função da alteração de carga horária da disciplina Pacotes Estatísticos II ocorrida com a mudança de grade, optou-se por calcular a média, para cada aluno, das notas entre Pacotes Estatísticos I e Pacotes Estatísticos II em ambas as grades, gerando uma única variável resposta contemplando ambas as disciplinas de Pacotes.

Um procedimento semelhante foi adotado para as disciplinas de Probabilidade, e para as disciplinas introdutórias de Estatística, sendo calculada uma nota média para cada uma dessas duas áreas.

Com estas adaptações, o banco de dados consistiu de notas de 25 disciplinas, ou grupos de disciplinas, como nas situações acima. Ao longo deste trabalho, cada uma destas disciplinas será referenciada conforme uma abreviação, conforme apresentado na Tabela 4.

Apesar destes procedimentos, o conjunto de dados ainda apresentou algum desbalanceamento, no sentido de apresentar alguns valores ausentes. O motivo mais recorrente para isso correspondeu a alunos que fizeram transferência de curso, ou que fizeram uso

Tabela 4: Abreviaturas utilizadas neste trabalho, para referenciar as disciplinas do curso de Estatística da UFOP.

Abreviação	Disciplina
MetCie	Metodologia Cientifica
GA	Geometria Analítica
Calc_I	Cálculo Diferencial e Integral I
Calc_II	Cálculo Diferencial e Integral II
Algebra	Introdução à Álgebra Linear
Progr	Programação de Computadores I
CalcNum	Cálculo Numérico
Pacotes	Pacotes Estatísticos
Estats	Estatísticas
Probs	Probabilidades
Inferencia	Inferência Estatística I e II
Mult	Estatística Multivariada
Regres	Análise de Regressão
ProcEst	Processos Estocásticos
Demog	Demografia
MNP	Métodos Não-Paramétricos
Amostr	Técnicas de Amostragem
PlanExp	Planejamento de Experimento
CEQ	Controle Estatístico de Qualidade
Series	Análise de Séries Temporais
ADC	Análise de Dados Categóricos
Sobrev	Análise de Sobrevivência
POM	Pesquisa de Opinião e Mercado
Lab	Laboratório Supervisionado
MONO	Monografia

de vagas residuais (como portadores de diploma de graduação). Em tais casos, é comum a solicitação de aproveitamento de disciplinas cursadas anteriormente, e assim as notas para tais disciplinas aproveitadas em geral não constam dos históricos escolares.

Em um primeiro momento, este conjunto de dados foi sumariado utilizando algumas técnicas de Estatística Descritiva, como cálculo de medidas resumo, geração de diagramas de caixa (*boxplots*), para a distribuição de frequências, e obtenção das correlações de Pearson. Com tais técnicas exploratórias buscou-se identificar alguns padrões iniciais no conjunto.

Em seguida, os dados foram submetidos a uma análise fatorial, conforme as etapas do Capítulo anterior, conforme recomendações de MATOS e RODRIGUES (2019).

Uma vez que o número de correlações entre as 25 disciplinas é elevado (300 correla-

ções), optou-se por utilizar um nível de significância mais conservador, igual a $\alpha = 0,01$, para se testar a significância de cada correlação individualmente.

Eventualmente, algumas disciplinas vieram a ser excluídas da análise, com base na sugestão de FIELD *et al.* (2012), citados por MATOS e RODRIGUES (2019), de que a maioria das correlações tenham magnitude acima de 0,3, em valor absoluto. Neste sentido, os valores das medidas de adequação KMO subsidiaram as eventuais exclusões de disciplinas.

O modelo fatorial foi então ajustado, com a definição do número de fatores sendo embasada nos critérios apresentados no Capítulo anterior, mas também levando em conta a maior ou menor facilidade de interpretação dos fatores obtidos com o modelo. Foi empregado o método de estimação da máxima verossimilhança, seguida de rotação oblíqua dos fatores, visando à identificação de eixos de aprendizado eventualmente associados entre si.

As análises foram realizadas utilizando a linguagem **R** (R CORE TEAM, 2024). Em relação às técnicas descritivas, foi utilizado o comando *summary()*, do núcleo básico da linguagem, para obtenção das estatísticas-resumo, e o comando *cor()* para obtenção das correlações de Pearson. Os diagramas de caixa foram gerados utilizando o pacote *ggplot2* (WICKHAM, 2016), e representações gráficas da matriz de correlações foram obtidas com o pacote *corrplot* (WEI e SIMKO, 2024).

Em relação à análise fatorial, esta foi realizada utilizando o pacote *psych* (REVELLE, 2024). Através deste pacote foi possível realizar o Teste de Bartlet, gerar o gráfico *scree*, obter as estatísticas KMO, bem como ajustar os modelos fatoriais aqui considerados.

4 Resultados e Discussão

Antes da análise fatorial propriamente dita, são apresentados a seguir alguns resultados de uma análise exploratória (estatística descritiva) dos dados, para uma caracterização preliminar sua, e identificação de alguns padrões mais evidentes.

4.1 Estatística Descritiva

Para uma caracterização preliminar do conjunto de dados utilizado neste estudo, três recursos de estatística descritiva foram empregados; o cálculo de medidas-resumo, a construção de *boxplots* (diagramas de caixa) para visualização da distribuição de frequências, e a obtenção da matriz de correlações de Pearson.

A Tabela 5 apresenta as medidas descritivas: média, desvio padrão, quartis, mínimo e máximo, referentes às notas de aprovação das disciplinas do curso de Estatística da UFOP consideradas neste estudo. Conforme se pode observar nesta Tabela, foram utilizadas linhas horizontais para separar as disciplinas em 8 grupos, para facilidade de visualização:

- Grupo 1: Dadas as suas especificidades, optou-se por definir este primeiro grupo contendo apenas a disciplina de Metodologia Científica;
- Grupo 2: Disciplinas da área de Matemática;
- Grupo 3: Disciplinas de natureza computacional;
- Grupo 4: Dado seu caráter interdisciplinar, optou-se por deixar o resultado das disciplinas de Estatística em único grupo;
- Grupo 5: Disciplinas estatísticas de caráter mais teórico (Probabilidades e Inferências);
- Grupo 6: Disciplinas estatísticas de caráter teórico ou aplicado mais geral, podendo ser utilizadas em outras áreas aplicadas da Estatística mais específicas (Análise

Multivariada, Regressão e Processos Estocásticos);

- Grupo 7: Disciplinas referentes a técnicas utilizadas em áreas aplicadas da Estatística;

- Grupo 8: Disciplinas com predominância de atividades extra-classe, relativas ao comprimento de habilidades específicas (Laboratório Supervisionado e Monografia).

Tabela 5: Medidas descritivas referentes às notas de aprovação das disciplinas do curso de Estatística da UFOP, reunidas em 8 grupos, conforme suas características.

Disciplina[1]	Grupo[2]	Média	Desvio Padrão	$Q_{0,25}$	$Q_{0,50}$	$Q_{0,75}$	Mínimo	Máximo
MetCie	1	8,356	0,776	8,000	8,450	8,925	6,0	10,0
GA	2	7,055	1,034	6,100	6,500	7,500	6,0	9,9
Calc_I		6,914	1,056	6,000	6,500	7,500	6,0	10,0
Calc_II		7,422	1,228	6,200	7,400	8,100	6,0	10,0
Algebra		7,296	1,183	6,200	7,000	8,150	6,0	10,0
Progr	3	7,415	1,267	6,200	7,100	8,300	6,0	10,0
CalcNum		7,550	1,252	6,500	7,300	8,500	6,0	10,0
Pacotes		8,565	0,932	8,00	8,600	9,300	6,0	10,0
Estats	4	7,682	1,035	6,800	7,600	8,537	6,0	9,7
Probs	5	7,556	1,035	6,763	7,400	8,238	6,0	9,85
Inferencia		7,782	0,927	7,000	7,400	8,500	6,05	9,9
Mult	6	8,498	0,966	7,900	8,600	9,300	6,0	10,0
Regres		7,743	1,159	6,875	7,650	8,725	6,0	10,0
ProcEst		6,873	0,931	6,000	6,700	7,300	6,0	9,7
Demog	7	8,432	1,019	7,850	8,650	9,000	6,0	10,0
MNP		8,146	1,202	7,200	8,250	9,200	6,0	10,0
Amostr		7,631	1,104	6,700	7,600	8,400	6,0	10,0
PlanExp		8,432	1,070	7,800	8,600	9,400	6,0	10,0
CEQ		8,275	0,978	7,900	8,300	9,000	6,0	10,0
Series		8,149	0,971	7,375	8,400	9,000	6,0	9,8
ADC		8,318	1,169	7,300	8,300	9,300	6,0	10,0
Sobrev		8,175	1,227	7,300	8,300	9,300	6,0	10,0
POM		9,244	0,935	8,500	9,700	10,000	7,0	10,0
Lab	8	9,349	0,793	9,000	10,000	10,000	7,0	10,0
MONO		9,548	0,654	9,300	9,800	10,000	6,0	10,0

[1] A codificação das disciplinas está detalhada na Tabela 4.

[2] Critérios de agrupamento definidos no texto.

A partir das informações contidas na Tabela 5, é possível verificar que os grupos que apresentaram maiores notas de aprovação foram, respectivamente, os grupos 8 (Laboratório e Monografia), 1 (Metodologia Científica), e 7 (disciplinas aplicadas). Nestes três

grupos a nota média de aprovação foi superior a 8. Em todos os demais grupos a nota média foi inferior a 8, mas superior a 7. O grupo que apresentou menor nota média de aprovação foi o grupo 2 (disciplinas de Matemática), igual a 7,17.

Em relação à variabilidade, aparentemente o grupo de disciplinas que apresentou maior dispersão foi o grupo 3 (disciplinas da área de Computação). Este grupo apresentou tanto o maior desvio padrão médio (1,15), como também a maior diferença interquartílica média (1,8). Também baseando-se nestas medidas, os grupos mais homogêneos (com menores desvios padrões e menores diferenças interquartílicas) foram os grupos 8 e 1.

