UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
BACHARELADO EM ESTATÍSTICA

Trabalho de Conclusão de Curso

ANÁLISE FATORIAL E UMA APLICAÇÃO EM PERFIL DE COMPRAS DE PEQUENOS VAREJISTAS

Marcia Terazzi Basso

Uberlândia
2016

Marcia Terazzi Basso

ANÁLISE FATORIAL E UMA APLICAÇÃO EM PERFIL DE COMPRAS DE PEQUENOS VAREJISTAS

Trabalho de conclusão de curso de graduação apresentado a Faculdade de Matemática da Universidade Federal de Uberlândia como requisito parcial para a obtenção do título de Bacharel em Estatística.

Orientadora: Patrícia Viana da Silva

Uberlândia
2016

Marcia Terazzi Basso

ANÁLISE FATORIAL E UMA APLICAÇÃO EM PERFIL DE COMPRAS DE PEQUENOS VAREJISTAS

Trabalho de conclusão de curso de graduação apresentado à Faculdade de Matemática da Universidade Federal de Uberlândia como requisito parcial para a obtenção do título de Bacharel em Estatística.

Aprovado em: ____ de _____ de _____.

BANCA EXAMINADORA

Prof. Dr. Tiago Moreira Vargas − Universidade Federal de Uberlândia

---

Prof. Dr. Rodrigo Lambert − Universidade Federal de Uberlândia

---

Profa. Patrícia Viana da Silva (Orientadora) − Universidade Federal de Uberlândia

Agradecimentos

Obrigada Patrícia Viana da Silva pela orientação nessa jornada tão difícil. Pela paciência de me ensinar e explicar dos conceitos mais simples até os mais complexos. Por ser tão gentil, alegre e divertida em um momento tão estressante como esse. Obrigada por ser a minha orientadora, e espero que você não me esqueça pois eu não irei esquecê-la.

Agradeço a minha família pelo o apoio por todos esses anos. Por ter acreditado nas minhas escolhas e estar sempre a meu lado.

Agradeço a todos os Professores e Coordenadores da Estatística por essa jornada na minha carreira. Vocês tornaram essa experiência possível e agradeço de coração a perseverança e a sua escolha de ser um educador em momentos tão difíceis.

Obrigada ao meu noivo Marc por ser a pessoa mais linda desse mundo. Não tenho nem palavras para descrever o que você fez por nós, então meu sincero obrigado.

E finalmente, agradeço a todos os meus amigos pelas risadas, tristezas pós provas e alegrias pós semestre. Sem vocês eu não chegaria até aqui.

Obrigada!

Resumo

Este trabalho consiste na aplicação da técnica de Análise Fatorial, com a finalidade de redução da dimensionalidade dos dados. Os dados utilizados para a aplicação da técnica descrevem o comportamento de compra de varejistas. Os dados consistem em doze variáveis qualitativas utilizadas no estudo. Com o cálculo da matriz correlação, as medidas de adequação global de KMO e MSA confirmam a adequação dos dados que permite prosseguir com a análise. Três métodos foram utilizados para a extração de fatores: o critérios de Kaiser, a porcentagem da variância explicada e o *Scree-plot*. Depois do cálculo dos quatro fatores, a rotação Varimax foi utilizada para melhorar na interpretação das cargas fatoriais. Concluindo que o primeiro fator demonstra as características e valores do pedido do varejista, o segundo fator com a capacidade e volume de compra, o terceiro fator representa a quantidade e o valor das devoluções e o quarto fator relacionado com a fidelidade do cliente. O modelo apresentou um bom ajuste aos dados segundo a avaliação das comunalidades e conseguiu explicar 85,537% da variância total.

Palavras-chave: Análise multivariada; Análise Componentes Principais; Análise Fatorial; Perfil de Compra de Varejistas.

Sumário

1 Introdução

O desenvolvimento da análise multivariada e principalmente a sua utilização foram limitados durante anos devido à complexidade dos cálculos envolvidos. Porém, com o aumento da capacidade de processamento de dados pelos computadores e a grande quantidade de informações disponibilizadas, o uso e o interesse por esses métodos foram renovados e retomados [13]. A utilização desses métodos ganhou importância atualmente, inclusive para grandes empresas, devido à necessidade de lidar com essa gama de informações e para solucionar os problemas relacionados.

Kendall [10] separa as técnicas de análise multivariada pelos seus dois principais objetivos: análise de dependência e análise de interdependência. Temos a primeira quando estudamos a dependência de uma ou mais variáveis em relação às outras, como exemplo à Análise de Variância Multivariada. Por outro lado, as técnicas de Análise Fatorial e de Componentes Principais se enquadram na análise de interdependência. Esta análise é realizada quando há interesse nas relações de um conjunto de variáveis entre si.

Considerada como a técnica mais antiga da história da Análise Multivariada a Análise Fatorial foi proposta por Charles Spearman [18] e por Karl Pearson [14, 15, 16] no início do século XX. Anos depois, Hotelling [7] foi responsável por propor o uso de Componentes Principais para a construção de uma matriz de fatores ortogonais.

A Análise Fatorial propõe resumir a estrutura de interrelações de um grande número de variáveis. A redução de dimensionalidade dos se deve através de novas variáveis latentes conhecidas como fatores. Os fatores podem se caracterizar de duas maneiras: ortogonais ou oblíquos. Um fator ortogonal é considerado quando os fatores não são correlacionados entre si. O oposto é chamado de fatores oblíquos que são considerados correlacionados.

A construção do modelo da Análise Fatorial consiste em dois métodos para o cálculo dos fatores: o método da Máxima Verossimilhança e o método dos Componentes Principais. Com o primeiro método é possível testar hipóteses e criar intervalos de confiança para grandes amostras. Entretanto, esse método exige normalidade multivariada dos dados. Motivo que, o método dos Componentes Principais é mais utilizado [2], pois, não há pressuposição das variáveis envolvidas.

Na Análise de Componentes Principais, os fatores podem ser encontrados a partir da matriz de covariâncias ou da matriz de correlação das variáveis originais. Uma vantagem desse método é a melhora na explicação da estrutura dos dados. É possível identificar subgrupos de variáveis que, pelo seu alto grau de interdependência, podem ser considerados como um só construto, ou seja, um conceito abstrato que explica o que muitas vezes não pode ser visto através de uma única variável [2].

A aplicação usada nesse trabalho consiste em analisar o comportamento de compra dos clientes de uma empresa atacadista e assim procurar soluções plausíveis para a sua fidelização. Como a conquista de novos clientes é entre cinco e sete vezes mais cara que manter aqueles já conquistados [17] a fidelização é muito vantajosa para a empresa. A Análise Fatorial, nesse caso, tem como foco analisar as variáveis utilizadas e identificar o perfil de compra da base de clientes.

2 Metodologia

Neste estudo utilizaremos a Análise Fatorial de forma exploratória analisando a relação entre as variáveis e identificando padrões de correlação. Consequentemente, isso vai condensar as informações em um número menor de novas dimensões de fatores.

A Análise Fatorial confirmatória, é um procedimento que se baseia em em testar hipóteses sobre a adequação de um modelo proposto para a estrutura de correlação de um conjunto de dados. Entretanto para esse modelo não será abordado nesse trabalho.

Barroso [2] define a Análise Fatorial como uma técnica de transformação linear de um conjunto de p variáveis em um conjunto menor de variáveis definidos como fatores, que explica uma parcela razoável da variabilidade total dos dados originais.

2.1 Modelo Teórico

Seja $\mathbf{X}$ a matriz de dados originais, onde temos p variáveis e n observações de uma população. Na qual,são representadas pelas variáveis $X_1, X_2, \ldots, X_p$. Formando assim uma matriz de ordem $n \times p$ dada a seguir:

$$\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_p}] = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

Duas matrizes muito importante são a base para poder entender a teoria da Análise Fatorial, a matriz de correlação e a matriz de covariância. Seja que a matriz de correlação é uma matriz quadrada cujos elementos são as correlações entre as variáveis analisadas. A matriz de covariância é uma matriz quadrada cujos elementos fora da diagonal principal são as covariâncias entre as variáveis e na diagonal principal são as variâncias de cada variável.

O modelo supõe que cada variável $\mathbf{X}_j$ é linearmente dependente de variáveis aleatórias não observáveis $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_m$, $(m < p)$ chamadas fatores comuns, $E(\mathbf{X}) = \boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_p)^T$ o vetor de médias de $\mathbf{X}$ e p fontes adicionais de variação $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_p$ chamadas erros ou, algumas vezes, fatores específicos [8].

Um modelo de Análise Fatorial ortogonal é dado por:

$$\begin{aligned} \mathbf{X}_1 - \boldsymbol{\mu}_1 &= \phi_{11}\mathbf{F}_1 + \ldots + \phi_{1m}\mathbf{F}_m + \boldsymbol{\xi}_1 \\ \mathbf{X}_2 - \boldsymbol{\mu}_2 &= \phi_{21}\mathbf{F}_1 + \ldots + \phi_{2m}\mathbf{F}_m + \boldsymbol{\xi}_2 \\ &\vdots \\ \mathbf{X}_p - \boldsymbol{\mu}_p &= \phi_{p1}\mathbf{F}_1 + \ldots + \phi_{pm}\mathbf{F}_m + \boldsymbol{\xi}_p \end{aligned}$$

ou seja,

$$\mathbf{X}_j - \boldsymbol{\mu}_j = \phi_{j1}\mathbf{F}_1 + \ldots + \phi_{jm}\mathbf{F}_m + \boldsymbol{\xi}_j \tag{1}$$

em que $\mathbf{X}_j$ é a j-ésima variável, $\phi_{j1}, \phi_{j2}, \ldots, \phi_{jm}$ são as cargas dos fatores para a j-ésima variável e $\mathbf{F}_1, \ldots, \mathbf{F}_m$ são m fatores comuns não correlacionados, com $m < p$.

