24 de setembro de 2023

ANÁLISE FATORIAL EXPLORATÓRIA

ME731 - ANÁLISE MULTIVARIADA

Ana Julia Cunha e Silva - RA: 236038

Súmario

Lista de Tabelas

Lista de Figuras

1 Introdução

Este relatório tem como propósito apresentar, com certo nível de detalhe teórico a tecnica multivariada Análise Fatorial Exploratória. Além de apresentar, este relatório irá fazer uma abordagem prática exemplificada com um banco de dados e o software RStudio, além de descrever e discutir vantagens e desvantagens da técnica em questão. A estrutura do documento se resume em, análise teórica, análise prática, dicussão sobre as vantagens e desvantagens do método e por fim conclusões sobre os resultados da análise prática e conclusão geral do método.

No Rstudio foram utilizados os pacotes *psych* e *corrplot* para fazer a análise prática, além do banco de dados que consiste na tabela 1.9 de [1] porém disponibilizado no google drive, sendo assim necessário acesso a internet para rodar o arquivo da linguagem r. Todos os valores utilizados para fazer as tabelas da parte prática foram arrendodados com 2 casas decimais.

2 Análise Fatorial: Teoria

Por volta do começo do século XX, Karl Pearson, Charles Spearman e outros interessados em definir e medir inteligência deram origem à análise fatorial, conforme dito em [1]. Desde então, a técnica continou sendo aplicada nas áreas de ciências humanas e em algumas outras, para medir variaveis não observáveis.

As variáveis não observaveis, ou variáveis latentes, são aquelas que não podemos mensurar diretamente, precisamos observar outras variáveis para então relaciona-las, por meio de um modelo matemático(um *modelo linear*), a ponto que seja possível inferir a variável latente de interesse. Por exemplo, a personalidade é uma variável que não conseguimos observar, em [2] temos o exemplo do nível socioeconômico. "Para mensurá-lo, fazemos perguntas sobre questões como ocupação e escolaridade dos pais, bens domésticos, bens culturais e recursos educacionais da casa"disseram MATOS e RODRIGUES em [2]. Essencialmente, um fator(variável latente) é uma combinação linear de algumas variáveis observadas, que não é possível ser medida direto. Será usado o termo fator para se referir à essas variáveis não observaveis.

Existem algumas métricas importantes relacionadas aos fatores, uma delas é a *Carga Fatorial*. A carga fatorial de uma variável nada mais é do que correlação dessa variável com os fatores de interesse, sendo que para fatores diferentes ela provavelmente terá valores diferentes. Como é a correlação da variável com o fator, se a carga for positiva a variável estará positivamente correlacionada ao fator e idem para o caso de correlação negativo, assim como explicado em [2]. Como dito anteriormente, um fator pode ser definido por um *modelo linear*, sendo que para isso usamos as cargas fatoriais que chamaremos de λ_i a carga fatorial da variável no i-ésimo fator(X_i).

Além da carga fatorial, ainda tem outra métrica interessante e importante na análise fatorial, o *Escore Fatorial*, que resumidamente um escore fatorial é a estimativa de um fator. Abordando por cima, em [2] é citado duas formas de calcular o escore, uma fazendo a média ponderada das variáveis observaveis com os pesos sendo as cargas fatoriais respectivas. A outra forma produz escores com a distribuição normal padrão($N(0, 1)$).

Existem dois tipos de análise fatorial, a exploratória e a confirmatória. A Análise Fatorial Exploratória é usada quando não se tem noção de quantidade dos fatores, pouco conhecimento dos dados e não tem hipóteses iniciais, assim ela é usada para formular essas hipóteses e encontrar os fatores.

Já a Análise Fatorial Confirmatória é usada quando já temos as hipóteses iniciais e bom conhecimento do banco de dados, portanto, assim como o nome mesmo já diz, esse método busca confirmar as hipóteses. O segundo método é bem mais complicado e rígido que o primeiro, visto que seu propósito é confirmar/validar hipóteses, é preciso ter muito conhecimento sobre os dados, como foram colhidos e a quantidade de fatores existentes. Por isso nesse trabalho, exploraremos a Análise Fatorial Exploratória.