Quanto às notas de aprovação mínimas e máximas, houve pouca variação no conjunto de dados, uma vez que estas medidas estiveram sempre próximas de seus limites possíveis (6 e 10). Algumas poucas exceções (especialmente em relação ao mínimo) foram as disciplinas de Laboratório Supervisionado e Pesquisa de Opinião e Mercado, com notas mínimas iguais a 7 (Tabela 5).

Em seguida, foram construídos diagramas de caixa (*boxplots*) das distribuições de frequência de cada disciplina, reunidas conforme estes 8 grupos. Inicialmente, na Figura 1 são apresentados os *boxplots* daqueles grupos contendo um único ou dois conjuntos de notas, quais sejam, os grupos 1 (Metodologia Científica), 4 (Estatísticas) e 8 (Laboratório e Monografia).

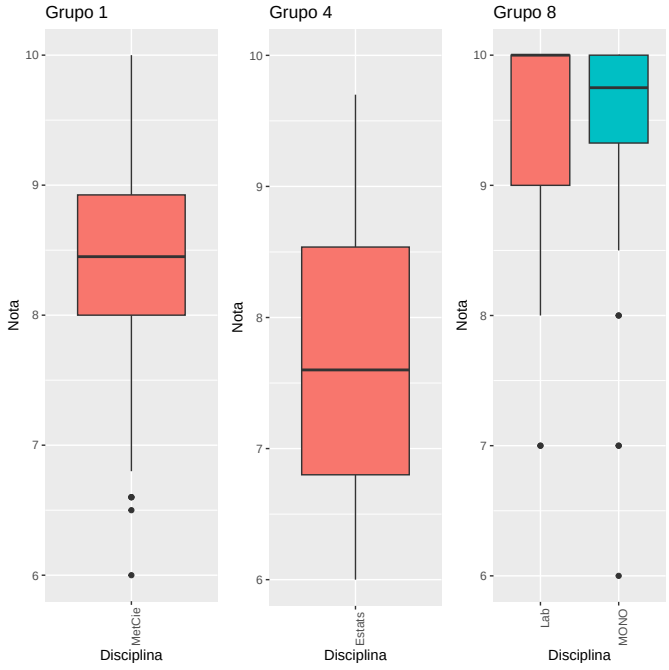

Figura 1: Diagramas de caixa das notas de aprovação de três grupos de disciplinas: grupo 1 (Metodologia Científica), grupo 4 (Estatísticas) e grupo 8 (Laboratório e Monografia).

Observa-se na Figura 1 que tanto a nota de Metodologia Científica como o desempenho médio nas disciplinas de Estatística apresentaram uma distribuição aproximadamente simétrica, com medianas próximas ao centro das caixas e *whiskers* de tamanhos semelhantes. Contudo, as disciplinas de Estatística apresentaram uma dispersão maior (maior caixa) e uma nota mediana consideravelmente menor. Já as disciplinas do grupo 8 (Laboratório e Monografia), por se tratarem das disciplinas com maiores notas, seus *boxplots* não apresentaram *whiskers* superiores, com uma clara assimetria à esquerda (ou seja, com elevada concentração de notas próximas do máximo).

A Figura 2 apresenta os *boxplots* da distribuição de frequências das disciplinas dos grupos 2 (Matemática) e 3 (área computacional), que podem ser considerados grupos de disciplinas de área básica, no contexto do curso.

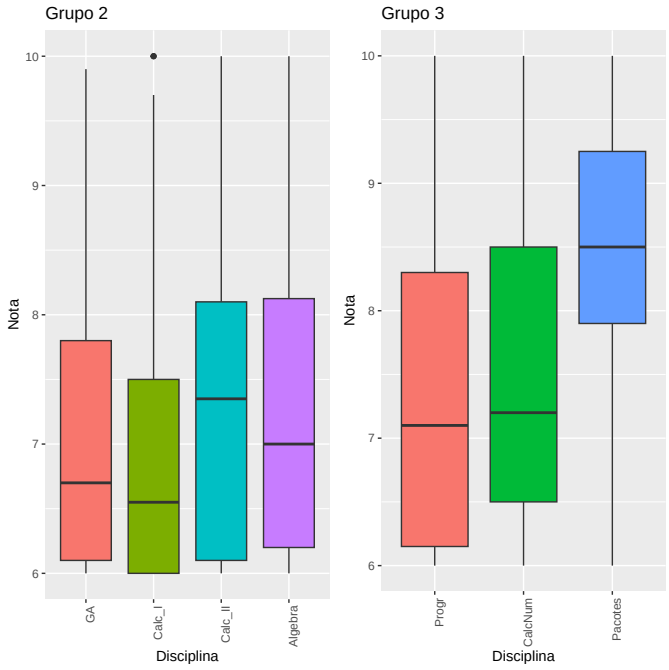

Figura 2: Diagramas de caixa das notas de aprovação de dois grupos de disciplinas: grupo 2 (Matemática) e grupo 3 (área computacional).

Observa-se na Figura 2 que, à exceção das disciplinas de Pacotes Estatísticos, todas as disciplinas apresentaram uma tendência de assimetria à direita, indicando uma tendência de a maioria dos graduados terem obtido uma nota de aprovação mais baixa. No grupo das disciplinas da área computacional as notas medianas estiveram acima de 7, notadamente o rendimento das disciplinas de Pacotes, acima de 8. Já nas disciplinas de Matemática, houve duas com nota mediana abaixo de 7: Geometria Analítica e Cálculo I. Tais disciplinas, na realidade, corresponderam àquelas com menor nota mediana, no elenco geral das disciplinas do curso (Tabela 5).

A Figura 3 apresenta os *boxplots* da distribuição de frequências das disciplinas dos grupos 5 (Probabilidades e Inferências) e grupo 6 (Multivariada, Regressão e Processos Estocásticos).

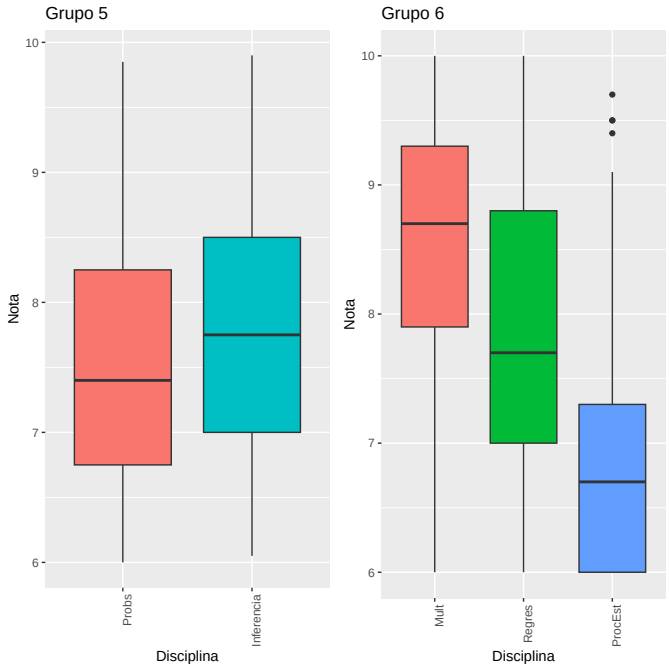

Figura 3: Diagramas de caixa das notas de aprovação de dois grupos de disciplinas: grupo 5 (Probabilidades e Inferências) e grupo 6 (Multivariada, Regressão e Processos Estocásticos).

Observa-se na Figura 3 que as disciplinas de Probabilidades, Inferências e de Regressão apresentaram distribuições com uma leve assimetria à direita. Nestas três disciplinas as notas medianas estiveram entre 7 e 8. A julgar pelo tamanho relativo da caixa, a disciplina de Regressão aparentemente apresentou maior variabilidade.

A disciplina de Processos Estocásticos apresentou uma assimetria à direita mais acentuada, evidenciada pela ausência de *whisker* inferior e pela ocorrência de *outliers* à direita, correspondendo aos alunos que se sobressaíram nesta disciplina. Já a disciplina de Estatística Multivariada apresentou uma distribuição assimétrica à esquerda, com tendência de concentração de notas de aprovação mais elevadas, e *whisker* superior menos pronunciado. A nota mediana desta disciplina esteve acima de 8,5 (Figura 3)

Finalmente, a Figura 4 apresenta os *boxplots* da distribuição de frequências das disciplinas do grupo 7, correspondendo ao conjunto de disciplinas de natureza aplicada em áreas específicas (Demografia, Métodos Não-Paramétricos, Amostragem, Planejamento de Experimento, Controle Estatístico de Qualidade, Séries Temporais, Análise de Dados Categóricos, Análise de Sobrevivência e Pesquisa de Opinião e Mercado).

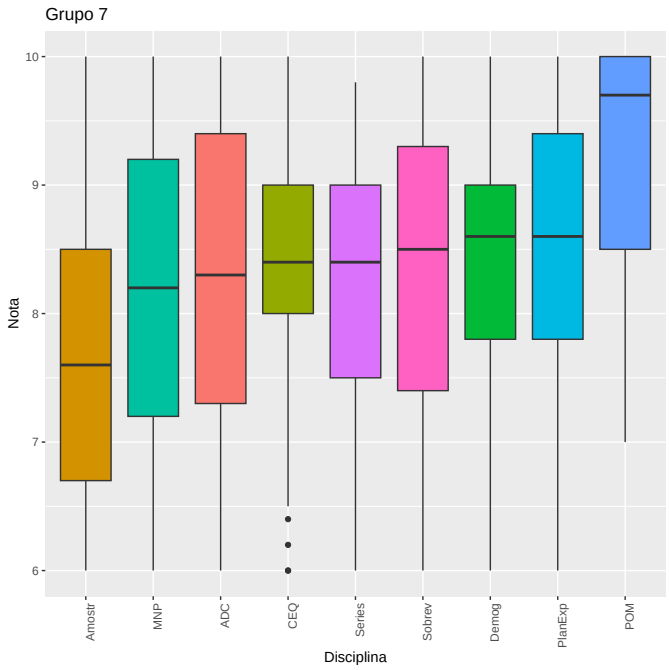

Figura 4: Diagramas de caixa das notas de aprovação do grupo 7 de disciplinas, de natureza aplicada em áreas específicas (ver codificação das disciplinas no texto).