A equação (1) é representada a seguir matricialmente:

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{\Phi F} - \boldsymbol{\xi} \tag{2}$$

sendo $\boldsymbol{X} - \boldsymbol{\mu}$ uma matriz com dimensões $p \times 1$, os fatores $\boldsymbol{F}$ com dimensões $m \times 1$, os fatores específicos $\boldsymbol{\xi}$ com dimensões $p \times 1$ e os pesos

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1m} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{p1} & \phi_{p2} & \cdots & \phi_{pm} \end{bmatrix}.$$

No modelo usual de análise fatorial ortogonal, fazemos as seguintes suposições sobre $\boldsymbol{F}$: $E(F_i) = 0$ e $Var(F_i) = 1, i = 1, \ldots, m$. Além disso, num modelo ortogonal, consideramos que:

1. $E(\boldsymbol{F}) = 0$;
2. $E(\boldsymbol{\xi}) = 0$;
3. $Cov(\boldsymbol{F}) = E[\boldsymbol{F}\boldsymbol{F}^\intercal] = I_m$, em que I_m é a matriz identidade de ordem m;
4. $Cov(\boldsymbol{\xi}) = \boldsymbol{\Psi} = diag\{\psi_1, \ldots, \psi_p\}$;
5. $\boldsymbol{F}$ e $\boldsymbol{\xi}$ são independentes, isto é $Cov(\boldsymbol{\xi}; \boldsymbol{F}) = E(\boldsymbol{\xi}\boldsymbol{F}^\intercal) = 0$;

A partir dessas suposições, é possível analisar o modelo proposto e interpretar suas componentes.

2.1.1 Matriz de Covariância

A partir da matriz $\boldsymbol{X}$ de dados podemos fazer uma estimativa da matriz de covariância $\boldsymbol{\Sigma}$ a partir da equação (4) que é dado por:

$$\boldsymbol{\Sigma} = Cov(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^\intercal),$$

sabemos que

$$\begin{aligned} \mathbf{X}\mathbf{X}^\intercal &= (\boldsymbol{\Phi}\mathbf{F} + \boldsymbol{\xi})(\boldsymbol{\Phi}\mathbf{F} + \boldsymbol{\xi})^\intercal \\ &= (\boldsymbol{\Phi}\mathbf{F} + \boldsymbol{\xi})[(\boldsymbol{\Phi}\mathbf{F})^\intercal + \boldsymbol{\xi}^\intercal)] \\ &= \boldsymbol{\Phi}\mathbf{F}(\boldsymbol{\Phi}\mathbf{F})^\intercal + \boldsymbol{\xi}(\boldsymbol{\Phi}\mathbf{F})^\intercal + \boldsymbol{\Phi}\mathbf{F}\boldsymbol{\xi}^\intercal + \boldsymbol{\xi}\boldsymbol{\xi}^\intercal \end{aligned}$$

e de acordo com a suposição 5 temos:

$$\begin{aligned} \boldsymbol{\Sigma} &= Cov(\boldsymbol{X}) = E(\mathbf{X}\mathbf{X}^\intercal) \\ &= \boldsymbol{\Phi}E(\mathbf{F}\mathbf{F}^\intercal)\boldsymbol{\Phi}^\intercal + E(\boldsymbol{\xi}\mathbf{F}^\intercal)\boldsymbol{\Phi}^\intercal + \boldsymbol{\Phi}E(\mathbf{F}\boldsymbol{\xi}^\intercal) + E(\boldsymbol{\xi}\boldsymbol{\xi}^\intercal) \\ &= \boldsymbol{\Phi}\mathbf{I}\boldsymbol{\Phi}^\intercal + \mathbf{0} + \mathbf{0} + \boldsymbol{\Psi} \\ &= \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \boldsymbol{\Psi}. \end{aligned}$$

Sendo assim definimos que a matriz de covariância é dada pela expressão (3):

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \boldsymbol{\Psi} \tag{3}$$

2.1.2 Cargas Fatoriais

As cargas fatoriais são interpretadas a partir da $Cov(\mathbf{X}_i, \mathbf{F}_j)$, pois,

$$Cov(X_i, \mathbf{F}_j) = Cov(\phi_{i1}\mathbf{F}_1 + \ldots + \phi_{ij}\mathbf{F}_j + \ldots + \phi_{im}\mathbf{F}_m + \xi_p; \mathbf{F}_j)$$

Utilizando as suposições 3 e 5, temos que,

$$Cov(X_i, \mathbf{F}_j) = Cov(\phi_{ij}\mathbf{F}_j; \mathbf{F}_j)$$

Assim, verifica-se que as cargas fatoriais são as covariâncias são medidas entre as variáveis observadas e os fatores comuns.

A correlação é dada por existente entre o fator e cada variável observada, da seguinte maneira:

$$Corr(\mathbf{X}_i, \mathbf{F}_j) = \frac{Cov(\mathbf{X}_i, \mathbf{F}_j)}{\sqrt{Var(\mathbf{X}_i)Var(\mathbf{F}_j)}} = \frac{\phi_{ij}}{\sigma_i} \tag{4}$$

2.1.3 Comunalidades e Especificidade

A partir da equação 1 podemos chegar ao mesmo resultado acima utilizado apenas propriedades da variância que torna mais fácil o entendimento temos:

$$\mathbf{X}_j - \boldsymbol{\mu}_j = \phi_{j1}\mathbf{F}_1 + \phi_{j2}\mathbf{F}_2 + \ldots + \phi_{jm}\mathbf{F}_m + \boldsymbol{\xi}_j$$

Aplicando as propriedades da variância e com base nas suposições 3, 4 e 5, nota-se que,

$$\begin{aligned} Var(\mathbf{X_j}) &= \phi_{j1}^2 Var(\mathbf{F}_1) + \phi_{j2}^2 Var(\mathbf{F}_2) + \ldots + \phi_{jm}^2 Var(\mathbf{F}_m) + Var(\boldsymbol{\xi}_j) \\ &= \phi_{j1}^2 + \phi_{j2}^2 + \ldots + \phi_{jm}^2 + \psi_j \end{aligned} \tag{5}$$

De (5) conclui-se que a parcela da variância de $\mathbf{X}_j$ explicada pelo fatores comuns é dada por $c_j^2 = \phi_{j1}^2 + \phi_{j2}^2 + \ldots + \phi_{jm}^2$. Está quantia é conhecida como comunalidade dessa variável. O termo ψ_j representa a parte da variância que não pode ser explicada pelos fatores comuns definido como especificidade.

A comunalidade é costumeiramente redefinida na forma,

$$\bar{c}_j^2 = \frac{c_j^2}{\sigma_j^2}$$

o que melhora sua interpretação, pois tal medida assume valores no intervalo de [0,1] e pode ser entendida como a proporção da variabilidade de $\mathbf{X}_j$ explicada pelos fatores comuns. Quanto mais próximo de 1 é $\bar{c}_j^2$ melhor é o ajuste do modelo.

Analogamente, quanto maior é o valor da especificidade maior é a variação não explicada pelos fatores comuns e pior é o ajuste do modelo. Espera-se valores pequenos de ψ_j para todas as variáveis afim de se obter o melhor ajuste.

2.1.4 Padronização das Variáveis

Na Análise Fatorial é altamente indicada a padronização das variáveis quando há diferentes magnitudes de variâncias. Isso porque alguns métodos de estimação, como por exemplo, a Análise de Componentes Principais é muito sensível a essas diferenças. Nesse caso, as variáveis devem ser padronizadas antes, o que equivale ao uso da matriz de correlações das variáveis originais. A decomposição sugerida na equação (3) deve ser feita sobre a matriz de correlação em vez da matriz de covariância. Assim:

$$\boldsymbol{\rho} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\intercal} + \boldsymbol{\Psi}$$

em que $\boldsymbol{\rho}$ representa a matriz de correlação dos dados.

Consequência de realizar a análise sobre a matriz de correlação temos as seguintes adaptações de resultados anteriores:

1. $\phi_{ij} = Corr(\mathbf{X}_j, \mathbf{F}_j)$, é a correlação entre os fatores comuns e as variáveis originai;
2. A transformação das comunalidades por $\bar{c}_j^2$ é desnecessária, uma vez que, $\phi_{j1}^2 + \phi_{j2}^2 + \ldots + \phi_{jm}^2 + \psi_j = \rho_j = Corr(\mathbf{X}_j, \mathbf{X}_j) = 1$. Assim, $c_j^2 = \phi_{j1}^2 + \phi_{j2}^2 + \ldots + \phi_{jm}^2$ já representa a proporção da variância de $\mathbf{X}_j$ explicada pelos fatores comuns.
3. A variância total das variáveis padronizadas é igual ao número de variáveis originais, $\sigma_T^2 = p$, assim $\lambda_j = \sum_{i=1}^{p} \frac{\phi_{ij}^2}{p}$ é a média dos ϕ_{ij}^2's e representa a proporção da variância total explicada pelo fator j.
4. A proporção da variabilidade dos dados padronizados explicada pelo conjunto de fatores é $\sum_{i=1}^{p} \frac{c_i^2}{p}$, ou seja, é a média das comunalidades .