2.1 Modelo Fatorial Ortogonal

Seguindo [4], seja $\mathbf{X} = (x_1, x_2, ..., x_p)'$ um vetor $(p \times 1)$ com as p-variávies observaveis $x_i,\ i = 1, ..., p$ e com $E(\mathbf{X}) = \mu$ e matriz de covariâncias $\mathbf{Var}(\mathbf{X}) = \sum$, $\mathbf{F} = (f_1, f_2, ..., f_k)'$ o vetor $(k \times 1)$ dos $f_j,\ j = 1, ..., k$ fatores comuns com $k < p$, $\mathbf{U}$ é o vetor de fatores únicos e $\mathbf{Q}$ a matriz $(p \times k)$ de cargas fatoriais, onde o elemento q_{ij} a carga fatorial da i-ésima variável no j-ésimo fator. Portanto o modelo fatorial ortogonal é:

$$x_i = \sum_{j=1}^{k} q_{ij} f_j + u_i + \mu_i \tag{1}$$

Em notação matricial temos:

$$\mathbf{X} = \mathbf{U} + \mathbf{QF} + \mu \tag{2}$$

Com as seguintes suposições:
(i) μ_i é a média da i-ésima variável
(ii) $\mathbf{F}$ e $\mathbf{U}$ são vetores aleatórios, não observaveis e não-correlacionados
(iii) $E(\mathbf{F}) = 0,\ Var(\mathbf{F}) = \mathbf{I}_k$, onde $\mathbf{I}_k$ é a matriz identidade $(k \times k)$
(iv) $Cov(u_l, u_g) = 0,\ l \neq g$, $Cov(\mathbf{U}, \mathbf{F}) = 0$ e $Cov(\mu, \mathbf{F}) = 0$

$(v)\ Var(\mathbf{U}) = \Psi = \begin{bmatrix} \psi_{11} & & \\ & \ddots & \\ & & \psi_{pp} \end{bmatrix}$

Além disso, da equação (1) e das suposições acima, conseguimos chegar em:

$$Var(x_i) = \sum_{j=1}^{k} q_{ij}^2 + \psi_{ii} \tag{3}$$

Onde $h_i^2 = \sum_{j=1}^{k} q_{ij}^2$ é chamado de comunalidade e ψ_{ii} é a variância específica.

Além disso o modelo fatorial ortogonal tem algumas propriedades relacionadas a variância de $\mathbf{X}$ e covariância das matrizes:

- $Var(\mathbf{X}) = \sum = \mathbf{QQ}' + \Psi$, onde Ψ foi definido na suposição (v)
- $Cov(\mathbf{X}, \mathbf{F}) = Cov(\mathbf{U} + \mathbf{QF} + \mu, \mathbf{F})$
  Como $\mathbf{Q}$ é constante, temos:
  $\Rightarrow\ Cov(\mathbf{U}, \mathbf{F}) + Cov(\mu, \mathbf{F}) + \mathbf{Q}Cov(\mathbf{F}, \mathbf{F}) = 0 + 0 + \mathbf{Q}Var(\mathbf{F}) = \mathbf{QI}_k = \mathbf{Q}$

2.2 Estimação e Máxima Verossimilhança

Na estimação, do modelo fatorial, temos interesse em estimar duas matrizes, a matriz com as cargas fatoriais($\mathbf{Q}$) e a matriz diagonal(Ψ) com a variância específica, levando em consideração principalmente a propriedade $Var(\mathbf{X}) = \mathbf{QQ}' + \Psi$ e definindo $\mathbf{S} = \hat{\mathbf{Q}}\hat{\mathbf{Q}}' + \hat{\Psi}$, assim estaremos estimando $\sum$ por meio de $\mathbf{S}$. Existem várias formas de estimar esses valores, uma delas é por Análise de Componentes Principais e outra, que será abordada, é por meio da *máxima verossimilhança*.