Observa-se na Figura 4 que a maioria das disciplinas do grupo 7 apresentaram nota mediana entre 8 e 9. As exceções corresponderam às disciplinas de Amostragem (com nota mediana abaixo de 8) e de Pesquisa de Opinião e Mercado (acima de 9,5). Em função desta tendência de notas mais elevadas, as distribuições tenderam a apresentar graus diferenciados de assimetria à esquerda. Também aqui a exceção foi a disciplina de Amostragem, com uma distribuição levemente assimétrica à direita.

Cabe uma observação final, referente aos *boxplots* apresentados aqui. Comparando os quantis destes gráficos (que definem a "caixa" de cada um) com os quantis da Tabela 5, eventualmente ocorrem algumas pequenas diferenças. Isto se deve ao fato de que a função *summary()* e o pacote *ggplot2* utilizam métodos diferentes de estimação de quantis.

4.1.1 Correlações

Como última técnica descritiva, foram calculados e apresentados os coeficientes de correlação de Pearson (Tabela 6). Embora este coeficiente esteja apresentado aqui como uma medida resumo de associação, é importante destacar, desde já, que a obtenção da matriz de correlações consiste em uma das primeiras etapas de uma análise fatorial, a ser discutida no próximo tópico.

Tabela 6: Matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP.

	MetCie	GA	Calc_I	Calc_II	Algebra	Progr	CalcNum	Pacotes	Estats	Probs	Inferencia	Mult	Regres	ProcEst	Demog	MNP	Amostr	PlanExp	CEQ	Series	ADC	Sobrev	POM	Lab	MONO
MetCie	1	0,085	0,082	-0,010	-0,020	0,124	-0,043	0,072	0,112	0,089	0,153	0,132	0,124	0,117	0,063	0,141	0,240	0,097	0,236	0,207	0,206	-0,127	-0,295	-0,233	-0,004
GA	0,085	1	0,104	0,135	0,287	0,293	0,128	0,149	0,330	0,412	0,292	0,160	0,286	0,173	0,128	0,217	0,309	0,360	0,183	0,269	0,226	0,225	0,118	0,132	0,081
Calc_I	0,082	0,104	1	0,075	0,077	0,157	0,176	0,130	0,279	0,135	0,152	0,167	0,132	0,076	0,166	0,116	0,235	0,086	0,246	0,068	0,239	0,112	-0,043	0,064	0,147
Calc_II	-0,010	0,135	0,075	1	0,249	0,207	0,201	0,223	0,250	0,227	0,321	0,262	0,206	0,168	0,239	0,315	0,144	0,163	0,207	0,198	0,270	0,138	0,083	0,243	0,294
Algebra	-0,020	0,287	0,077	0,249	1	0,095	0,113	0,334	0,349	0,231	0,338	0,285	0,331	0,238	0,149	0,368	0,142	0,246	0,283	0,266	0,363	0,131	-0,051	0,216	0,127
Progr	0,124	0,293	0,157	0,207	0,095	1	0,202	0,345	0,435	0,302	0,264	-0,016	0,170	0,184	0,108	0,302	0,202	0,165	0,150	0,154	0,182	0,023	0,093	0,142	0,103
CalcNum	-0,043	0,128	0,176	0,201	0,113	0,202	1	0,287	0,387	0,448	0,307	0,136	0,104	0,255	0,186	0,250	0,251	0,229	0,258	0,172	0,266	-0,016	0,294	-0,149	0,172
Pacotes	0,072	0,149	0,130	0,223	0,334	0,345	0,287	1	0,428	0,278	0,385	0,254	0,142	0,378	0,251	0,556	0,359	0,227	0,333	0,206	0,245	0,187	0,152	0,322	0,225
Estats	0,112	0,330	0,279	0,250	0,349	0,435	0,387	0,428	1	0,589	0,481	0,345	0,371	0,511	0,301	0,641	0,494	0,465	0,554	0,346	0,559	0,240	0,053	0,112	0,207
Probs	0,089	0,412	0,135	0,227	0,231	0,302	0,448	0,278	0,589	1	0,438	0,324	0,395	0,328	0,288	0,411	0,284	0,315	0,353	0,266	0,445	0,258	0,267	-0,012	0,221
Inferencia	0,153	0,292	0,152	0,321	0,338	0,264	0,307	0,385	0,481	0,438	1	0,281	0,447	0,497	0,164	0,613	0,484	0,469	0,508	0,509	0,651	0,282	0,212	0,214	0,320
Mult	0,132	0,160	0,167	0,262	0,285	-0,016	0,136	0,254	0,345	0,324	0,281	1	0,285	0,206	0,257	0,413	0,213	0,214	0,424	0,276	0,380	0,128	-0,001	0,197	0,137
Regres	0,124	0,286	0,132	0,206	0,331	0,170	0,104	0,142	0,371	0,395	0,447	0,285	1	0,249	0,133	0,447	0,397	0,428	0,477	0,351	0,487	0,453	-0,195	0,050	-0,035
ProcEst	0,117	0,173	0,076	0,168	0,238	0,184	0,255	0,378	0,511	0,328	0,497	0,206	0,249	1	0,129	0,422	0,461	0,337	0,419	0,330	0,468	0,217	0,076	0,097	0,108
Demog	0,063	0,128	0,166	0,239	0,149	0,108	0,186	0,251	0,301	0,288	0,164	0,257	0,133	0,129	1	0,203	0,295	0,117	0,426	0,301	0,314	0,127	0,068	0,036	0,244
MNP	0,141	0,217	0,116	0,315	0,368	0,302	0,250	0,556	0,641	0,411	0,613	0,413	0,447	0,422	0,203	1	0,468	0,354	0,521	0,491	0,545	0,356	-0,050	0,294	0,239
Amostr	0,240	0,309	0,235	0,144	0,142	0,202	0,251	0,359	0,494	0,284	0,484	0,213	0,397	0,461	0,295	0,468	1	0,471	0,567	0,455	0,566	0,302	-0,085	0,054	0,174
PlanExp	0,097	0,360	0,086	0,163	0,246	0,165	0,229	0,227	0,465	0,315	0,469	0,214	0,428	0,337	0,117	0,354	0,471	1	0,471	0,388	0,451	0,257	0,142	0,005	0,104
CEQ	0,236	0,183	0,246	0,207	0,283	0,150	0,258	0,333	0,554	0,353	0,508	0,424	0,477	0,419	0,426	0,521	0,567	0,471	1	0,409	0,660	0,178	-0,208	0,058	0,155
Series	0,207	0,269	0,068	0,198	0,266	0,154	0,172	0,206	0,346	0,266	0,509	0,276	0,351	0,330	0,301	0,491	0,455	0,388	0,409	1	0,552	0,232	0,050	0,052	0,140
ADC	0,206	0,226	0,239	0,270	0,363	0,182	0,266	0,245	0,559	0,445	0,651	0,380	0,487	0,468	0,314	0,545	0,566	0,451	0,660	0,552	1	0,344	0,014	0,145	0,338
Sobrev	-0,127	0,225	0,112	0,138	0,131	0,023	-0,016	0,187	0,240	0,258	0,282	0,128	0,453	0,217	0,127	0,356	0,302	0,257	0,178	0,232	0,344	1	0,131	0,144	0,370
POM	-0,295	0,118	-0,043	0,083	-0,051	0,093	0,294	0,152	0,053	0,267	0,212	-0,001	-0,195	0,076	0,068	-0,050	-0,085	0,142	-0,208	0,050	0,014	0,131	1	0,180	0,206
Lab	-0,233	0,132	0,064	0,243	0,216	0,142	-0,149	0,322	0,112	-0,012	0,214	0,197	0,050	0,097	0,036	0,294	0,054	0,005	0,058	0,052	0,145	0,144	0,180	1	0,253
MONO	-0,004	0,081	0,147	0,294	0,127	0,103	0,172	0,225	0,207	0,221	0,320	0,137	-0,035	0,108	0,244	0,239	0,174	0,104	0,155	0,140	0,338	0,370	0,206	0,253	1

Por se tratar de um número elevado de coeficientes (300), optou-se também por representar a matriz de correlações em duas formas gráficas. A Figura 5 corresponde a uma representação gráfica da matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP.

Na Figura 5, a cada valor de correlação corresponde uma circunferência, com um setor circular cujo ângulo é proporcional à magnitude da correlação. Setores circulares com a cor azul representam correlações positivas, e setores em vermelho representam correlações negativas.

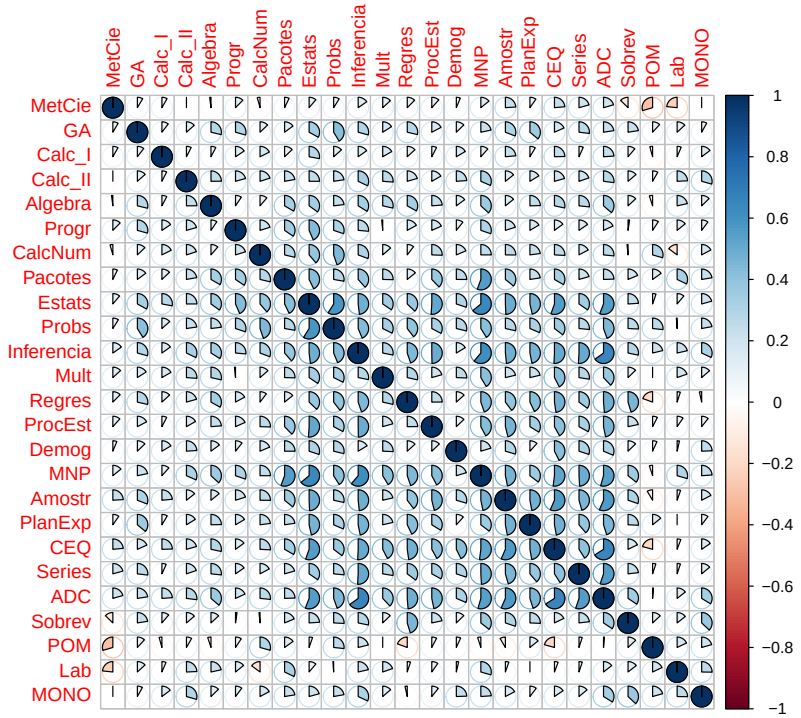

Figura 5: Representação gráfica da matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP. Ângulos dos setores circulares são proporcionais aos valores de correlação; setores em azul representam correlações positivas, e setores em vermelho representam correlações negativas.