2.2 Procedimentos Gerais para a Análise Fatorial

Definem-se na literatura três estágios básicos para a construção da análise de fatores:

1) Obtenção dos fatores pelas suas cargas fatoriais;

2) Escolha da quantidade de fatores;

3) Rotação das cargas fatoriais;

2.2.1 Método para Obtenção de Fatores

1. **Análise de Componentes Principais**

   A Análise de Componentes Principais é uma das técnicas mais utilizadas para a obtenção dos fatores, pois este método não exige suposiçoes sobre a distribuição das variáveis originais. Está técnica depende apenas da matriz de covariância ($\boldsymbol{\Sigma}$) ou da matriz de correlação ($\boldsymbol{\rho}$) das variáveis originais $X_1, X_2, \ldots, X_p$ [2].

Considere a estrutura de dados como vista na seção (2.1), $\mathbf{X}$ de ordem $(n \times p)$ com matriz de covariância $\mathbf{\Sigma}$ de uma população.

A matriz de covariância $\mathbf{\Sigma}$ pode ser fatorada pelo processo de expansão de matrizes simétricas, denominado decomposição espectral. Dessa forma, $\mathbf{\Sigma}$ será representada por pares de autovalores λ_i, $i = 1, 2, \ldots, n$ e seus correspondentes autovetores normalizados $bm\tilde{a}_i$.

$$Cov(\mathbf{X}) = \mathbf{\Sigma} = \mathbf{\Gamma\Lambda\Gamma}^\intercal$$

em que $\mathbf{\Lambda}$ é uma matriz diagonal com os autovalores de $\mathbf{\Sigma}$ e $\mathbf{\Gamma} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p]$ é a matriz dos respectivos autovetores. Também podemos escrever da seguinte forma:

$$\mathbf{\Sigma} = \lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^\intercal + \ldots + \lambda_m\boldsymbol{\alpha}_m\boldsymbol{\alpha}_m^\intercal + \ldots + \lambda_p\boldsymbol{\alpha}_p\boldsymbol{\alpha}_p^\intercal. \tag{6}$$

As equações (6) e (3) indicam que a Análise de Componentes Principais faz a seguinte aproximação:

$$\mathbf{\Sigma} \approx \lambda_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^\intercal + \ldots + \lambda_m\boldsymbol{\alpha}_m\boldsymbol{\alpha}_m^\intercal + \ldots + \lambda_p\boldsymbol{\alpha}_p\boldsymbol{\alpha}_p^\intercal = \mathbf{\Phi\Phi}^\intercal$$

A análise de componentes principais propõe determinar a matriz $\mathbf{\Psi}$ utilizando a diagonal principal de $\mathbf{\Sigma} - \mathbf{\Phi\Phi}^\intercal$. Dessa forma temos:

$$\mathbf{\Psi} = diag\{\sigma_1^2 - \sum_{j=1}^{p}\phi_{1j}^2, \ldots, \sigma_p^2 - \sum_{j=1}^{p}\phi_{pj}^2\}$$

e ainda $\mathbf{\Sigma} \approx \mathbf{\Phi\Phi}^\intercal + \mathbf{\Psi}$. Assim temos

$$\mathbf{\Phi} = [\sqrt{\lambda_1}\boldsymbol{\alpha}_1, \ldots, \sqrt{\lambda_m}\boldsymbol{\alpha}_m] = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p]$$

em que $\boldsymbol{\phi}_j = (\phi_{1j}, \ldots, \phi_{pj})$.

Como os autovetores sao ortonormais, ou seja, $\boldsymbol{\alpha}_i^\intercal\boldsymbol{\alpha}_i = 1$ e $\boldsymbol{\alpha}_i^\intercal\boldsymbol{\alpha}_j = 0$, se $i \neq j$, temos $\sum_{j=1}^{p}\phi_{pj}^2 = (\sqrt{\lambda_j}\boldsymbol{\alpha}_i)^\intercal(\sqrt{\lambda_j}\boldsymbol{\alpha}_i) = \lambda_j$. Assim, o autovalor representa a parte da variância total das variáveis originais que é explicada pelo fator j.

Os componentes principais são p combinações lineares de $\boldsymbol{X}^\intercal = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ obtidos a partir da matriz de covariância Σ das p variáveis originais.

$$\begin{aligned}
\mathbf{Y}_1 &= \boldsymbol{a}_1^\intercal\mathbf{X} = a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + \ldots + a_{1p}\mathbf{X}_p \\
\mathbf{Y}_2 &= \boldsymbol{a}_2^\intercal\mathbf{X} = a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \ldots + a_{2p}\mathbf{X}_p \\
&\vdots \\
\mathbf{Y}_p &= \boldsymbol{a}_p^\intercal\mathbf{X} = a_{p1}\mathbf{X}_1 + a_{p2}\mathbf{X}_2 + \ldots + a_{pp}\mathbf{X}_p
\end{aligned}$$

Como os componentes principais são combinações lineares de $\boldsymbol{X}$, temos $Var(\mathbf{Y}_i) = Var(\boldsymbol{a}_p^\intercal \boldsymbol{X}) = \boldsymbol{a}_p^\intercal \boldsymbol{\Sigma} \boldsymbol{a_i}$ e, ainda $Cov(\mathbf{Y}_i, \mathbf{Y}_j) = Cov(\boldsymbol{a}_i^\intercal \boldsymbol{X}, \boldsymbol{a}_j^\intercal \boldsymbol{X}) = \boldsymbol{a}_i^\intercal \boldsymbol{\Sigma} \boldsymbol{a_j}$.

Restrições são necessárias para garantir que a soma das variâncias dos componentes seja igual à soma das variâncias das variáveis originais e que os componentes sejam não correlacionados. Dessa forma, o primeiro componente deve ser a combinação linear $\boldsymbol{a}_1^\intercal \boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_1^\intercal \boldsymbol{X})$ restrito a que $\boldsymbol{a}_1^\intercal \boldsymbol{a}_1 = 1$. Cada um dos demais componente i, $i = 2, \ldots, p$ deve ser a combinação linear $\boldsymbol{a}_i^\intercal \boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_i^\intercal \boldsymbol{X})$ restrito a que $Cov(\boldsymbol{a}_i^\intercal \boldsymbol{X}, \boldsymbol{a}_j \boldsymbol{X}) = 0$, para qualquer $j < i$.

Seja $(\lambda_i, \tilde{\boldsymbol{a}}_i)$, $i = 1, 2, \ldots, p$ pares de autovalores e autovetores ortogonais padronizados da matriz de covariância dos dados originais, ordenados de modo que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, o i-ésimo componente principal é dado pela combinação linear:

$$\mathbf{Y_i} = \tilde{\mathbf{a}}_\mathbf{i}\mathbf{X} = \tilde{\mathbf{a}}_{\mathbf{i1}}\mathbf{X_1} + \tilde{\mathbf{a}}_{\mathbf{i2}}\mathbf{X_2} + \ldots + \tilde{\mathbf{a}}_{\mathbf{ip}}\mathbf{X_p}, \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{p}.$$

com

(a) $Var(\mathbf{Y}_i) = \tilde{\boldsymbol{a}}_{\boldsymbol{i}}^\intercal \boldsymbol{\Sigma} \tilde{\boldsymbol{a}}_{\boldsymbol{i}} = \lambda_i$, $i = 1, 2, \ldots, p$;

(b) $Cov(\mathbf{Y}_i, \mathbf{Y}_j) = \tilde{\boldsymbol{a}_i}^\intercal \boldsymbol{\Sigma} \tilde{\boldsymbol{a}_j} = 0, i \neq j$;

Como vimos anteriormente, a soma das variâncias permanece a mesma após a transformação linear:

$$\sum_{j=1}^{p} Var(X_j) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{j=1}^{p} Var(\mathbf{Y}_j)$$

Com isso, verifica-se que a variância total dos dados originais será a mesma dos componentes principais que é a soma dos autovalores. A proporção da variância total para o i-ésimo componente principal é dado por:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}, \quad i = 1, 2, \ldots, p.$$

Os autovetores $\tilde{a_i}$ apresentam as seguintes propriedades:

(a) São normalizados, ou seja, a soma dos quadrados dos coeficientes é igual a 1;

$$\sum_{j=1}^{p} \tilde{a_{ij}}^2 = 1, \quad (\tilde{a_i}^\intercal . \tilde{a_i} = 1)$$

(b) São ortogonais entre si, que significa que o produto escalar entre si é nulo;

$$\sum_{j=1}^{p} \tilde{a_{ij}} . \tilde{a_{kj}} = 0, \quad (\tilde{a_i}^\intercal . \tilde{a_k} = 0 \quad para \quad i \neq k)$$

Para apresentar o metodo de obtenção das cargas fatorias por Componentes Principais, vamos utilizar a matriz de covariância amostral $\mathbf{S}$, simétrica e de ordem $p \times p$.

$$\mathbf{S} = \begin{pmatrix} \hat{V}ar(\mathbf{x}_1) & \hat{C}ov(\mathbf{x}_1\mathbf{x}_2) & \cdots & \hat{C}ov(\mathbf{x}_1\mathbf{x}_p) \\ \hat{C}ov(\mathbf{x}_2\mathbf{x}_1) & \hat{V}ar(\mathbf{x}_2) & \cdots & \hat{C}ov(\mathbf{x}_2\mathbf{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{C}ov(\mathbf{x}_p\mathbf{x}_1) & \hat{C}ov(\mathbf{x}_p\mathbf{x}_2) & \cdots & \hat{V}ar(\mathbf{x}_p) \end{pmatrix}$$

A matriz de covariância só deve ser usada se as variáveis originais estiverem na mesma escala de medida, caso contrário, é necessário padronizar as variáveis $X_1, \ldots, X_p$ transformando-as em variáveis que possuam média **zero** e a variância **um**.