Para estimação das cargas fatoriais($\mathbf{Q}$) e da variância dos fatores únicos(Ψ) a notação será baseada em [4] e *tr()* corresponde ao traço da matriz dentro do parentêses. Se $\mathbf{X} \sim N_p(\mu, \sum)$, onde N_p é uma normal p-variada, conseguimos estimar $\hat{\mathbf{Q}}$ e $\hat{\mathbf{\Psi}}$ pelo método de máxima verossimilhança. Como $\mu = \bar{\mathbf{X}}$, sendo $\bar{\mathbf{X}}$ a média amostral e a matriz inversa de $\sum = \sigma$, temos que a *log-verossimilhança* pode ser definida por:

$$l(\mathbf{X}; \sum) = -\frac{n}{2}[log|2\pi \sum| + tr(\sigma \mathbf{S})] \tag{4}$$

Assim, substituindo $\sigma = (\mathbf{QQ}' + \Psi)^{-1}$ e $\sum = \mathbf{QQ}' + \Psi$ temos que a *log-verossimilhança* é proporcional e por sua vez pode ser definida por:

$$l(\mathbf{X}; \mathbf{Q}, \Psi) = -\frac{n}{2}[log|2\pi(\mathbf{QQ}' + \Psi)| + tr((\mathbf{QQ}' + \Psi)^{-1}\mathbf{S})] \tag{5}$$

2.3 Rotação de Fatores

Para interpretar os fatores, é possivel fazer gráficos de pontos, onde a quantidade de fatores é a quantidade de eixos então se 2 fatores, teremos uma visão em $\mathbb{R}^2$, para k-fatores teremos que visualizar em $\mathbb{R}^k$. Com isso a visão gráfica para mais de 2 fatores começa a ficar complicada. Entretanto para 2 fatores, podemos rotaciona-los até termos uma visualização melhor, pois nem sempre a posição cartesiana é a melhor para visualizar os fatores. Abaixo um exemplo de como rotacionar os fatores pode auxiliar na visualização.

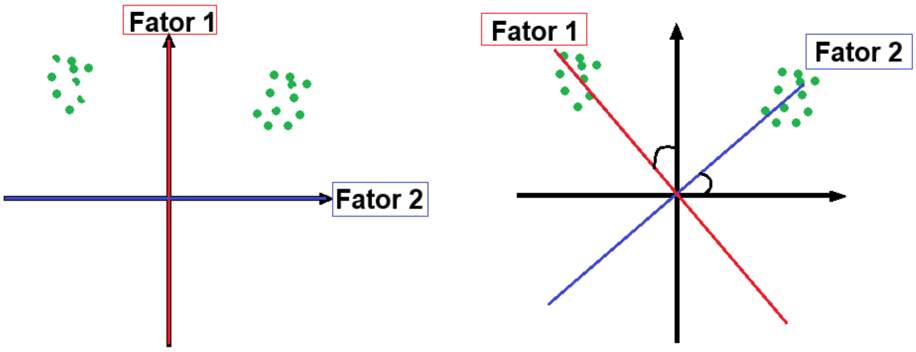

Figura 1: Exemplo Visual de uma Rotação Fatorial

Trabalhando com o caso de 2 fatores, uma das técnicas de rotação possível e explicada em [4] é a *Varimax*, que busca simplificar a visualização dos fatores encontrando o θ que maximize a variância dos quadrados das cargas em cada coluna da matriz de cargas rotacionada($\hat{\mathbf{Q}}^*$) que será abordada abaixo.

Assim, seja $\mathbf{G}(\theta)$ a matriz de rotação horária com ângulo θ, $\mathbf{G}(\theta) = \begin{bmatrix} cos(\theta) & sen(\theta) \\ -sen(\theta) & cos(\theta) \end{bmatrix}$

A matriz de cargas fatoriais rotacionadas é dada por $\hat{\mathbf{Q}}^* = \hat{\mathbf{Q}}\mathbf{G}(\theta)$.