A Figura 6 representa também a matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP, mas apresentado apenas as correlações significativas a 1% de probabilidade. Nesta Figura, são dispostas circunferências cujos raios são proporcionais às magnitudes das correlações. As circunferências de cor azul correspondem a correlações positivas significativas, enquanto que a única circunferência vermelha desta Figura representa uma correlação negativa significativa.

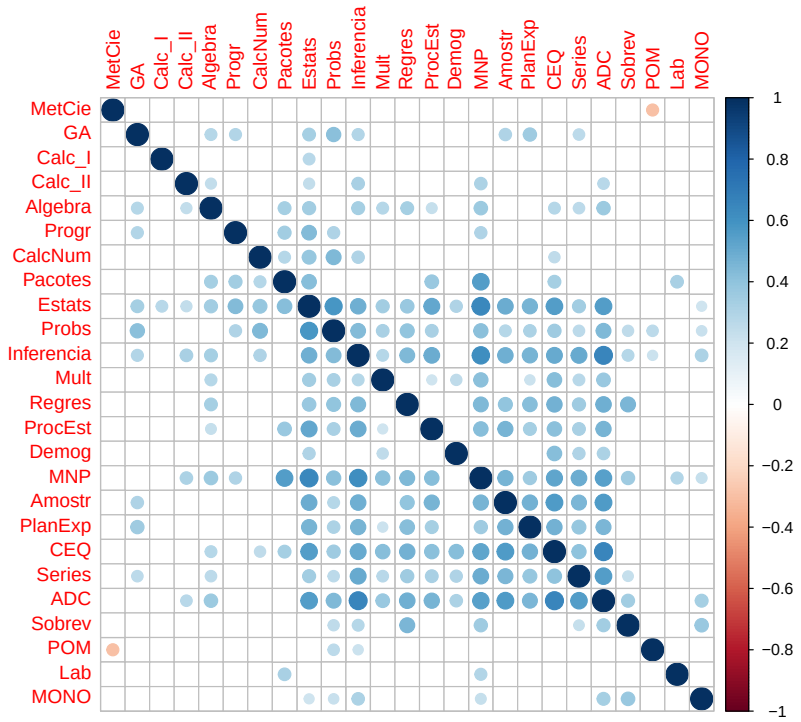

Figura 6: Representação gráfica da matriz de correlações entre notas de aprovação das disciplinas do curso de Estatística da UFOP, contendo apenas correlações significativas a 1% de probabilidade. Os raios das circunferências são proporcionais às magnitudes das correlações; circunferências azuis correspondem a correlações positivas, e a circunferência vermelha representa uma correlação negativa.

Considerando tanto a Tabela 6 como as Figuras 5 e 6, verifica-se que alguns poucos coeficientes de correlação apresentaram estimativas negativas, envolvendo as disciplinas Metodologia Científica, Pequisa de Opinião e Mercado (POM), Laboratório Supervisionado e Monografia. Contudo, consultando a Figura 6, verifica-se que o único coeficiente significativo negativo foi aquele entre as disciplinas Metodologia Científica e POM. Na realidade, este foi o único coeficiente envolvendo a disciplina de Metodologia que foi significativo. Isto ilustra o caráter e a natureza bastante específicos desta disciplina, em relação ao elenco total das disciplinas do curso. Dado que o restante (a imensa maioria) dos outros coeficientes de correlação significativos foram positivos (Figura 6), parece mais razoável admitir que esta correlação significativa negativa envolvendo a disciplina de Metodologia tenha sido consequência de um Erro Tipo I.

Consultando a Figura 6, também é interessante notar que a disciplina com maior número de correlações significativas (20) foi o grupo das Estatísticas (que envolve Estatística I e II, bem como Estatística Descritiva, para os alunos da grade curricular 2). As discipli-

nas de Estatística poderiam ser consideradas, por assim dizer, disciplinas que sintetizam o pensamento estatístico, e abordando os princípios inferenciais básicos, a partir das metodologias de análise mais simples. Sendo assim, trata-se de um resultado interessante, evidenciando o caráter central que tais disciplinas desempenham no curso.

Em seguida, as disciplinas com maior número de correlações significativas (18) foram as Inferências e a de Métodos Não-Paramétricos. No caso das Inferências, também seria natural admitir a possibilidade de várias correlações significativas, uma vez que fornecem a base para a análise de dados, e assim manifestando-se em outras disciplinas. Já no caso da disciplina de Métodos Não-Paramétricos as razões de tal comportamento não são tão evidentes, uma vez que os métodos não-paramétricos são pouco abordados em outras disciplinas. Contudo, conforme será discutido mais adiante no tópico de ajustamento de modelos fatoriais, uma das razões possíveis deste comportamento estaria relacionada ao fato de a disciplina de Métodos Não-Paramétricos ser uma das primeiras disciplinas de caráter aplicado a fazer uso intensivo de recursos computacionais, notadamente a linguagem **R**, uma característica que estará presente em várias outras disciplinas.

Ainda considerando a Figura 6, tem-se que outra disciplina que apresentou uma única correlação significativa foi o Cálculo I. As notas desta disciplina apenas se correlacionaram significativamente com as notas do grupo de disciplinas de Estatística.

Entre as disciplinas da área de Matemática, houve apenas duas correlações significativas (dentre as 6), envolvendo a disciplina de Álgebra Linear e as disciplinas de Geometria Analítica e Cálculo II, embora as estimativas não tenham sido muito elevadas (0,287 e 0,249, respectivamente, conforme Tabela 6). Entre as disciplinas da área computacional, duas das três correlações foram significativas, da disciplina de Pacotes com as disciplinas de Programação e Cálculo Numérico, embora as magnitudes também não tenham sido elevadas (0,345 e 0,287, respectivamente). Estas baixas correlações ilustram a dificuldade da formação de conjuntos individualizados de disciplinas, eventualmente representando eixos temáticos de aprendizado, baseando-se apenas em um agrupamento apriorístico conforme o conteúdo das disciplinas.

Finalmente, de uma maneira geral, vale a pena destacar aquelas correlações significativas de maior magnitude (acima de 0,6). Corresponderam às correlações entre: Controle Estatístico de Qualidade e Análise de Dados Categóricos (0,660), Inferências e Análise de Dados Categóricos (0,651), Métodos Não-Paramétricos e Estatísticas (0,641), e Métodos Não-Paramétricos e Inferências (0,613).

4.2 Análise Fatorial

Neste tópico, foi realizada uma análise fatorial, conforme as etapas descritas no Capítulo de Referencial Teórico, com a expectativa de se identificar eixos temáticos de aprendizado no Curso de Bacharelado em Estatística da UFOP, tendo por base as notas de aprovação de seus graduados nas diferentes disciplinas. Estas etapas estão descritas a seguir.

Adequação do Tamanho da Amostra

A base de dados original do presente estudo contém $n = 118$ observações (referentes aos 118 graduados no curso) de $p = 25$ variáveis (correspondentes às notas de aprovação em 25 disciplinas ou pequenos grupos de disciplinas).

HAIR *et al.* (2009) apontam ser difícil um bom ajuste de um modelo fatorial com menos de 50 observações, sendo recomendado pelo menos 100. A presente base de dados atende a este critério, mas os mesmos autores recomendam que o número de observações seja pelo menos cinco vezes o número de variáveis consideradas, e que são ainda preferíveis proporções de dez para um, ou mesmo maiores. Por esta recomendação, a base deveria conter no mínimo 125 observações. Isto sugeriu que a eventual retirada de algumas variáveis, mediante algum tipo de critério, poderia contribuir para a melhoria do ajuste.

Matriz de Correlações e Teste de Bartlet

A matriz de correlações entre as notas de aprovação de 25 disciplinas (ou pequenos grupos de disciplinas) foi obtida no item anterior de análise exploratória, tendo sido apresentada na Tabela 6 e nas Figuras 5 e 6. Trata-se de correlações de Pearson, e esta matriz é o ponto de partida para uma análise fatorial.

O Teste de Bartlet tem por hipótese nula a existência de uma matriz de correlações igual à matriz identidade. Com a presente base de dados, o teste apresentou uma estatística $\chi^2 = 904,67$, com um valor-p igual a $6,44 \times 10^{-62}$, com evidência muito forte de que a hipótese nula seja falsa.

Em situações práticas, é plausível admitir que em várias, senão na maioria destas situações o Teste de Bartlet será significativo, pois bastam algumas poucas correlações significativas para que a hipótese nula tenda a ser rejeitada. Isto não implica, necessariamente, que a estrutura de associações entre as variáveis seja tal que possibilite um bom

ajustamento na análise fatorial.

Em função disto, indo além do Teste de Bartlet, FIELD *et al.* (2012), citados por MATOS e RODRIGUES (2019), recomendam que a maioria dos elementos da matriz de correlações (fora da diagonal) tenham magnitude acima de 0,3, em valor absoluto. A distribuição de frequência das 300 correlações da presente base, considerando todas as 25 disciplinas (ou, mais precisamente, os 25 grupos de notas) está apresentada na Tabela 7.