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{S(x_j)}, i = 1, 2, \ldots, n \quad e \quad j = 1, 2, \ldots, p$$

em que $\bar{x_j}$ e $S(x_j)$ são respectivamente, a estimativa da média e o desvio padrão da variável j e são dados por:

$$\bar{x_j} = \frac{\sum_{i=1}^{n} x_{ij}}{n} \quad e$$

$$S(x_j) = \sqrt{\hat{V}ar(x_j)}, \quad j = 1, 2, \ldots, p, \text{ em que, } \hat{V}ar(x_j) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}{n-1}$$

Após a padronização o resultado é uma nova matriz de dados:

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1p} \\ Z_{21} & Z_{22} & \cdots & Z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{np} \end{pmatrix}$$

Utilizar a matriz padronizada $\mathbf{Z}$ é equivalente ao uso da matriz $\mathbf{R}$, correlações amostral dos dados, para obter os Componentes Principais. A matriz $\mathbf{R}$ é a mais utilizada para a análise que a matriz de covariância, pois, não há pressuposição na unidade de medida das variáveis e a interpretação das comunalidades é direta. Uma importante observação é que o resultado dos componentes principais pode ser diferente caso a análise seja feita com a matriz $\mathbf{R}$ e com a matriz de covariância. A matriz de correlação é dada por:

$$\mathbf{R} = \begin{pmatrix} 1 & r(x_1, x_2) & \cdots & r(x_1, x_p) \\ r(x_2, x_1) & 1 & \cdots & r(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ r(x_p, x_1) & r(x_p, x_2) & \cdots & 1 \end{pmatrix}$$

O cálculo para determinar os componentes principais é dado resolvendo a equação característica da matriz $\mathbf{R}$ (ou $\mathbf{S}$), isto é:

$$det[\mathbf{R} - \boldsymbol{\lambda}\mathbf{I}] = 0$$

resultando em p autovalores, que são as p raízes da equação $det[\mathbf{R} - \boldsymbol{\lambda}\mathbf{I}] = 0$. É recomendado que na matriz $\boldsymbol{X}$ de dados tenha o valor de n (tamanho da amostra) igual ou maior a $p+1$, isto é, tenha mais indivíduos do que variáveis nos dados, caso contrario a solução da combinação linear não será possível.

Sejam os autovalores $\lambda_1, \lambda_2, \cdots, \lambda_p$ as raízes da equação características da matriz $\boldsymbol{R}$ (ou $\boldsymbol{S}$), então sabemos que o primeiro autovalor terá o maior valor do que o seguinte, e assim por diante, isto é:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

Para cada autovalor λ_i existe um autovetor correspondente dado como $\tilde{a_i}$:

$$\tilde{a_i} = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{bmatrix}$$

Os autovetores $\tilde{a_i}$ apresentam as seguintes propriedades:

(a) São normalizados, ou seja, a soma dos quadrados dos coeficientes é igual a 1;

$$\sum_{j=1}^{p} \tilde{a_{ij}}^2 = 1, \quad (\tilde{a_i}^{\intercal}.\tilde{a_i} = 1)$$

(b) São ortogonais entre si, que significa que o produto escalar entre si é nulo;

$$\sum_{j=1}^{p} \tilde{a_{ij}}.\tilde{a_{kj}} = 0, \quad (\tilde{a_i}^{\intercal}.\tilde{a_k} = 0 \quad para \quad i \neq k)$$

Insuma, podemos reescrever a nossa matriz de covariância da seguinte forma:

$$\Sigma = \lambda_1 \tilde{a_1}\tilde{a_1}^{\intercal} + \lambda_2 \tilde{a_2}\tilde{a_2}^{\intercal} + \ldots + \lambda_p \tilde{a_p}\tilde{a_p}^{\intercal}$$

$$\begin{bmatrix} \sqrt{\lambda_1 \tilde{a_1}} & \vdots & \sqrt{\lambda_2 \tilde{a_2}} & \vdots & \ldots & \vdots & \sqrt{\lambda_p \tilde{a_p}} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1 \tilde{a_1}^{\intercal}} \\ \sqrt{\lambda_2 \tilde{a_2}^{\intercal}} \\ \vdots \\ \sqrt{\lambda_p \tilde{a_p}^{\intercal}} \end{bmatrix} \tag{7}$$

Os autovetores determinam as direções de variabilidade máxima e os autovalores especificam as variâncias. Quando os primeiros autovalores são muito maiores do que os restantes, a maior parte da variância total pode ser explicada em um número menor que p dimensões.

Pelo modelo da Análise Fatorial dado pela equação (1), podemos escrever o nosso modelo fatorial da seguinte maneira:

$$X_j - \mu_j = a_{j1}\mathbf{F}_1 + \ldots + a_{jm}\mathbf{F}_m + \xi_j,$$

em que, $a_{ji} = \sqrt{\lambda_i \tilde{a_{ij}}}$.

2. Método da Máxima Verossimilhança

Em 1940 Lawley [11] propôs o método da Máxima Verossimilhança para Análise Fatorial. No entanto, esse método não é muito utilizado por suas dificuldades nos cálculos e computacionais. Hoje já é possível encontrar procedimentos rápidos e eficientes para a obtenção dos estimadores por este método [5].

A principal vantagem de utilizar o Método da Máxima Verossimilhança é possibilidade de se desenvolverem testes de hipóteses, com o objetivo de testar a adequacidade do modelo, o que não é possível através do método de Componentes Principais, devido às suas características não estatísticas.

Além das suposições habituais do modelo fatorial, supomos que os vetores aleatórios F (fatores comuns) e ξ (fatores específicos) têm distribuição normal multivariada. Sendo assim temos:

$$\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

,com

$$\boldsymbol{\Sigma} = \Phi\Phi^{\intercal} + \boldsymbol{\Psi}$$

Um problema com esse modelo é que ele não é identificável, já que há infinitas matrizes $\boldsymbol{\phi}$ que satisfaz essa igualdade. Com isso surgiu às restrições de identificabilidade. Uma restrição conveniente do ponto de vista computacional é fazer com que, $\boldsymbol{\phi^{\intercal}\Psi^{-1}\phi}$ seja uma matriz diagonal, facilitando nos cálculos.

Sejam $X_1, \ldots, X_n$ uma amostra aleatória independentes. As estimativas de máxima verossimilhança de $\boldsymbol{\mu, \phi}$ e $\boldsymbol{\Psi}$ são obtidas a partir da maximização da função de verossimilhança a seguir:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma^{-1}}(\boldsymbol{x_i} - \boldsymbol{\mu})\right] \tag{8}$$

Não há uma solução explícita para os estimadores, o que exige o uso de métodos numéricos para a maximização da função acima. O estudo de tais métodos está acima do nível deste texto (vide Anderson [1], para maiores detalhes).

Como dito anteriormente, a vantagem de utilizar o método da máxima verossimilhança é que a inferência estatística nos garante sua consistência e normalidade assintótica, que nos leva a construção de intervalos de confiança e testes de hipótese para grandes amostras.

Para avaliar a escolha do número de fatores pode-se fazer o teste de hipótese:

$$\begin{cases} H_0: & \boldsymbol{\Sigma} = \boldsymbol{\Phi\Phi^{\intercal}} + \boldsymbol{\Psi} \\ H_1: & \boldsymbol{\Sigma} \neq \boldsymbol{\Phi\Phi^{\intercal}} + \boldsymbol{\Psi} \end{cases}$$

Sejam,

$$\boldsymbol{S_n} = \sum_{i=1}^{n} \frac{(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^{\intercal}}{n} \quad e \quad \Sigma = \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Phi}} + \hat{\boldsymbol{\Psi}}$$

em que, $\hat{\mathbf{\Phi}}$ e $\hat{\mathbf{\Psi}}$ são os estimadores da máxima verossimilhança de $\mathbf{\Phi}$ e $\mathbf{\Psi}$, respectivamente. A estatística do teste é dada por:

$$TRV = -2ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\boldsymbol{S}}_{\boldsymbol{n}}|}\right)$$

Sob a hipótese nula, TRV segue uma distribuição χ^2_g, ou seja, uma distribuição qui-quadarado com g graus de liberdade definido a seguir:

$$g = \frac{1}{2}\{(p-m)^2 - p - m\}$$

2.2.2 Escolha do Número de Fatores

A determinação do número de fatores para Zwick e Velicer [20] é a decisão mais importante do investigador. A ocorrência de erros nesta fase afetará a interpretação de todos os resultados subsequentes. A este respeito, Velicer e Jackson [19] mostraram que a superextração (a extração de um número maior de fatores dos que realmente o necessario) é um erro tão sério como a subextração (a extração de um número menor de fatores) ambos levarão a resultados e conclusões distorcidas.