Por fim, para escolher qual o melhor θ para aplicar a técnica, usamos a seguinte formula e maximizamos ela, então encontraremos θ ideal:

$$\frac{1}{p}\sum_{j=1}^{k}[\sum_{i=1}^{p}(\frac{\hat{q}^*_{ij}}{\hat{h}^*_i})^4 - (\frac{1}{p}\sum_{j=1}^{p}(\frac{\hat{q}^*_{ij}}{\hat{h}^*_i})^2)^2] \tag{6}$$

Onde $\frac{\hat{q}^*_{ij}}{\hat{h}^*_i}$ é a estimativa de q^*_{ij} que é a carga fatorial rotacionada da i-ésima variável no j-ésimo fator.

3 Vantagens

Uma das vantagens mais importantes da análise fatorial é a **redução da dimensionalidade**, onde reduzimos as várias variáveis em poucos fatores **simplificando o estudo estatístico**.

Na análise fatorial, as variáveis são explicadas pelos fatores, sendo essa característica o que podemos chamar de **representatividade dos dados**. Algo que não ocorre na análise de componentes principais, onde as variáveis explicam os componentes, sendo assim não é possível explicar que tal comportamento de tal variavel ocorre por causa do componente, mas é possível falar que a variável tal tem o respectivo comportamento por causa do fator na análise fatorial.

4 Desvantagens

Uma grande desvantagem do método é o **tamanho do banco de dados**, normalmente recomenda-se ter no mínimo uma quantidade de dados 5 vezes maior que a quatidade de variáveis no banco, sendo assim o método precisa de muitos dados para ser bem utilizado.

A **dificuldade mais elevada**, tanto matemática quanto computacional, para aplicação do método em relação à análise de componentes principais.

O modelo é muito sensível a mudanças na rotação, podendo se tornar uma desvantagem pois dependendo da rotação feita a interpretação pode variar bastante. Sendo assim a **interpretação** é um tanto quanto **subjetiva** e facilmente influênciável.

5 Análise Fatorial Exploratória: Aplicação Prática

O banco de dados utilizado é a tabela 1.9 de [1], que conta com os recordes das provas de corrida na modalidade feminina de 55 países diferentes. Sendo assim, o banco conta com 55 observações, que apesar de não ser o ideal para a aplicação do método, ainda é valido e 8 varíaveis sendo elas:

- **COUNTRY**: O país em questão.
- **X100m**: Recorde na prova de 100 metros em segundos.
- **X200m**: Recorde na prova de 200 metros em segundos.
- **X400m**: Recorde na prova de 400 metros em segundos.
- **X800m**: Recorde na prova de 800 metros em minutos.
- **X1500m**: Recorde na prova de 1500 metros em minutos.
- **X3000m**: Recorde na prova de 3000 metros em minutos.
- **Marathon**: Recorde na prova de maratona em minutos.

Para a análise trabalharemos com todas as váriavies exceto **COUNTRY**, pois é a única variável que não é numérica. Começando fazendo uma breve análise das estatísticas básicas encontramos os seguintes dados:

Tabela 1: Estatísticas Básicas dos Dados

Variável	**Média**	**DesvPadrão**	**Mínimo**	**Máximo**
X100m	11,62	0,45	10,79	12,90
X200m	23,58	1,15	21,52	27,10
X400m	53,61	2,99	47,99	63,60
X800m	2,08	0,11	1,89	2,33
X1500m	4,40	0,40	3,87	5,81
X3000m	9,36	1,11	3,92	13,04
Maratona	173,25	30,46	142,72	306,00

Na tabela acima nota-se que os valores médios das provas mais curtas(100, 200 e 400 metros) estão com valores muito maiores do que as provas mais longas, isso porquê a unidade de medida é diferente. Enquanto nas primeiras provas a medição é em segundos, as demais provas tem medição em minutos.

Pelo correlograma abaixo não conseguimos encontrar nenhum padrão ou tirar qualquer conclusão muito mais profunda do que a correlação entre as variáveis é estritamente positiva. Logo para análises mais profundas faremos o teste de adequação *Kaiser-Meyer-Olkin(KMO)* e o teste de esfericidade de *Bartlett* para confirmar a necessidade de uma análise fatorial.