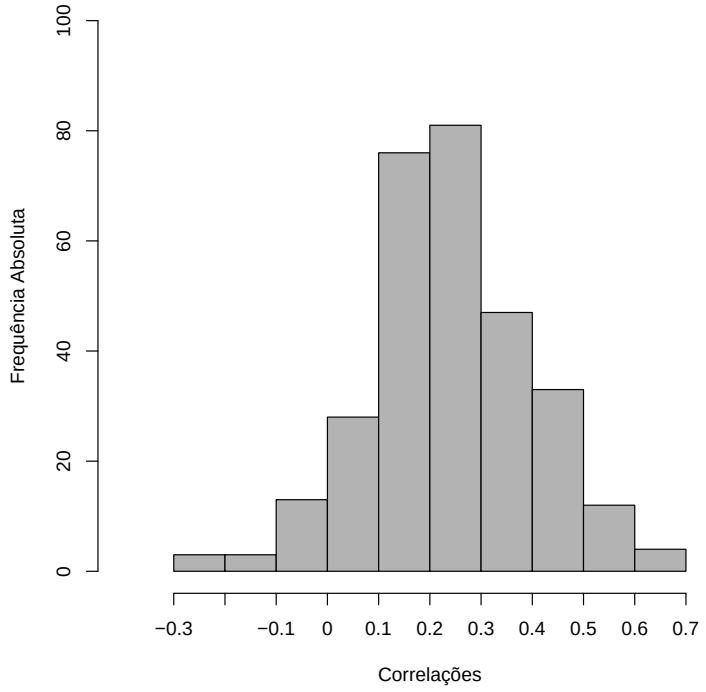

Figura 7: Histograma da distribuição de frequência das 300 correlações de Pearson, envolvendo todas as 25 disciplinas do banco de dados utilizado.

Conforme a Figura 7, fica evidente que a maioria das correlações apresentou magnitude abaixo de 0,30, ferindo a recomendação de FIELD *et al.* (2012), citados por MATOS e RODRIGUES (2019). Na realidade, apenas 96 dentre as 300 correlações foram superiores a 0,30 (ou seja, apenas 32% delas).

Este resultado sugeriu a necessidade de que algumas das variáveis (disciplinas) devessem ser retiradas do banco de dados, de maneira a aumentar a magnitude média das correlações bem como a possibilidade de um melhor ajustamento a um modelo fatorial. A decisão sobre quais variáveis poderiam ser retiradas do banco foi subsidiada pelo cálculo da medida KMO, discutida no tópico a seguir.

Medida de Kaiser-Meyer-Olkin (KMO)

Conforme descrito no Capítulo de Referencial Teórico, a medida KMO é uma medida de adequação da base de dados para uma análise fatorial. Conforme apontado, é possível calcular uma medida KMO global para o conjunto de dados, bem como uma medida KMO para cada uma das variáveis.

FIELD *et al.* (2012), citados por MATOS e RODRIGUES (2019), sugerem critérios empíricos para avaliar a magnitude das medidas KMO, os quais foram reproduzidos na Tabela 2. Segundo esta recomendação, valores entre 0,7 e 0,8 indicariam uma boa adequação da base de dados, e valores entre 0,8 e 0,9 uma adequação ótima. A Tabela 7 apresenta as medidas KMO para a presente base de dados, para cada uma das 25 disciplinas, bem como a medida KMO geral.

Tabela 7: Medidas KMO referentes a cada uma das $n = 25$ variáveis, bem como medida KMO global.[1]

MetCie: 0,59	GA: 0,75	Calc_I: 0,75	Calc_II: 0,90	Algebra: 0,77
Progr: 0,73	CalcNum: 0,73	Pacotes: 0,77	Estats: 0,86	Probs: 0,84
Inferencia: 0,87	Mult: 0,87	Regres: 0,74	ProcEst: 0,89	Demog: 0,76
MNP: 0,84	Amostr: 0,90	PlanExp: 0,85	CEQ: 0,88	Series: 0,88
ADC: 0,90	Sobrev: 0,59	POM: 0,40	Lab: 0,55	Mono: 0,61
Geral: 0,80				

[1] A codificação das variáveis (disciplinas) está detalhada na Tabela 4.

A julgar pelo valor da medida KMO geral apresentado na Tabela 7 (0.80), a adequação geral da base de dados seria de boa a ótima, conforme a recomendação de FIELD *et al.* (2012), citados por MATOS e RODRIGUES (2019). Algumas variáveis, contudo, apresentaram medidas KMO abaixo de 0,7, sugerindo, conforme a recomendação dos mesmos autores (ver Tabela 2), que estariam contribuindo pouco para a adequação da base de dados ao ajustamento de um modelo fatorial. Estas variáveis corresponderam ao rendimento acadêmico nas disciplinas Metodologia Científica, Análise de Sobrevivência, Pesquisa de Opinião e Mercado, Laboratório Supervisionado e Monografia. Desta forma, tais variáveis foram retiradas, passando o banco a contar com 20 variáveis (disciplinas).

A retirada destas 5 variáveis fez com que, dentre as 190 correlações remanescentes, 78 (41%) fossem superiores a 0,30. Ou seja, mesmo com esta retirada, a maioria das correlações continuou menor que 0,30, embora esta proporção tenha diminuído. Contudo, optou-se por não se fazer novas retiradas de variáveis neste momento, deixando-se para

fazê-las, eventualmente, quando da busca do ajustamento de um modelo fatorial adequado.

Com esta nova matriz de correlações, o teste de Bartlett continuou altamente significativo ($\chi^2 = 704,73$, valor-p $3,05 \times 10^{-60}$). As novas medidas KMO estiveram agora todas acima de 0,7 (Tabela 8).

Tabela 8: Medidas KMO referentes ao novo banco de dados, contendo $n = 20$ variáveis.[1]

GA: 0,75	Calc_I: 0,74	Calc_II: 0,90	Algebra: 0,83	Progr: 0,77
CalcNum: 0,82	Pacotes: 0,74	Estats: 0,83	Probs: 0,80	Inferencia: 0,87
Mult: 0,83	Regres: 0,88	ProcEst: 0,88	Demog: 0,77	MNP: 0,82
Amostr: 0,86	PlanExp: 0,85	CEQ: 0,89	Series: 0,91	ADC: 0,91
Geral: 0,84				

[1] A codificação das variáveis (disciplinas) está detalhada na Tabela 4.

Determinação do Número de Fatores

O pacote *psych* da linguagem **R**, utilizado aqui para a realização da análise fatorial, possui uma função denominada *scree()* que gera um gráfico apresentando tanto os autovalores da matriz de correlações original, bem como os autovalores de uma análise fatorial. Este gráfico está apresentado na Figura 8.

Avaliando a Figura 8, pelo critério de Kaiser deveriam ser retidos 5 componentes, em uma análise de componentes principais, e um único fator, em uma análise fatorial. No entanto, ao se tentar obter um número de fatores que corresponderia a um ponto de curvatura máxima nas "curvas" deste gráfico, percebe-se a dificuldade desta identificação, uma vez que (ao menos) entre os fatores 2 e 7 houve uma tendência quase linear de decréscimo da magnitude dos autovalores da análise fatorial. Esta dificuldade em se encontrar um ponto de curvatura máxima fez com que o gráfico *scree* não tenha elucidado claramente qual seria um número de fatores adequado, que aparentemente poderia ser algo entre 2 e 7. Em função disso, e levando-se em conta o critério de Kaiser para a análise de componentes principais, optou-se por trabalhar com um número inicial de 5 fatores, com a possibilidade de alterá-lo, em função da maior ou menor facilidade de interpretação dos fatores obtidos com o modelo ajustado.

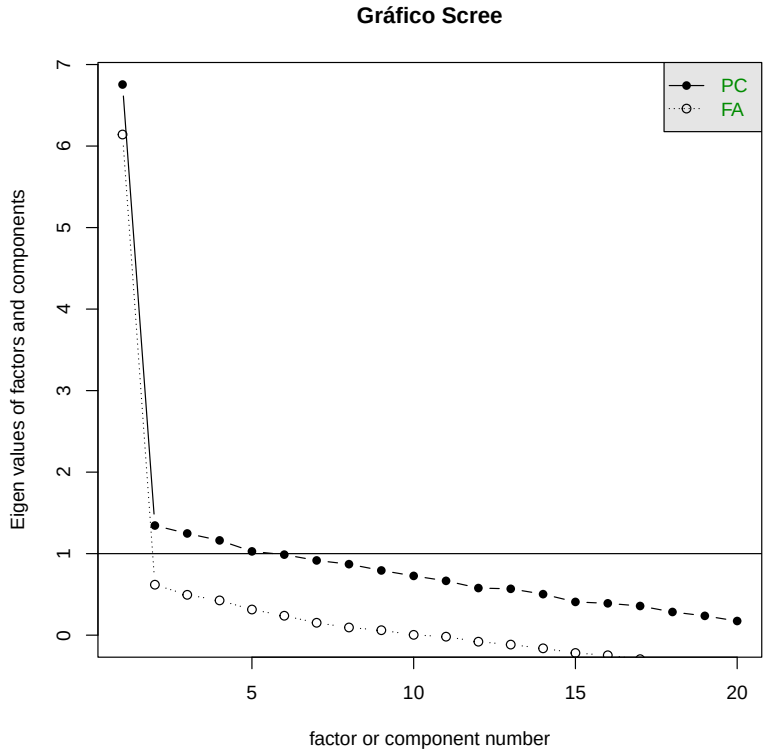

Figura 8: Gráfico scree apresentando os autovalores da matriz de correlações original, os quais estão associados a uma análise de componentes principais ("PC"), bem como os autovalores de uma análise fatorial ("FA").

Ajuste de Modelos Fatoriais

Finalmente, neste item são apresentados os resultados que nortearam a busca pelo ajustamento de um modelo fatorial adequado.

Um dos pontos de partida deste trabalho, e que motivou a escolha pela técnica da análise fatorial, foi a perspectiva de se encontrar alguns poucos eixos temáticos (os fatores) que pudessem refletir eixos de aprendizado no curso de Bacharelado em Estatística da UFOP, com a possibilidade de que pudessem eventualmente estar correlacionados, como seria de se esperar em eixos de aprendizado. Assim, uma das escolhas feitas antes das análises foi pelo uso de uma rotação oblíqua. Além disso, dado que as variáveis observadas eram tipicamente quantitativas contínuas, também foi feita a escolha pelo uso da máxima verossimilhança, admitindo distribuição normal para as variáveis.