Na literatura existem vários critério que auxiliam na determinação do número de fatores, porém esses critérios podem conduzir a resultados diferentes mesmo aplicados no mesmo conjunto de dados e também, não existe um critério considerado melhor que um outro. Por esses motivos é indicado utilizar mais de um critério. O analista deve procurar um número de fatores em comum entre os diferentes critérios.

Neste estudo focaremos nos quatros critérios mais utilizados para a definição do número de fatores. Esses critérios podem ser utilizados como ponto de partida para a obtenção de uma solução final na análise.

1. **Critério de Kaiser**

   O Critério de Kaiser consiste em considerar apenas aqueles autovalores que são maiores que o valor 1 [9]. Lembrando que o autovalor corresponde à quantidade da variância explicada por um fator, então sendo que um autovalor igual a 1 representa a variância explicada de uma única variável e abaixo da 1 o fator estaria explicando menos que uma das variáveis originais.

   Quando análise é realizada sobre a matriz de covariância, ao invés de usar o valor 1 como referencia o ponto de corte deve ser a média de todos os autovalores.

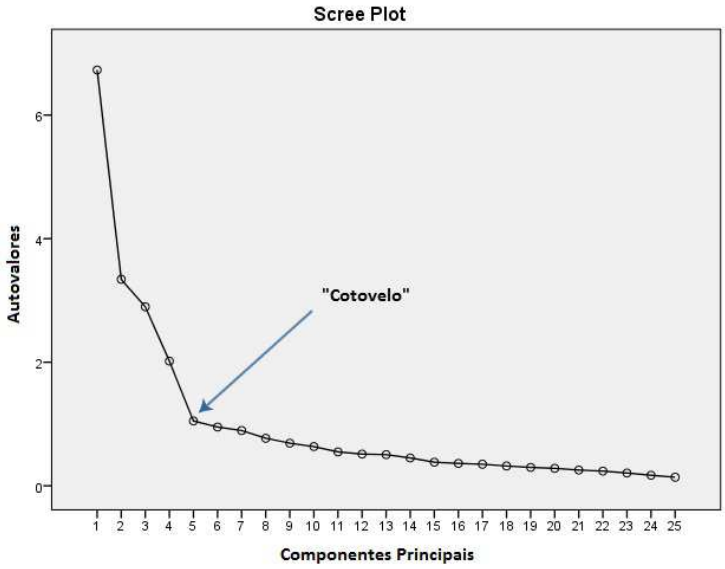
2. **Critério da Porcentagem da Variância Explicada**

   Este critério é um pouco mais subjetivo. Determina-se o número de fatores que explique uma porcentagem pré-definida da variabilidade global. É comum adotar um valor de explicação de 70% como mínimo. Porém pode mudar de acordo a área do problema.

3. ***Scree-Test***

   Este método é baseado no gráfico dos autovalores, também conhecido como Scree-Plot. O eixo y representa os valores dos autovalores e o eixo x mostra o número sequencial dos

Figura 1: Scree-Plot - Exemplo



fatores. O teste consiste em separar os fatores triviais do início de fatores não triviais por intermédio de uma inspeção visual do gráfico. A decisão subjetiva está baseada no uso de uma linha reta colocada ao longo da parte do fundo do gráfico onde os pontos formam uma linha aproximadamente reta. Os pontos acima da linha reta são associados com fatores não triviais, enquanto os pontos restantes representam os fatores triviais. Esse método também é descrito popularmente como buscar o "cotovelo" no gráfico. Este cotovelo é o marco de onde começaria a reta no gráfico. O problema desse critério e um possivel mau julgamento dos pesquisadores em identificar esse "cotovelo". Um exemplo é dado na figura 1 onde a um indício de se escolher cinco fatores pois, nesse ponto é onde começam a formar a linha reta ou conhecido como "cotovelo".

4. **Métodos Inferenciais**

   Quando as variáveis originais seguem uma distribuição normal pode se utilizar métodos que consistem no desenvolvimento de testes estatísticos que se alicerçam na suposição de normalidade, então não se deve utilizar tais critérios quando os dados não são normais. Este estudo não vai focar esses tipos de métodos porém dentre esses métodos destacamos o de Bartlett [3] que verifica a adequação do modelo de Análise Fatorial (pelo o método da máxima verossimilhança) para presentar a estrutura de dependência dos dados.

2.2.3 Rotações dos Fatores

Para procurar uma melhor interpretação dos fatores, é prática comum fazer uma rotação ou uma transformação dos fatores. Geometricamente, a operação de multiplicar a matriz de cargas fatoriais Φ por uma matriz ortogonal equivale a fazer uma rotação de eixos, resultando na prática novos fatores, porém esses novos fatores estão rotacionados.

Existem dois tipos gerais de rotação: as rotações ortogonais, nas quais os fatores continuarAM a ser não correlacionados, suas comunalidades e especificidades continuam preservadas;

ou rotações oblíquas que geram fatores correlacionados. Essas últimas não serão o foque desse trabalho.

Dentro das rotações ortogonais existem diversos tipos de procedimentos, por exemplo, a Varimax, Quartimax, Equamax, Orthomax e Parsimax. Neste estudo vamos focar apenas a rotação Varimax.

Este método de rotação ortogonal foi proposto por Kaiser [9]. A idéia do método é maximizar a variância das cargas fatoriais para cada fator por meio do aumento das cargas altas e a diminuição das cargas baixas, ou seja, propõe fazer com que cada variável seja altamente correlacionada com apenas um fator, e nos outros fatores essa correlação será baixa.

Sejam $\hat{\boldsymbol{a_{ij}}}$, $i = 1, \ldots, p$, $j = 1, \ldots, m$ as cargas fatoriais rotacionadas e c_i a comunalidade de $\boldsymbol{X}_i$. Podemos definir que:

$$\beta_{ij} = \frac{\hat{a_{ij}}^2}{\hat{c_j}^2} \quad e \quad \bar{\beta}_{ij} = \sum_{i=1}^{p} \frac{\beta_{ij}}{p}$$

Podendo β_{ij} ser interpretado como a proporção da comunalidade de $\boldsymbol{X}_i$ que é explicada pelo o fator j. Concluindo assim que a matriz de rotação $\boldsymbol{T}$ será escolhida baseada na maximização do seguinte conceito:

$$V = \sum_{j=1}^{m} V_j,$$

em que,

$$V_j = \sum_{i=1}^{p} \frac{(\beta_{ij} - \bar{\beta}_j)^2}{p}.$$

O método numérico de maximização e os seus detalhes não serão abordado nesse estudo. Porém a ideia é que V_j é a variância amostral de β_{ij}, em que $i = 1, \ldots, p$. Logo ao maximizar V, consequentemente V_j será maximizado. Como esperado V_j sumirá um valor alto quando houver valores muito altos para alguns β_{ij} e baixos para os demais. Esse processo que faz com que a rotação Varimax seja muito útil em detectar quais são as variáveis que mais se relacionam com um determinado fator.

2.3 Estimação dos *scores* Fatoriais

Quando a Análise Fatorial é preliminar a algum outro tipo de análise multivariada, ou quando o seu uso principal é para construção de índices, é necessário, procurar descrever os fatores em termos das variáveis observadas. Para isto, estimam-se os valores, denominados *scores* fatoriais, de cada fator para cada indivíduo.

Serão considerados dois métodos de estimação dos *scores* dos fatores, o método dos mínimos quadrados ponderados e o método da regressão. Conforme apresentado por Fachel [5] estes dois métodos serão descritos, supondo que os fatores sejam correlacionados.

Ambos os métodos apresentam os seguintes pressupostos [8]:

1. As estimativas das cargas fatoriais $\hat{a_{ij}}$ e das variâncias específicas ξ_j são tomadas como valores paramétricos;

2. Os métodos envolvem transformações lineares dos dados originais padronizados, e geralmente, utilizam as cargas estimadas nos cálculos dos *scores*.

2.3.1 Método dos Mínimos Quadrados

Relembrando o modelo de uma Análise Fatorial ortogonal, temos para a observação i que:

$$\boldsymbol{X_i} - \boldsymbol{\mu} = \boldsymbol{\Phi}\mathbf{F}_i + \boldsymbol{\xi_i}$$

O modelo acima se assemelha a um modelo de regressão linear, comparando as partes percebemos que $\boldsymbol{X_i} - \boldsymbol{\mu}$ desempenha o papel da variável dependente, o $\boldsymbol{\Phi}$ da matriz de variáveis independentes, $\mathbf{F}_i$ o do vetor de parâmetros e $\boldsymbol{\xi_i}$ o vetor de erros.