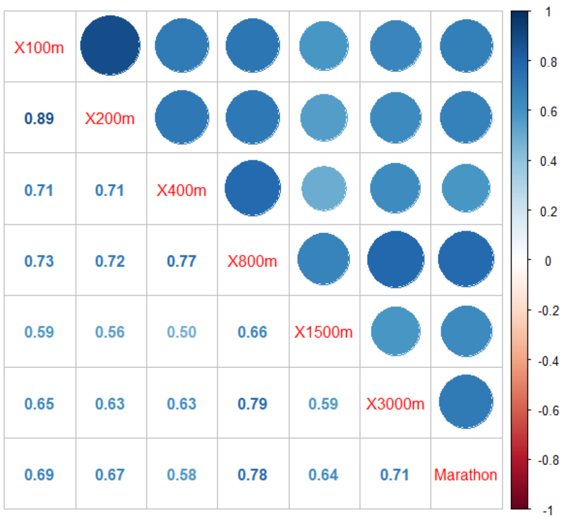

Figura 2: Correlograma dos Dados

5.1 Testes

O **teste de Bartlett** tem como hipótese nula a matriz de correlação dos dados ser igual a matriz identidade, portanto se o p-valor for muito baixo, poderemos rejeitar a hipótese nula e teremos uma indicação de que seja recomendável fazer uma análise fatorial nos dados. Ao aplicar o teste, obtivemos um p-valor da ordem de 10^{-54}, sendo assim muito pequeno($4,43 \cdot 10^{-54}$ mais precisamente) e definindo $p < 0,05$, podemos rejeitar a hipótese nula.

Por fim, para confirmar que o método utilizado deve ser a análise fatorial, faremos o **teste KMO**. Esse teste tem como propósito checar se a análise fatorial é apropriada ou não, para isso ele calcula dois tipos de resposta, o primeiro tipo é o valor global chamado de *Overall MSA* que representa todas as váriaveis dos dados, o segundo tipo é o valor especifico de cada variável chamado de *MSA for each item*. A seguir os valores específicos(MSA for each item) do teste KMO:

Tabela 2: Teste KMO - Resultados Específicos

Valor Específico						
X100m	**X200m**	**X400m**	**X800m**	**X1500m**	**X3000m**	**Maratona**
0,85	0,85	0,90	0,86	0,95	0,94	0,92

O resultado global foi $Overall\ MSA = 0,89$ sendo assim, com base no padrão escolhido temos mais uma evidência de que esses dados são recomendados para uma análise fatorial, visto que todos os valores são proximmos de 1,0 e maiores que 0,8 o que supera o valor minímo(0,7) recomendado no padrão. Foi escolhido seguir o padrão de [2] onde os valores(tanto global quanto específicos) recomendam a análise fatorial quando são maiores que 0,7 e não recomendam quando são menores que 0,5.

5.2 Modelos Fatoriais

Para descobrir quantos fatores seriam necessários, aplicou-se a função de componentes principais usando a matriz de correlações invés da matriz de covariâncias, para poder se adaptar a análise com 6 fatores e sem rotação, apenas para conseguir plotar o scree plot(gráfico scree).

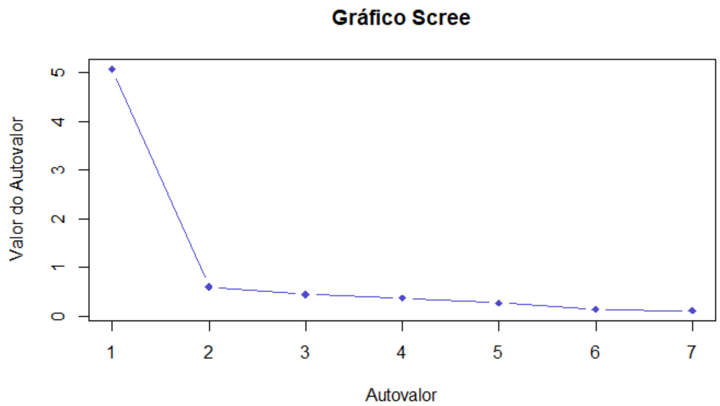

Figura 3: Gráfico Scree

Tabela 3: Valores dos Autovalores para 6 Fatores

Valor 1	**Valor 2**	**Valor 3**	**Valor 4**	**Valor 5**	**Valor 6**	**Valor 7**
5.07	0,60	0,44	0,37	0,27	0,14	0,11