No entanto, são apresentados aqui alguns resultados dos ajustes feitos sem rotação (utilizando tanto a máxima verossimilhança como o método dos fatores principais), e com rotação ortogonal (utilizando máxima verossimilhança e o método de rotação varimax). Estes resultados, bem como aqueles correspondentes do ajuste utilizando máxima veros-

similhança e rotação oblíqua (com o método oblimin), estão apresentados na Tabela 9. Conforme apontado no item anterior, em todos estes ajustes foram considerados modelos com 5 fatores.

Tabela 9: Resultados do ajuste de modelos fatoriais considerando $p = 20$ variáveis e 5 fatores.[1]

	MV Sem Rotação (QMR = 0,04148)				
Parâmetro	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
Variância Explicada Pelo Fator	4,95	2,02	0,81	0,69	0,68
Proporção da Variância Total	0,25	0,10	0,04	0,03	0,03
Proporção Acumulada da Variância Total	0,25	0,35	0,39	0,42	0,46
Proporção da Variância Explicada	0,54	0,22	0,09	0,08	0,07
Proporção Acumulada da Variância Explicada	0,54	0,76	0,85	0,93	1,00
	FP Sem Rotação (QMR = 0,03719)				
Parâmetro	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
Variância Explicada Pelo Fator	6,28	0,83	0,71	0,63	0,58
Proporção da Variância Total	0,31	0,04	0,04	0,03	0,03
Proporção Acumulada da Variância Total	0,31	0,36	0,39	0,42	0,45
Proporção da Variância Explicada	0,70	0,09	0,08	0,07	0,06
Proporção Acumulada da Variância Explicada	0,70	0,79	0,87	0,94	1,00
	MV Com Rotação Ortogonal (QMR = 0,04148)				
Parâmetro	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
Variância Explicada Pelo Fator	3,35	1,68	1,49	1,32	1,31
Proporção da Variância Total	0,17	0,08	0,07	0,07	0,07
Proporção Acumulada da Variância Total	0,17	0,25	0,33	0,39	0,46
Proporção da Variância Explicada	0,37	0,18	0,16	0,14	0,14
Proporção Acumulada da Variância Explicada	0,37	0,55	0,71	0,86	1,00
	MV Com Rotação Oblíqua (QMR = 0,04148)				
Parâmetro	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
Variância Explicada Pelo Fator	2,55	1,89	1,87	1,73	1,11
Proporção da Variância Total	0,13	0,09	0,09	0,09	0,06
Proporção Acumulada da Variância Total	0,13	0,22	0,32	0,40	0,46
Proporção da Variância Explicada	0,28	0,21	0,20	0,19	0,12
Proporção Acumulada da Variância Explicada	0,28	0,49	0,69	0,88	1,00

[1] MV: método da máxima verossimilhança; FP: método dos fatores principais; QMR: quadrado médio residual.

A Tabela 9 contém algumas informações globais interessantes, referentes ao ajuste destes modelos fatoriais. Observou-se aqui que o método dos fatores principais (FP) promoveu um ajuste com um quadrado médio residual (QMR) ligeiramente menor que aquele

da máxima verossimilhança (MV). Porém, com as três variações do método MV (sem rotação, com rotação ortogonal, e com rotação oblíqua), o valor do QMR não se alterou, o que decorre do fato de que a rotação (seja ortogonal ou oblíqua) não interfere no total da variância explicada pelos fatores, e consequentemente não altera a variação residual.

Nesta Tabela, a primeira linha de cada modelo se refere à variância explicada por cada fator. Estas grandezas correspondem à soma dos quadrados das cargas fatoriais, para cada fator, ao longo de todas as variáveis, e dá um indicativo da importância de cada fator. Considerando os ajustamentos sem rotação, observa-se que em ambos os métodos o primeiro fator desempenhou uma destacada importância. Com o método MV este parâmetro foi igual a 4,95, explicando 0,25 da variância total (Tabela 9). Com o método FP este parâmetro foi ainda maior (6,28), representando 0,31 da variância total. Este primeiro fator contabilizou 54% (0,54) da variância explicada pelo conjunto dos fatores com o método MV, e 70%, com o método FP. Ou seja, ao menos para este conjunto de dados, o método MV foi capaz, mesmo sem rotação, de distribuir um pouco mais a importância relativa de cada fator.

Consultando o último valor da proporção acumulada da variância total, verifica-se também que ambos os métodos produziram valores muito semelhantes (0,46 e 0,45 para os métodos MV e FP, respectivamente), sugerindo qualidades de ajuste semelhantes, apesar da pequena redução do QMR ocorrida com o método FP.

Independentemente desta semelhança, chama a atenção estes baixos valores de proporção acumulada da variância total, inferiores a 0,5. Em todas as tentativas realizadas neste estudo, isto foi uma tendência geral. Isto aponta para um primeiro fato relevante no estudo do fenômeno em questão, avaliado pelo conjunto das notas de aprovação nas diferentes disciplinas. Provavelmente se trata de um fenômeno com variabilidade muito elevada, com alto grau de variação casual. Pois a atribuição de notas é função de uma série de fatores até certo ponto casuais, como diferentes professores ministrando a mesma disciplina, e diferentes turmas de alunos, com diferentes perfis e trajetórias. Isto é importante para evidenciar as limitações do eventual modelo fatorial ajustado, mas que, por outro lado, o modelo permitiria descrever a estrutura da associação entre as variáveis, quando esta associação está presente, contribuindo para a compreensão do fenômeno.

Com a utilização de rotação dos fatores, espera-se que a importância relativa dos fatores tenda a ser mais distribuída, o que de fato se pode constatar nesta mesma Tabela 9, comparando-se os resultados das três variações com o método MV. A rotação ortogonal, em relação ao modelo sem rotação, promoveu uma sensível redução das variâncias explica-

das pelos fatores 1 e 2, aumentando-se aquelas dos demais. Os fatores 1 e 2 contabilizaram 54% e 22% no ajuste sem rotação, ao passo que com rotação ortogonal estas proporções reduziram para 37% e 18%, respectivamente. Os demais fatores, por sua vez, aumentaram estas proporções para acima de 10%.

Com a rotação oblíqua, observou-se uma redução expressiva da proporção da variância explicada do primeiro fator, passando de 37%, com a rotação ortogonal, para 28%. Isto se refletiu no aumento desta proporção para os fatores 2, 3 e 4, ilustrando como a rotação oblíqua tem o potencial de identificar diferentes eixos temáticos com importância mais distribuída, o que pode permitir uma melhor compreensão do fenômeno estudado.

Em seguida, investigou-se mais detalhadamente a qualidade do ajuste do modelo utilizando o método MV e rotação oblíqua, em particular sobre a importância das diferentes variáveis no ajustamento do modelo fatorial. Na Tabela 10 são apresentadas as estimativas das cargas fatoriais de cada variável sobre cada fator, bem como a comunalidade e a especificidade (variância específica) de cada uma.

Nas duas últimas colunas da Tabela 10 são apresentados os valores de comunalidade e de variância específica, conforme definidos em 2.17 e 2.18. Para cada variável, a soma destas duas grandezas é igual a 1, que corresponde à variância total, lembrando que as variáveis são padronizadas. Desta forma, a comunalidade h^2 pode ser diretamente interpretada como uma proporção da variância de uma variável que é compartilhada com outras variáveis. Se uma variável apresenta comunalidade baixa, isto seria um indicativo de que estaria contribuindo pouco para a qualidade do ajustamento do modelo fatorial.

Em particular, chama a atenção as comunalidades baixas para os Cálculos I e II, sugerindo a possibilidade de estas variáveis serem retiradas do modelo. Se considerarmos também como baixas aquelas comunalidades próximas de 0,2, também seriam candidatas à remoção as disciplinas de Álgebra Linear e Cálculo Numérico. Em um primeiro momento buscou-se retirar apenas as disciplinas de Cálculo (resultados não apresentados aqui), e observou-se que as comunalidades de Álgebra Linear e Cálculo Numérico pouco alteraram seus valores, permanecendo próximas de 0,20. Desta maneira, optou-se pela remoção de todas estas 4 variáveis com menores comunalidades, na expectativa de uma melhoria no ajustamento do modelo fatorial, e assim tentar caracterizar de maneira mais clara a natureza da associação, naquele grupo de variáveis que apresentam tal associação.

Em relação às cargas fatoriais apresentadas na Tabela 10, elas representam os coeficientes que multiplicam cada fator, em uma dada variável. Podem ser interpretadas como correlações (ou, mais rigorosamente, como grandezas análogas a coeficientes de regressão

Tabela 10: Cargas fatoriais, comunalidade (h^2) e especificidade (ψ) de cada variável, no ajustamento de um modelo fatorial com $p = 20$ variáveis e 5 fatores F1, F2, ... F5, utilizando máxima verossimilhança e rotação oblíqua.[1]

Variável	F1	F2	F3	F4	F5	h^2	ψ
GA	0,33	-0,02	-0,17	0,23	0,25	0,279	0,721
Calc_I	0,07	-0,03	0,18	0,01	0,18	0,098	0,902
Calc_II	0,10	0,22	0,07	0,06	0,07	0,147	0,853
Algebra	0,19	0,21	0,12	0,03	0,04	0,202	0,798
Progr	0,04	0,06	-0,08	0,10	**0,59**	0,412	0,588
CalcNum	0,03	-0,02	0,10	0,28	0,25	0,225	0,775
Pacotes	-0,13	**0,40**	0,17	-0,09	0,35	0,354	0,646
Estats	0,00	0,28	0,23	0,32	0,34	0,659	0,341
Probs	0,01	-0,01	0,00	**0,99**	0,01	0,995	0,005
Inferencia	**0,65**	0,18	-0,05	0,13	-0,03	0,627	0,373
Mult	-0,14	0,22	**0,53**	0,17	-0,16	0,410	0,590
Regres	0,36	0,22	0,08	0,12	-0,11	0,336	0,664
ProcEst	**0,39**	0,09	0,07	0,16	0,09	0,354	0,646
Demog	-0,07	-0,10	**0,51**	0,07	0,13	0,265	0,735
MNP	0,06	**0,97**	-0,01	0,00	0,01	0,995	0,005
Amostr	**0,58**	0,05	0,14	-0,11	0,20	0,516	0,484
PlanExp	**0,39**	-0,06	0,34	-0,05	0,20	0,420	0,580
CEQ	0,24	0,04	**0,69**	-0,01	0,02	0,721	0,279
Series	**0,51**	0,18	0,11	0,06	-0,11	0,469	0,531
ADC	**0,49**	0,03	0,30	0,20	-0,02	0,653	0,347

[1] A codificação das variáveis (disciplinas) está detalhada na Tabela 4. Valores em negrito correspondem a cargas próximas ou acima de 0,4.

múltipla, devido à rotação oblíqua) entre cada variável e um dado fator. A magnitude das cargas fatoriais (em valor absoluto) é que nos possibilita indicar a que fator uma dada variável está mais fortemente vinculada, auxiliando na interpretação do significado de cada fator.