Quando determinamos que $\boldsymbol{Psi}$ é conhecido, recomenda-se a utilização do método dos mínimos quadrados ponderados ao invés do método de mínimos quadrados ordinários. Nesse método, o previsor de $\mathbf{F}_i$ será aquele que minimizar a equação a seguir:

$$Q(\mathbf{F}_i) = (\boldsymbol{X_i} - \boldsymbol{\mu} - \boldsymbol{\Phi}\mathbf{F}_i)^\intercal \boldsymbol{\Psi}^{-1}(\boldsymbol{X_i} - \boldsymbol{\mu} - \boldsymbol{\Phi}\mathbf{F}_i)$$

O previsor será encontrado a partir,

$$\hat{\mathbf{F}_i} = (\boldsymbol{\Phi}^\intercal \boldsymbol{\Psi}^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}_\intercal \boldsymbol{\Psi}^{-1}(\boldsymbol{X_i} - \boldsymbol{\mu})$$

2.3.2 Método da Regressão

Para estimar o $\mathbf{F}_i$ é necessário fazer a previsão através da esperança condicional de $\mathbf{F}$ dado $\boldsymbol{X_i}$. Para isso é necessário que $\boldsymbol{\mu}$,$\boldsymbol{\Phi}$ e $\boldsymbol{\Psi}$ sejam conhecidos, além da suposição que os fatores $\mathbf{F}$ e o vetor de erros ($\boldsymbol{\xi}$) sejam normalmente distribuídos. Os detalhes dos resultados desse método não serão apresentados nesse estudo. Por fim temos que o previsor de $\mathbf{F}_i$ será dado por:

$$\hat{\mathbf{F}_i} = \boldsymbol{\Phi}^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{X_i} - \boldsymbol{\mu}) = \boldsymbol{\Phi}^\intercal(\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \boldsymbol{\Psi})^{-1}(\boldsymbol{X_i} - \boldsymbol{\mu})$$

2.4 Estudo de Viabilidade da Análise Fatorial

Por fim, é necessário apresentar medidas adicionais que ajudam a entender a viabilidade da aplicação de uma Análise Fatorial a um conjunto de dados. Serão apresentadas três medidas para auxiliar nessa viabilidade.

2.4.1 Matriz Anti-Imagem

A principal premissa de uma Análise Fatorial é que exista uma estrutura de dependência clara entre as variáveis envolvidas. Sendo essa estrutura a matriz de covariância ou de correlação. Se tal estrutura existe podemos implicar que uma variável pode, dentro de certos limites, ser prevista pelas demais. Para isso calculamos os coeficientes de correlação parcial entre os pares de variáveis, eliminando o efeito das demais variáveis. O ideal é que os valores obtidos sejam baixos. A matriz resultante desse processo é chamada de matriz anti-imagem, e ela é construída com esses coeficientes com sinais invertidos.

2.4.2 KMO: Kaiser-Meyer-Olkin

O coeficiente de KMO parte do princípio da matriz anti-imagem, ou seja, as correlações parciais entre as variáveis. O coeficiente é dado por:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2}$$

em que, a_{ij}^2 é a correlação parcial entre duas variáveis $\mathbf{X}$, eliminando o efeito das demais variáveis.

O coeficiente de KMO é interpretado na literatura baseado na proposta feita por Kaiser e Rice (1974) apresentada na Tabela (1) a seguir.

Tabela 1: Interpretação do KMO

KMO	Intepretação
0,90-1,00	Excelente
0,80-0,90	Ótimo
0,70-0,80	Bom
0,60-0,70	Regular
0,50-0,60	Ruim
0,50-0,60	Inadequado

2.4.3 MSA: Measure of Sampling Adequacy

Similar ao KMO essa medida verifica se existe uma estrutura fatorial nos dados. Porém a MSA deve ser calculada separadamente para cada variável, pois o objetivo é verificar se uma dada variável pode ser explicada pelas as demais. Valores baixos indicam que a variável analisada por ser retirada da análise sem maiores prejuízos. Essa medida é dada pela seguinte formula:

$$MSA = \frac{\sum_{j=1}^{p} r_{ij}^2}{\sum_{j=1}^{p} r_{ij}^2 + \sum_{j=1}^{p} a_{ij}^2}$$

Para termos uma noção do desempenho do conjunto das variáveis, pode-se calcular a média dos MSA_i tendo assim um ponto de referência, como indicado a seguir,

$$\overline{MSA} = \sum_{i=1}^{p} \frac{MSA_i}{p}.$$

3 Resultados

A pesquisa em primeiro momento consistiu em organizar os dados, para depois iniciar a análise com auxílio do software JMP 11 e o R versão 2.15.2. Este estudo tem como fim identificar os principais fatores que possam agrupar subconjuntos de variáveis mais relacionadas. O exemplo utilizado para aplicação dessa análise foi dados referente ao comportamento de compra de varejistas.

O primeiro passo da análise consiste em verificar a distribuição do conjunto de dados. Para isso foi utilizado o teste multivariado de normalidade de Shapiro-Wilk para verificar a multinormalidade dos dados. Em seguida, medidas de adequação da amostra o KMO e medida de adequação de cada variável MSA foram calculadas para analisa a estrutura de correlação dos dados. Essa análise é importante para verificar se os dados são apropriados à aplicação da Análise Fatorial.

Na construção do modelo, três critérios foram utilizados para identificar o número de fatores a serem extraídos: análise do Scree Plot, porcentagem da variância explicada e o critério de Kaiser. Procedeu-se a Análise Fatorial sendo desenvolvida a extração dos fatores pelo método das componentes principais.

A base analisada nesse estudo é composta por 158.812 varejistas de todo o Brasil. Para a seleção dessa base foi necessário selecionar clientes considerados ativos durante um período pré estabelecido de Janeiro a Agosto de 2016. Um cliente ativo é definido por ter feito pelo menos uma compra durante esse período.

Durante o desenvolvimento do projeto algumas variáveis foram alteradas devido à dificuldade na obtenção dos dados ou pela detecção de uma nova variável que possa contribuir para melhorar a determinação dos potenciais. A construção da Análise Fatorial consiste em 12 variáveis, sendo elas:

1. ***DIVISAO_FORNECEDOR***: Baseado no histórico de compra do cliente, essa variável representa a quantidade distinta de industrias (fornecedor) que o cliente comprou durante esses oito meses.

2. ***VLR_LIMITE_CRE***: O valor do limite de crédito disponível para o cliente.

3. ***FREQUENCIA***: A frequência mensal de compra do cliente.

4. ***DIAS_COMPRA***: A quantidade de dias que durante esses oito meses o cliente fez um pedido.

5. ***TOTAL_PEDIDOS***: A quantidade total de pedidos feitos pelo o cliente durante esses oitos meses.

6. ***QTDITENS***: A quantidade de itens distintos que o cliente comprou durante esses oitos meses.

7. ***QTD_CATEG***: A quantidade de categorias distintas que o cliente comprou durante esses oitos meses. Uma categoria é um grupo de diferentes itens com características em comum.

8. ***FAT_TOTAL***: O faturamento total que o cliente trouxe para a empresa, ou seja, o total de compras que o cliente fez durante esses oito meses.

9. ***VLR_IMPOSTO***: O total de impostos associado a um produto pago pelo o cliente.

10. ***NUM_DEVOLUCAO***: A quantidade total de devoluções efetuadas pelo o cliente.

11. ***VLR_DEVOLUCAO***: O valor das devoluções efetuadas pelo o cliente .

12. ***ANOS_CADASTRO***: Em anos, o tempo de cadastro dos clientes.

As variáveis utilizadas para esse estudo definem de um modo geral o comportamento de compra do cliente. Sendo essas variáveis informações em valores de venda e até a quantidade de algumas medidas importantes que depois possam caracterizar a compra do cliente.

Uma estatística descritiva foi realizada nas variáveis para dar uma noção geral do comportamento das variáveis e também é necessário entender a distribuição dessas variáveis para determinar medidas importante para a execução da Análise Fatorial. A tabela (2) mostra as estatísticas descritivas das 12 variáveis.

Podemos concluir dois pontos importantes com essa informação. O primeiro é que as variáveis estão em unidade de medidas muito diferentes e desproporcionais impossibilitando o uso da matriz de covariância para a Análise Fatorial, é notório isso entre por exemplo a mediana das variáveis QTD_CATEG e FAT_TOTAL com valores de 12 e 3.945,86 respectivamente. O segundo fato importante do resultado da estatística descritiva é a grande assimetria em quase todas as variáveis dos nossos dados, que é um indício de utilizar o método dos componentes principais na extração dos fatores por não exigir uma distribuição nos dados.

Tabela 2: Estatística Descritiva: Variáveis Quantitativas

Variável	Média	Desvio Padrão	Minímo	Percentil 25%	Mediana	Percentil 75%	Máximo
DIVISAO_FORNECEDOR	15,40	14,15	1	4	11	22	141
VLR_LIMITE_CRE	12.815,41	15.631,17	0	5.000	7.942	11.474	600.000
FREQUENCIA	25,42	29,57	1	5	15	35	438
DIAS_COMPRA	15,84	14,59	1	5	12	23	802
TOTAL_PEDIDOS	28,16	35,21	1	6	16	38	865
QTDITENS	330,80	595,66	0	50	159	402	43614
QTD_CATEG	17,19	16,72	1	5	12	25	189
FAT_TOTAL	6.165,59	7.972,02	0	1.611,905	3.945,86	8.095,07	318.959,15
VLR_IMPOSTO	901,67	1.224,44	0	214,475	535,21	1.148,595	6.028,59
NUM_DEVOLUCAO	0,18	0,57	0	0	0	0	27
VLR_DEVOLUCAO	637,49	59.236,17	0	0	0	0	27.968,37
ANOS_CADASTRO	8,60	8,09	0	2	6	13	44

Para verificar a normalidade dos dados, foi aplicado o teste multivariado de normalidade Shapiro-Wilk em uma amostra de 3.000 observações dos nossos dados. Conforme a Tabela (3) o teste obteve-se um p-valor menor $0,001$, esse valor indica que se rejeita a hipótese nula de que não há normalidade nós dados o nível de 5%. Confirmando a utilização do método de componentes principais na aplicação da analise fatorial.