Pela regra de Kraiser escolhemos a quantidade de fatores equivalente a quantidade de autovalores com valor maior que 1, entretanto se fisessemos isso para estes dados teriamos apenas um fator. Entretanto afim de mostrar como o método funciona com mais de um fator, usaremos dois, até porquê os dois primeiros fatores explicam

$\frac{5,067}{7} \cdot 100\% = 72,39\%$ e $\frac{0.086}{7} \cdot 100\% = 8,6\%$ dos dados separadamente, juntos explicam pouco mais de 80% dos dados. O calculo do percentual que cada fator explica dos dados é calculado pela razão entre o valor do autovalor pela quantiddade de variáveis no modelo.

Com a quantidade de fatores definida, foi utilizada a função própria para análise fatorial no R para enfim serem calculadas as cargas fatoriais e finalizar o estudo.

Foram feitos dois modelos, o primeiro sem nenhum método de rotação e o segundo com o método *varimax*. Considere **MR1** o Fator 1 e **MR2** o Fator 2.

Tabela 4: Cargas Fatoriais Com e Sem Rotação

Sem Rotação			**Com Rotação**		
Variável	**Fator 1**	**Fator2**	**Variável**	**Fator 1**	**Fator2**
X100m	0,886	-0,298	X100m	0,470	0,808
X200m	0,881	-0,368	X200m	0,420	0,857
X400m	0,783		X400m	0,557	0,552
X800m	0,920	0,232	X800m	0,845	0,432
X1500m	0,696	0,151	X1500m	0,623	0,345
X3000m	0,808	0,209	X3000m	0,745	0,375
Maratona	0,821	0,169	Maratona	0,729	0,414

Analisando primeiro o modelo, o sem rotação, não conseguimos enchergar motivo para fazer uma análise com dois fatores visto que um fator sozinho explicava 69% da variância dos dados e o segundo fator explicava apenas $5,4\%$. Olhando a figura 4 tamém não conseguimos enchergar motivos para um segundo fator, é notável que as variáveis 1 e 2 estão um pouco mais distantes das outras, porém nada alarmante. Na figura 5 todos as varíavies tem cargas altas próximas de 1, o que apenas favorece a ideia de ter apenas 1 fator.

Entretanto ao olhar na tabela nota-se algo curioso, as duas primeiras variáveis(X100m e X200m) tiveram cargas negativas no segundo fator e a terceira variável(X400m) sem nenhum valor, o que pode ser um indicativo de que uma rotação seja necessária.

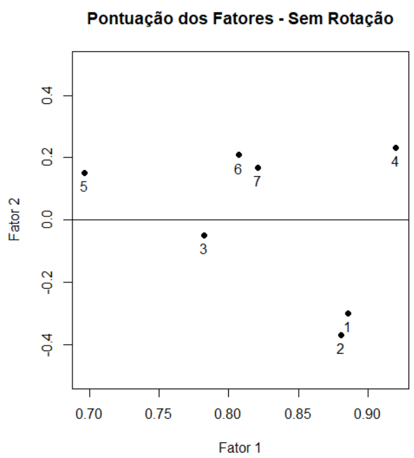

Figura 4: Pontuação dos Fatores - Sem Rotação

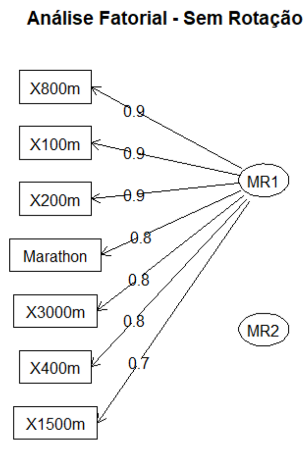

Figura 5: Fatores - Sem Rotação

Após aplicar a rotação varimax, nota-se que o segundo fator começa a fazer sentido, tanto na tabela quanto nos gráficos. As variáveis 1 e 2 tem cargas altas no segundo fator e baixas no primeiro, porém a variável 3 teve carga muito parecida em ambos os fatores, sendo suavemente maior no primeiro fator. Sendo assim, com essa rotação temos uma outra interpretação dos dados porém como citado sobre a variável 3 talvez outro tipo de rotação seja melhor para interpretação dos fatores latentes.