Conforme visto no Referencial Teórico, HAIR et al. (2009) apresentam critérios para a identificação de cargas fatoriais relevantes, conforme o tamanho da amostra. Estes critérios foram reproduzidos na Tabela 1. Segundo estes critérios, para o tamanho da amostra do presente estudo seriam dignas de atenção cargas fatoriais próximas a 0,5 ou superiores. Contudo, dado que a proposta deste estudo consistiu na obtenção de um modelo fatorial que possa explicar ao menos parte do fenômeno, apesar do alto nível de variação casual, optou-se por se considerar magnitudes próximas a 0,4 ou superiores como dignas de atenção, com a expectativa de facilitar a interpretação dos fatores. Os valores que atendessem a este critério foram destacados em negrito na Tabela 10.

Embora ainda não estejamos fazendo, neste momento, uma interpretação definitiva dos fatores, é conveniente destacar alguns aspectos já manifestados aqui, em relação aos fatores, aspectos esses elencados a seguir.

O fator 1 reúne disciplinas que poderiam ser caracterizadas como de caráter teórico-aplicado, reunindo Inferência, Processos Estocásticos, Amostragem, Séries Temporais e Análise de Dados Categóricos.

O fator 2 é uma dimensão com a qual a disciplina de Métodos Não-Paramétricos apresenta correlação muito elevada, sugerindo que esta disciplina teria características que a colocariam em uma dimensão à parte, e possivelmente associada à disciplinas de Pacotes Estatísticos. Chama a atenção também a altíssima comunalidade desta disciplina, reforçando sua importância no ajuste do modelo fatorial.

O fator 3 reúne a disciplina de Análise Multivariada com as disciplinas de caráter mais aplicado Demografia e Controle Estatístico de Qualidade.

O fator 4 chama a atenção por ser um fator com importância destacada para as disciplinas de Probabilidade, caracterizando, assim, um eixo de natureza mais teórica.

Finalmente, o fator 5 corresponderia a um eixo de natureza mais computacional, dada a predominância da importância da disciplina de Programação de Computadores.

A Tabela 11 apresenta os resultados do ajustamento do modelo fatorial realizado a seguir, ou seja, após a remoção das disciplinas de Cálculo, Álgebra Linear e Cálculo Numérico, resultando em um conjunto com 16 variáveis.

Comparando-se as Tabelas 10 (20 variáveis) e 11 (16 variáveis), verifica-se, em linhas gerais, uma tendência de manutenção da natureza dos fatores obtidos com a análise fatorial utilizando 20 variáveis, com uma pequena mudança quando ao ordenamento dos fatores. Na primeira análise (Tabela 10), o fator reunindo a disciplina de Análise Multivariada com disciplinas de caráter aplicado correspondeu ao terceiro fator quanto à variância explicada, tendo sido listado como fator 3 na Tabela 10. Na análise com 16 variáveis (Tabela 11), este passou a ser o segundo fator. Esta "troca" de ordenamento foi feita com o fator agrupando Métodos Não Paramétricos e as disciplinas de Pacotes Estatísticos.

O primeiro fator continuou agrupando disciplinas que poderiam ser caracterizadas como teórico-aplicadas. E os fatores 4 e 5 se mantiveram como fatores tendo as disciplinas de Probabilidade e de Programação, respectivamente, como eixos principais.

Finalmente, buscou-se realizar uma última análise fatorial, reduzindo o número de

Tabela 11: Cargas fatoriais, comunalidade (h^2) e especificidade (ψ) de cada variável, no ajustamento de um modelo fatorial com $p = 16$ variáveis e 5 fatores F1, F2, ... F5, utilizando máxima verossimilhança e rotação oblíqua.[1]

Variável	F1	F2	F3	F4	F5	h^2	ψ
GA	0,32	-0,15	-0,02	0,23	0,30	0,29	0,706
Progr	0,01	-0,05	0,04	0,08	**0,65**	0,47	0,533
Pacotes	-0,16	0,18	**0,40**	-0,09	0,33	0,35	0,653
Estats	-0,02	0,24	0,28	0,32	0,31	0,64	0,360
Probs	0,01	0,00	-0,02	**1,00**	0,02	1,00	0,005
Inferencia	**0,62**	-0,01	0,20	0,14	-0,02	0,63	0,373
Mult	-0,17	**0,50**	0,22	0,18	-0,16	0,38	0,615
Regres	0,32	0,11	0,23	0,12	-0,10	0,33	0,667
ProcEst	0,35	0,11	0,09	0,16	0,09	0,35	0,645
Demog	-0,11	**0,52**	-0,11	0,07	0,12	0,26	0,741
MNP	0,06	-0,02	**0,98**	0,00	0,01	1,00	0,005
Amostr	**0,54**	0,21	0,05	-0,10	0,18	0,52	0,481
PlanExp	0,33	**0,40**	-0,06	-0,05	0,21	0,43	0,569
CEQ	0,16	**0,77**	0,03	0,00	0,00	0,76	0,243
Series	**0,47**	0,14	0,19	0,07	-0,07	0,46	0,536
ADC	**0,43**	0,34	0,05	0,22	-0,02	0,64	0,364

[1] A codificação das variáveis (disciplinas) está detalhada na Tabela 4. Valores em negrito correspondem a cargas próximas ou acima de 0,4.

fatores de 5 para 4, na expectativa de que essa nova configuração pudesse contribuir para uma melhor interpretação dos eixos temáticos manifestados através dos fatores. Os resultados desta última análise com 4 fatores estão apresentados nas Tabelas 12 e 13.

Tabela 12: Resultados do ajuste de um modelo fatorial considerando $p = 16$ variáveis e 4 fatores.

Parâmetro	Fator 1	Fator 2	Fator 3	Fator 4
Variância Explicada Pelo Fator	2,77	1,87	1,55	1,54
Proporção da Variância Total	0,17	0,12	0,10	0,10
Proporção Acumulada da Variância Total	0,17	0,29	0,39	0,48
Proporção da Variância Explicada	0,36	0,24	0,20	0,20
Proporção Acumulada da Variância Explicada	0,36	0,60	0,80	1,00

Conforme a Tabela 12, verifica-se que a proporção acumulada da variância total explicada pelos 4 fatores foi igual a 0,48, um valor um pouco maior do que o do ajustamento considerando 20 variáveis, e 5 fatores com rotação (0,46, Tabela 9). Também pode-se observar que a contribuição de cada fator para a proporção da variância explicada foi relativamente distribuída, variando de 0,20 (fatores 3 e 4) a 0,36 (fator 1).

A Tabela 13 apresenta as cargas fatoriais, comunalidade e especificidade de cada variável, no ajustamento do modelo fatorial com $p = 16$ variáveis e 4 fatores.

Tabela 13: Cargas fatoriais, comunalidade (h^2) e especificidade (ψ) de cada variável, no ajustamento de um modelo fatorial com $p = 16$ variáveis e 4 fatores F1, F2, ... F4, utilizando máxima verossimilhança e rotação oblíqua.[1]

Variável	F1	F2	F3	F4	h^2	ψ	Hofmann
GA	0,33	-0,17	0,13	0,22	0,22	0,781	2,7
Progr	0,03	-0,10	**0,47**	0,11	0,24	0,756	1,2
Pacotes	-0,08	0,07	**0,68**	-0,07	0,43	0,573	1,1
Estats	0,06	0,18	**0,44**	0,34	0,62	0,381	2,3
Probs	0,02	0,01	-0,01	**0,99**	1,00	0,005	1,0
Inferencia	**0,78**	-0,04	0,00	0,08	0,64	0,358	1,0
Mult	-0,10	**0,45**	0,11	0,18	0,30	0,702	1,5
Regres	**0,41**	0,13	0,03	0,09	0,31	0,689	1,3
ProcEst	**0,42**	0,04	0,16	0,12	0,37	0,629	1,5
Demog	-0,18	**0,49**	0,07	0,10	0,23	0,771	1,4
MNP	0,38	0,00	**0,51**	0,03	0,62	0,384	1,9
Amostr	**0,56**	0,19	0,11	-0,13	0,47	0,529	1,4
PlanExp	0,30	0,36	0,10	-0,06	0,38	0,624	2,2
CEQ	0,10	**0,84**	0,01	-0,01	0,82	0,184	1,0
Series	**0,60**	0,11	-0,01	0,02	0,46	0,539	1,1
ADC	**0,49**	0,30	-0,03	0,18	0,63	0,367	2,0

[1] A codificação das variáveis (disciplinas) está detalhada na Tabela 4. Valores em negrito correspondem a cargas próximas ou acima de 0,4.

Esta nova configuração em 4 fatores, de fato, aparentemente propiciou uma simplificação de interpretação, e a discussão deste ajuste é feita a seguir.

Conforme se observa na Tabela 13, a maior alteração percebida (comparando-se as Tabelas 11 e 13) foi a formação de um único fator (fator 3) contendo a disciplina de Métodos Não Paramétricos e as disciplinas de Programação e de Pacotes Estatísticos. Neste mesmo fator, considerando ainda uma magnitude de carga fatorial próxima ou superior a 0,4 como uma referência, observou-se a inclusão das disciplinas de Estatísticas neste fator. Ou seja, este fator parece apontar para uma caracterização que reúne tanto

aspectos de disciplinas mais aplicadas quanto aspectos de natureza computacional.