Tabela 3: Teste de Normalidade Multivariada de Shapiro-Wilk

$Estatistica$	$P-Valor$
$0,2025$	$< 0,001*$

Como analisado anteriormente as variáveis são medidas em magnitudes diferentes então a Análise Fatorial será aplicada à matriz de correlações que se encontra na Tabela (4). A Figura (2) representa um gráfico da matriz de correlação. Quanto mais escuro a cor azul mais forte são as correlações entre as variáveis.

Figura 2: Correlograma da Matriz de Correlação

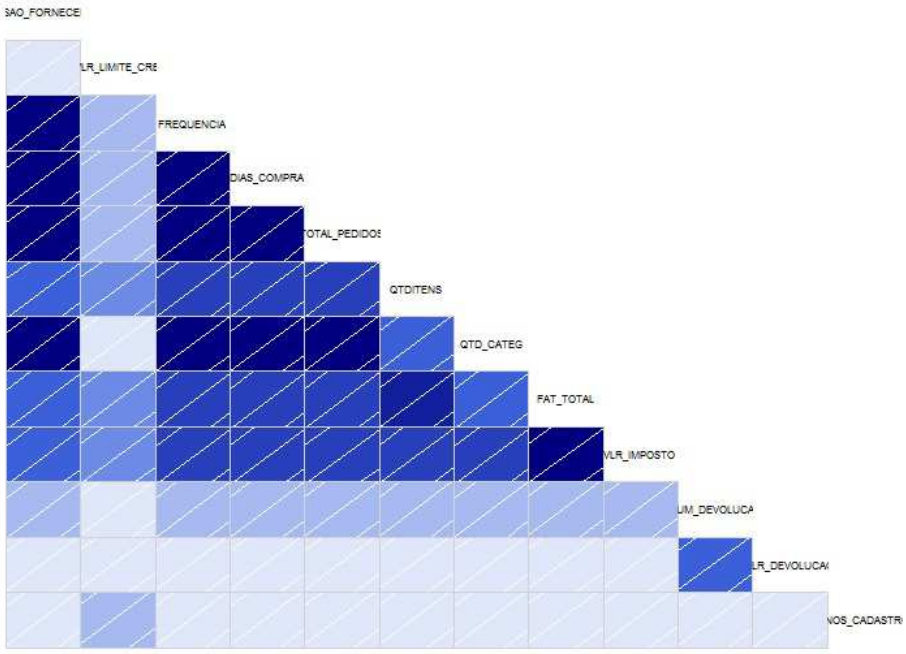


Tabela 4: Matriz de Correlação

DIVISAO_FORNECEDOR	1,000	0,131	**0,940**	**0,907**	**0,904**	**0,510**	**0,992**	**0,541**	**0,560**	0,217	0,050	0,094
VLR_LIMITE_CRE	0,131	1,000	0,165	0,172	0,179	0,290	0,137	0,340	0,340	0,131	0,087	0,279
FREQUENCIA	**0,940**	0,165	1,000	**0,990**	**0,988**	**0,580**	**0,951**	**0,612**	**0,610**	0,238	0,057	0,113
DIAS_COMPRA	**0,907**	0,172	**0,990**	1,000	**0,998**	**0,590**	**0,924**	**0,630**	**0,620**	0,250,	0,064	0,112
TOTAL_PEDIDOS	**0,904**	0,179	**0,988**	**0,998**	1,000	**0,620**	**0,922**	**0,647**	**0,630**	0,253	0,068	0,111
QTDITENS	**0,511**	0,288	**0,580**	**0,595**	**0,616**	1,000	**0,542**	**0,788**	**0,580**	0,248	0,126	0,080
QTD_CATEG	**0,992**	0,137	**0,951**	**0,924**	**0,922**	**0,540**	1,000	**0,560**	**0,570**	0,228	0,056	0,091
FAT_TOTAL	**0,541**	0,340	**0,612**	**0,630**	**0,647**	**0,790**	**0,560**	1,000	**0,870**	0,233	0,133	0,089
VLR_IMPOSTO	**0,564**	0,340	**0,614**	**0,625**	**0,632**	**0,580**	**0,575**	**0,867**	1,000	0,233	0,123	0,107
NUM_DEVOLUCAO	0,217	0,131	0,238	0,2500	0,253	0,250	0,228	0,233	0,230	1,000	**0,532**	0,057
VLR_DEVOLUCAO	0,050	0,087	0,057	0,064	0,068	0,130	0,056	0,133	0,120	**0,532**	1,000	0,013
ANOS_CADASTRO	0,094	0,279	0,113	0,112	0,111	0,080	0,091	0,089	0,110	0,057	0,013	1,000

Uma medida global de adequação amostral é dada pelas estatísticas de KMO e MSA que parte do princípio da matriz anti-imagem, ou seja, as correlações parciais entre as variáveis. Obtevemos a estatística de KMO de 0,825, que pela Tabela (1) o índice indica que segundo Barroso [2] os dados estão ótimos para a realização do modelo fatorial. Foi obtido o índice de MSA para cada variável (Tabela (5) e conclui que é aceitável prosseguir com a análise fatorial, pois todas as variáveis possuem um valor maior que 0,5.

Tabela 5: Estatística de MSA

Variavel	*MSA*
DIVISAO_FORNECEDOR	0,821
VLR_LIMITE_CRE	0,844
FREQUENCIA	0,959
DIAS_COMPRA	0,824
TOTAL_PEDIDOS	0,837
QTDITENS	0,809
QTD_CATEG	0,8308
FAT_TOTAL	0,760
VLR_IMPOSTO	0,797
NUM_DEVOLUCAO	0,720
VLR_DEVOLUCAO	0,563
ANOS_CADASTRO	0,678

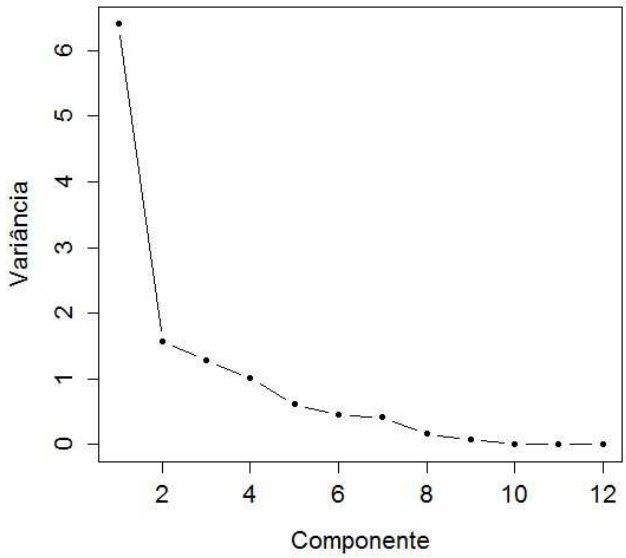
Na tabela (6) são apresentados os autovalores. Neste caso, já podemos ter um indício da quantidade de fatores a serem extraídos na análise. Utilizando o critério de Kaiser, todos os autovalores maiores que o valor um devem ser incluídos na Análise Fatorial, ou seja, neste caso seriam 4 fatores extraídos. O outro critério, o método da variância explicada, se seguido o valor de atingir pelo menos 70% da variância explicada, apenas três fatores seriam extraídos, entretanto esse critério é muito subjetivo pois o valor sugerido de 70% muda conforme diferentes literatura.

Tabela 6: Matriz de Autovalores

Componente	*Autovalor*	*Porcentagem*	*PorcentagemAcumulada*
1	6,400	53,335	53,335
2	1,569	13,082	66,417
3	1,282	10,690	77,107
4	1,011	8,430	85,537
5	0,610	5,085	90,622
6	0,453	3,780	94,402
7	0,416	3,468	97,871
8	0,164	1,368	99,239
9	0,075	0,630	99,869
10	0,008	0,071	99,940
11	0,005	0,048	99,989
12	0,001	0,011	100,000

Entre os dois métodos a melhor escolha de extração estaria entre 3 ou 4 fatores para a análise, para decidirmos um último método foi utilizado, o gráfico Scree Plot (Figura (3). Pelo o gráfico é possível identificar o "cotovelo"no 3º componente. A partir dos três critérios foi decidido a extração de quatro fatores, pois quatro é um valor em comum em todos os critérios como o critério de Kaiser é baseado em critérios matemáticos o que o torna mais confiável.