Nas figuras conseguimos visualizar melhor quais variáveis pertencem a quais fatores, na figura 6 conseguimos ver 2 grupos bem claros, apenas com a variável 3 um pouco distante das outras componentes de seu grupo.

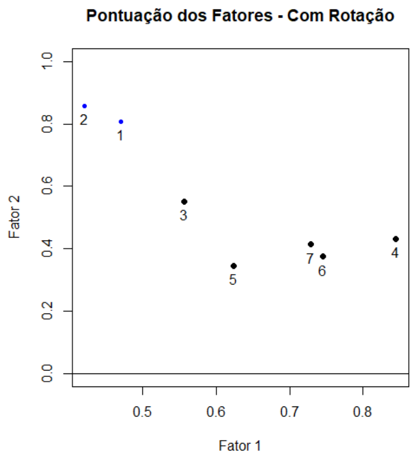

Figura 6: Pontuação dos Fatores - Com Rotação

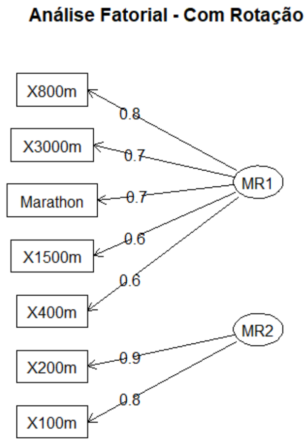

Figura 7: Fatores - Com Rotação

5.3 Conclusão Prática

Após a aplicação do método, nos dois modelos diferentes conseguimos ver a importância da rotação na interpretação. No modelo sem rotação temos que apenas um fator é necessário, ou seja apenas uma variável latente, o que é complicado formular alguma hipótese sobre o que seria essa variável latente.

Entretanto, quando rotacionamos os dados, encontramos dois fatores explicados pelos dados. O primeiro fator tem relação com as provas mais longas, como a maratona e prova de 3000 metros, sendo assim podemos dizer que o fator seria a resistência das atletas. Já o segundo fator tem relação só com as duas provas mais curtas, sendo assim poderiamos dizer que a variável latente em questão é a "explosão". Essas hipóteses vem ao pensar que provas mais longas requerem maior resistência(aguentar correr naquela velocidade por mais tempo) entquanto provas mais curtas requerem uma explosão de velocidade maio e mais atenção aos segundos que demora para concluir a prova.

6 Conclusão

O método de análise fatorial exploratória se mostrou um ótimo método de redução de dimensionalidade e formulação de hipóteses. Infelizmente não é perfeito, tem como principal desvantagem a necessidade de uma grande amostra ou banco de dados, pois a quantidade de dados é proporcional a quantidade de variáveis. Entretanto tem grande vantagem por manter a representatividade dessas variáveis no método. A interpretação é um ponto que pode ser tanto positivo, por sua facilidade, quanto negativo, por sua subjetividade.

Quanto ao exemplo prático, podemos concluir que é um método bem simples de ser implementado, com poucas linhas de código e muita interpretação gráfica. Tem dois tipos de testes que indicam se é uma abordagem apropriada ou não, e testes bem simples de compreender. Podemos notar o impacto que a rotação tem na interpretação da conclusão ou formulação de hipóteses e ainda podemos ver como a visualização gráfica é benéfica para poucos fatores.

7 Referências Bibliográficas

[1] Johnson, R. A. e Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Sexta Edição, Prentice Hall, Nova Jersey.

[2] Matos, D. A. S. e Rodrigues, E. C. (2019). Análise Fatorial. Primeira Edição, Enap, Brasília.

[3] Frost, J. (2022). Factor Analysis Guide with an Example. Link (Acessado em 17/09/2023).

[4] Härdle, W. K. e Simar, L. (2019). Applied Multivariate Statistical Analysis. Quinta Edição, Springer Nature.