Os demais fatores apresentaram uma interpretação semelhante à dos ajustes anteriores. Sendo assim, uma caracterização possível para a interpretação dos 4 fatores deste último ajustamento poderia ser resumida como segue.

Fator 1: eixo teórico-aplicado. Neste fator são reunidas as disciplinas de Inferência, Regressão, Processos Estocásticos, Amostragem, Séries Temporais e Análise de Dados Categóricos. Em relação às disciplinas de Regressão e Processos Estocásticos houve um sensível aumento de suas cargas fatoriais, quando comparadas com aquelas da Tabela 11, estando agora acima de 0,40.

Fator 2: eixo aplicado. Embora a disciplina de Análise Multivariada pudesse em princípio ser caracterizada como uma disciplina teórico-aplicada, por sua potencial utilização em diferentes áreas, ocorre que a maioria das demais disciplinas da grade fazem uso de análises univariadas, e assim este caráter mais teórico não estaria se manifestando em outras disciplinas com frequência. Pode-se dizer que isto lhe conferiria um caráter aplicado, no sentido de apresentar técnicas de análise em um campo específico de aplicação. Assim, não havendo controvérsia em se caracterizar Análise Multivariada como uma disciplina de caráter mais aplicado, tem-se aqui a interpretação proposta para este fator 2 como a de um eixo aplicado. A disciplina de Planejamento de Experimento apresentou a sua maior carga fatorial neste fator (embora inferior a 0,40), corroborando esta caracterização.

Fator 3: eixo computacional-aplicado. Este é o fator que parece reunir com clareza a dimensão computacional do fenômeno, sendo que esta dimensão tem claramente uma natureza aplicada, pela presença das disciplinas de Estatísticas e de Métodos Não Paramétricos. De fato, estas disciplinas são as primeiras disciplinas mais aplicadas que os alunos cursam, que fazem uso de recursos computacionais, sendo que estas disciplinas são cursadas em semestres iguais ou próximos aos das disciplinas de Programação e de Pacotes.

Fator 4: eixo teórico. Aqui manteve-se a clara tendência de as disciplinas de Probabilidade representarem um eixo temático à parte, correspondendo a uma dimensão teórica fundamental no pensamento estatístico, que é a Teoria de Probabilidades. Chama a atenção a altíssima comunalidade deste grupo de disciplinas (0,995, arredondada para 1,00 na Tabela 13).

Em relação à disciplina de Geometria Analítica, não apresentou nenhuma carga fatorial acima de 0,40 (Tabela 13). Além disso, apresentou uma baixa comunalidade, igual a

0,22. Apesar disto, optou-se por mantê-la neste ajuste fatorial, em virtude de ter apresentado comunalidades maiores nos ajustes anteriores, próximas de 0,30, e por ser a única disciplina da área de Matemática remanescente no modelo. Esta disciplina apresentou maior carga fatorial para o fator 1 (0,33).

Em relação às Tabelas anteriores, a Tabela 13 apresenta uma coluna adicional, contendo os índices de complexidade de Hofmann. Estes índices seriam uma estimativa do número de fatores aos quais uma determinada variável pertence. Em uma análise fatorial ideal, todas as variáveis do modelo teriam índices próximos de 1. Porém, considerando o conjunto de 16 variáveis que foram retidas no modelo, algumas apresentaram valores sensivelmente maiores que 1, o que poderiam apontar para um caráter multifatorial destas variáveis, para o modelo fatorial ajustado em questão.

O maior índice de Hofmann observado foi para a disciplina Geometria Analítica (2,7), sugerindo que pudesse estar vinculada a até 3 destes fatores. A julgar pelas magnitudes das cargas fatoriais, este fatores seriam os fatores 1, 4 e 2, evidenciando que a disciplina teria aspectos teóricos, aplicados e teórico-aplicados.

A disciplina de Planejamento de Experimento também não apresentou nenhuma carga fatorial acima de 0,40, mas também optou-se por mantê-la no modelo, dado o valor moderado de sua comunalidade (0,38). Baseando-se no índice de Hofmann, esta disciplina seria bifatorial, compondo os fatores 1 e 2, com os quais a disciplina apresentou as maiores cargas fatoriais (0,30 e 0,36).

Outras disciplinas que apresentaram índices próximos a 2, sugerindo uma natureza bifatorial, foram as disciplinas de Estatísticas (expressando-se nos fatores 3 e 4), Métodos Não Paramétricos (fatores 1 e 3) e Análise de Dados Categóricos (fatores 1 e 2).

Conforme apontado no Referencial Teórico, uma das grandes vantagens da análise fatorial com rotação oblíqua é a possibilidade de identificação de fatores que sejam correlacionados entre si, o que eventualmente pode auxiliar sua interpretação. Assim, houve aqui o interesse em se investigar se os 4 fatores obtidos com o ajustamento do último modelo fatorial estariam de fato correlacionados, e, em caso positivo, a magnitude destas correlações.

Assim, embora no presente estudo não tenha havido particular interesse na obtenção dos escores de cada bacharéu para cada um dos fatores, tais escores foram obtidos apenas para o cálculo das correlações entre os 4 fatores. Estas correlações estão apresentadas na Tabela 14.

Observa-se na Tabela 14 que as correlações entre os 4 fatores poderiam ser consideradas como moderadas, sendo que a maior correlação correspondeu àquela entre o fator 1 (eixo teórico-aplicado) com o fator 2 (eixo aplicado).

Tabela 14: Coeficientes de correlação de Pearson entre os escores dos fatores de uma análise fatorial com 16 variáveis e 4 fatores.

	Fator 1	Fator 2	Fator 3	Fator 4
Fator 1	1,00	0,58	0,47	0,48
Fator 2	0,58	1,00	0,46	0,35
Fator 3	0,47	0,46	1,00	0,36
Fator 4	0,48	0,35	0,36	1,00

5 Considerações Finais

Neste trabalho, foi utilizado um conjunto de variáveis correspondentes às notas de aprovação nas diferentes disciplinas do curso de Bacharelado em Estatística da UFOP. Estas variáveis, após uma caracterização descritiva, foram submetidas a uma análise fatorial, com a expectativa de que pudessem ser detectados fatores que correspondessem a eixos de aprendizado, no processo de formação dos bacharéis do curso.

Evidentemente, a natureza das variáveis (notas de aprovação) propiciaria apenas uma descrição parcial do processo de aprendizado, uma vez que este é um fenômeno complexo, e naturalmente deve se manifestar apenas parcialmente nas notas de aprovação das diferentes disciplinas do curso. Além disso, a baixa (ou moderada) proporção da variação explicada pelos fatores aqui observada, considerando diferentes modelos fatoriais, sugere que o fenômeno apresenta uma grande variabilidade residual, ou casual, o que faz sentido, uma vez que a avaliação do rendimento acadêmico está sujeita à influência de um grande número de variáveis não controladas, ou não previstas pelos modelos aqui utilizados. Basta citar que uma mesma disciplina por vezes é ministrada por diferentes docentes, além de ser ofertada para diferentes turmas, com diferentes perfis e trajetórias de aprendizado.

Todos estes elementos não controlados devem ter contribuído para a elevada variação residual, tendo por referência os modelos fatoriais aqui utilizados. Em função desta alta variabilidade residual, optou-se por considerar como cargas fatoriais dignas de atenção aquelas próximas ou superiores a 0,40. Com este procedimento, houve a expectativa de se tentar identificar alguma estrutura que descrevesse (ao menos em parte) a variação presente nos dados, com ciência da existência desta elevada variação residual ou casual. Esta parte da variação explicada pelo modelo corresponderia à fração da variação relacionada à associação existente (ainda que não elevada) entre as variáveis. Com isso, foi possível identificar quatro eixos de aprendizado, que aqui foram caracterizados como: teórico-aplicado, aplicado, computacional-aplicado, e finalmente um eixo teórico.

Assim, embora o presente conjunto de dados não tenha apresentado um grau de as-

sociação entre as variáveis que o classificasse como um conjunto apropriado (conforme a tendência da literatura) para um bom ajustamento de modelos fatoriais, esta metodologia permitiu uma melhor compreensão do fenômeno em questão, pela identificação de fatores apresentando uma interpretação coerente com a de eixos de aprendizado, ao menos aproximadamente.

Portanto, o presente estudo de caso ilustra o potencial da técnica de análise fatorial, mesmo quando o nível de associação entre as variáveis não é muito elevado. Também é nossa expectativa de que o presente material possa contribuir enquanto referência de consulta, ilustrando os procedimentos e as etapas de uma análise fatorial.

6 Referências Bibliográficas

HOFMANN, R. J. Complexity and simplicity as objective indices of factor solutions. **Multivariate Behavioral Research**: n.13, p.247-250, 1978.

JOHNSON, A.R;WICHERN, W.D. **Applied Multivariate Statistical Analysis**: 6.ed. New Jersey.Editora Upper Saddle River, n.755,2007.

MATOS, D.A.S.; RODRIGUES, E.C. **Análise fatorial**. Brasília: Enap, 2019. 74 p.

HAIR, J.F.Jr.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E.; TATHAM, R.L. **Análise Multivariada de Dados**. Porto Alegre, Bookman, 2009. 688p. 6.ed.

HÄRDLE, W.K.; SIMAR, L. **Applied Multivariate Statistical Analysis**. Springer, 2019. 558p.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`. 2024.

REVELLE, W. **psych: Procedures for Psychological, Psychometric, and Personality Research**. Northwestern University, Evanston, Illinois. R package version 2.4.12, 2024. URL `https://CRAN.R-project.org/package=psych`.

UNIVERSIDADE FEDERAL DE OURO PRETO. **Projeto Pedagógico do Curso de de Bacharelado em Estatística da Universidade Federal de Ouro Preto**. Ouro Preto, 2017, 74p. URL `https://sites.ufop.br/coest/ppc`.

WEI, T. e SIMKO, V. **R package 'corrplot'**: Visualization of a Correlation Matrix (Version 0.95). 2024. URL `https://github.com/taiyun/corrplot`.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag, New York, 2016.