Figura 3: ScreePlot



Para a extração dos fatores foi utilizado o método de componentes principais, já que não exige uma distribuição nos dados. A Tabela (7) apresenta as cargas fatoriais dos 4 fatores. Hair [6], acentua que para uma amostra de 100 observações ou mais, as cargas fatoriais são consideradas significativas para fins de interpretação quando apresentam um valor maior que 0,55. Percebemos que apenas o primeiro fator se relaciona fortemente com algumas variáveis, os outros fatores contém relações fracas ou moderadas com as outras variáveis dificultando a interpretação e a importancia de cada fator. Nesta situação, será utilizado a rotação dos fatores Varimax pois melhora a interpretação da relação entre os fatores e acentua a relação das variáveis com apenas um fator.

Tabela 7: Cargas Fatoriais - Antes da Rotação

Variavel	*Fator*1	*Fator*2	*Fator*3	*Fator*4
DIVISAO_FORNECEDOR	**0,909**	−0,253	−0,147	0,157
VLR_LIMITE_CRE	0,292	0,463	**0,584**	0,105
FREQUENCIA	**0,954**	−0,211	−0,106	0,122
DIAS_COMPRA	**0,953**	−0,187	−0,096	0,101
TOTAL_PEDIDOS	**0,958**	−0,174	−0,088	0,085
QTDITENS	**0,730**	0,213	0,183	−0,319
QTD_CATEG	**0,924**	−0,241	−0,145	0,142
FAT_TOTAL	**0,789**	0,251	0,268	−0,389
VLR_IMPOSTO	**0,768**	0,215	0,244	−0,296
NUM_DEVOLUCAO	0,330	**0,638**	**−0,477**	0,142
VLR_DEVOLUCAO	0,143	**0,715**	**−0,502**	0,08
ANOS_CADASTRO	0,149	0,224	**0,476**	**0,745**

Outra medida importante da Análise Fatorial são as comunalidades de cada variável que representam a proporção de variância explicada pelos fatores comuns em uma variável. As comunalidades de cada variável são apresentadas na Tabela 8 e todas as comunalidades apresentaram um valor aceitável acima de 0,5 sugerido por Hair [6], o que sugere um bom ajuste do modelo nos nossos dados.

Outro aspecto importante na interpretação das comunalidades é entender quais as variáveis que mais impactam no modelo fatorial. Os maiores valores pertence a frequência de compra, a quantidade de pedidos e o tipo de itens comprados pelo o cliente são as variáveis que apresentam as maiores comunalidades, indicando que elas são as variáveis de maior importância nesse estudo, pois elas apresentam a maior porção da variância compartilhada com todas as outras variáveis consideradas.

Tabela 8: Comunalidades

Variavel	*Comunalidades*
DIVISAO_FORNECEDOR	0,912
VLR_LIMITE_CRE	0,640
FREQUENCIA	0,966
DIAS_COMPRA	0,953
TOTAL_PEDIDOS	0,956
QTDITENS	0,612
QTD_CATEG	0,933
FAT_TOTAL	0,758
VLR_IMPOSTO	0,696
NUM_DEVOLUCAO	0,744
VLR_DEVOLUCAO	0,783
ANOS_CADASTRO	0,599

Assim que a matriz fatorial de carga já tenha sido calculada pelo método de componentes principais, o processo de interpretação prossegue com o exame da matriz rotacionada para

detectar as cargas significantes e comunalidades adequadas. Para o critério de rotação usado para análise foi utilizado a rotação ortogonal Varimax, pois o objetivo é verificar se existe alguma diferença do comportamento das variáveis nos distintos fatores.

Os resultados da rotação Varimax são apresentados logo abaixo na tabela 9. Percebe-se que as variáveis que se relacionam com o primeiro fator estão ligadas com as características do pedido como a quantidade de pedidos e itens, o valor da compra e frequência de compra, o segundo fator está associado a capacidade ou volume de compra do comerciante, o quanto ele pode comprar e o quanto foi comprado, o terceiro fator representa a quantidade e o valor das devoluções do cliente, e por fim o quarto fator representa a fidelidade do cliente, quanto maior o tempo de casa, maior o limite de credito, uma relação importante que define a fidelidade do comerciante.

Tabela 9: Cargas Fatoriais - Depois da Rotação Varimax

$Variavel$	$Fator1$	$Fator2$	$Fator3$	$Fator4$
$DIVISAO\_FORNECEDOR$	$\mathbf{0,944}$	$0,202$	$0,053$	$0,033$
VLR_LIMITE_CRE	$-0,057$	$\mathbf{0,546}$	$0,059$	$\mathbf{0,589}$
$FREQUENCIA$	$\mathbf{0,947}$	$0,282$	$0,061$	$0,048$
$DIAS\_COMPRA$	$\mathbf{0,928}$	$0,308$	$0,070$	$0,045$
$TOTAL\_PEDIDOS$	$\mathbf{0,921}$	$0,329$	$0,074$	$0,043$
$QTDITENS$	$0,400$	$\mathbf{0,735}$	$0,113$	$0,001$
QTD_CATEG	$\mathbf{0,948}$	$0,224$	$0,060$	$0,029$
FAT_TOTAL	$0,396$	$\mathbf{0,863}$	$0,089$	$0,014$
$VLR\_IMPOSTO$	$0,423$	$\mathbf{0,770}$	$0,087$	$0,056$
$NUM\_DEVOLUCAO$	$0,173$	$0,108$	$\mathbf{0,849}$	$0,050$
$VLR\_DEVOLUCAO$	$-0,026$	$0,068$	$\mathbf{0,885}$	$-0,004$
$ANOS\_CADASTRO$	$0,112$	$-0,055$	$0,013$	$\mathbf{0,916}$

4 Conclusão

Conclui-se com esse trabalho que o uso de técnicas multivariadas como a Análise Fatorial pode auxiliar na redução da dimensão dos dados, pois é possível agrupar variáveis mais correlacionadas, fazendo com que a perda da informações seja a menor possível.

Com os resultados apresentados neste estudo foi possível identificar as variáveis mais representativas no comportamento de compra. Baseado em diversos critério da literatura foi indicado a extração de 4 fatores para a interpretação das cargas fatoriais de cada variável na composição dos fatores, formando assim, quatro grupos distintos de variáveis onde elas mais se relacionam.

Em aplicações de análises futuras sugere utilizar outras possíveis variáveis que possam contribuir mais para o modelo. Outra sugestão é refazer a Análise Fatorial separado por regiões ou tipo de atividade comercial do varejista. Assim, os grupos de comerciantes seriam os homogêneos possível e podendo melhor o resultados da análise.

Referências

[1] ANDERSON, T.W. (1984). *An Introduction to Multivariate Statistical Analysis* New York: John Wiley & Sons, volume (2).

[2] BARROSO, L. P., ARTES, R. (2003). Análise Multivariada. Lavras: Região Brasileira da Sociedade Internacional de Biometria, volume (1), 150.

[3] BARTLETT, M. S. (1937). *Properties of sufficiency and statistical tests.* Proceedings of the Royal Statistical Society, 160, 268-282.

[4] COOLEY, W. W., LOHNES, P. R. (1971). *Multivariate data analysis.* New York, John Wiley & Sons, 364.

[5] FACHEL, J. M. G. (1976). Análise Fatorial. São Paulo, USP, 81.

[6] HAIR, J. F., ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. (2009). Análise multivariada de dados. Porto Alegre: Bookman, volume (6), 100-148.

[7] HOTELLING, H. (1933). *Analysis of a Complex of Statistical Variables into Principal Components.* Journal of Educational Psychology, 24, 498-520.

[8] JOHNSON, R. A., WICHERN, D. W. (1998). *Applied multivariate statistical analysis.* Upper Saddle River, NJ: Prentice Hall, 352.

[9] KAISER, H. F. (1958). *The applicantion of electronic computers to factor analysis.* Educational and Psychological Measurement, 20, 141-151.

[10] KENDALL, M. G., BUCKLAND, W. R. (1957). *A Dictionary of Statistical Terms.* London: Longman Group, 121.

[11] LAWLEY, D.N. (1940). *The estimation of factor loadings by the method of maximum likelihood.* Proceedings of the Royal Society of Edinburgh, 60, 64-82.

[12] MARRIOT, F.H.C. (1974). *The interpretation of multiple observations.* New York, Academic Press, 117.

[13] MENEZES, A.C.F., FAISSOL, S.; FERREIRA, M.L. (1978). Análise da matriz geográfica: estruturas e inter-relações. Tendências atuais na geografia urbano-regional. Rio de Janeiro, 67-109.

[14] PEARSON, K., FILON, L. N. G. (1898). *On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation.* Chicago Statistical University, 19, 229-311.

[15] PEARSON, K. (1901). *On lines and planes of closest fit.* Philosophical Magazine, 6, 559–572.

[16] PEARSON, K., MAUL, M. (1927). *The Sampling Errors in the Theory of a Generalized Factor.* Biometrika, 19, 246-292.

[17] SOUKI, O. (2006). As 7 chaves da fidelização de clientes. São Paulo, 29, 245.

[18] SPEARMAN. (1904). *General Intelligence Objectively Determined and Measured.* The American Journal of Psychology, 2, 201–292.

[19] VELICER, W. F., JACKSON, D. N. (1982). *A comparison of component and factor patterns: A Monte Carlo approach.* Multivariate Behavioral Research, 17, 371-388.

[20] ZWICK, WILLIAM R., WAYNE F. (1986). *Comparison of five rules for determining the number of components to retain.* Psychological Bulletin, 99, 432-42.