Análise Multivariada Aplicada à Contabilidade

Prof. Dr. Marcelo Botelho da Costa Moraes
www.marcelobotelho.com
mbotelho@usp.br

Turma: 2º / 2016

Agenda – Aula 11/15

- Análise Fatorial
  - Uso
  - Modelagem
  - Formas de Análise
  - Extração e Análise dos Fatores
  - Caso prático

Análise Fatorial

Análise Fatorial

- Busca sintetizar as relações observadas entre um conjunto de variáveis inter-relacionadas

- Representar um conjunto de variáveis originais observadas por um meio de um menor número de fatores intrínsecos

- Redução dos dados em fatores

Análise Fatorial

- Descobrir e analisar a estrutura de um conjunto de variáveis inter-relacionadas, de modo a construir uma escala de medida para fatores (intrínsecos) que, de alguma forma (mais ou menos explícita), controla as variáveis originais (MAROCO, 2007)
- As variáveis que compõem um determinado fator devem ser altamente correlacionadas
- Objetivo de atribuir um escore a constructos (fatores)

Análise Fatorial

- Um fator é a combinação linear das variáveis originais

- Os fatores também representam as dimensões latentes (constructos) que resumem ou explicam o conjunto original de variáveis observadas

Análise Fatorial

Suposições existentes na Análise Fatorial

- Normalidade (multivariada) e linearidade
- Matriz de correlações com valores significativos (número substancial de valores superiores a 0,30)

Análise Fatorial

Prévia da Análise Fatorial

- Verificar viés e *outliers*
  - Afetam a variância, desvio padrão, covariância e correlação
- Amostra deve ser igual ou superior a 100 observações
  - Mínimo de 5 vezes mais observações do que o número de variáveis (recomendável 10 observações por variável)

Análise Fatorial

Tipos de Análise Fatorial

- Análise exploratória
  - O pesquisador tem pouco ou nenhum conhecimento prévio acerca da estrutura de fatores (foco da aula)
- Análise confirmatória
  - Caso particular de equações estruturais, já que o pesquisador possui algum conhecimento prévio sobre como as variáveis se comportam e relacionam, assim, assume que a estrutura de fatores é conhecida

Análise Fatorial

Etapas da Análise Fatorial

- Análise da matriz de correlações e adequação da utilização da análise fatorial
- Extração dos fatores iniciais e determinação do número de fatores
- Rotação dos fatores
- Interpretação dos fatores

Modelagem da Análise Fatorial

- Desenvolvido por Spearman em 1904
- Notou que as correlações das notas dos estudantes poderiam ser quantificadas de maneira mais simples
- Criou a hipótese de que o desempenho dos alunos em várias disciplinas são inter-relacionados, e essas inter-relações podem ser explicadas pelo nível de inteligência geral dos estudantes

Modelagem da Análise Fatorial

	Clássicos	Francês	Inglês	Matemática	Discriminação de tom	Música
Clássicos	1,00	0,83	0,78	0,70	0,66	0,63
Francês	0,83	1,00	0,67	0,67	0,65	0,57
Inglês	0,78	0,67	1,00	0,64	0,54	0,51
Matemática	0,70	0,67	0,64	1,00	0,45	0,51
Discriminação de tom	0,66	0,65	0,54	0,45	1,00	0,40
Música	0,63	0,57	0,51	0,51	0,40	1,00

Modelagem da Análise Fatorial

- Por exemplo, para a linha Clássicos e Inglês

$$\frac{0,83}{0,67} \approx \frac{0,70}{0,64} \approx \frac{0,66}{0,54} \approx \frac{0,63}{0,51} \approx 1,2$$

Modelagem da Análise Fatorial

- Spearman sugeriu que cada um dos seis testes de inteligência (variáveis) pudesse ser descrito como

$$X_i = a_i F + \varepsilon_i$$

- X_i é o i-ésimo escore da variável analisada depois de efetuada a padronização (média zero e desvio padrão 1 – *Z* scores)
- *F* é o fator aleatório comum a todas as variáveis (inteligência)
- ε_i é um componente aleatório específico para cada teste de inteligência
- a_i é a constante chamada de **carga fatorial** (**loading**), que mede a importância dos fatores na composição de cada variável (correlação)

Modelagem da Análise Fatorial

- A variância de X_i é dada por

$$Var(X_i) = Var(a_i F + \varepsilon_i)$$
$$= Var(a_i F\ ) + Var(\varepsilon_i)$$
$$= {a_i}^2 Var(F) + Var(\varepsilon_i)$$
$$= {a_i}^2 + Var(\varepsilon_i)$$

- Em que a_i é uma constante, F e ε_i são independentes e a variância de F é igual a 1. Como $Var(X_i) = 1$, tem-se

$$1 = {a_i}^2 + Var(\varepsilon_i)$$

- O quadrado da a_i (carga fatorial) representa a proporção da variância de X_i, que é explicada pelo fator comum (**comunalidade**)

Modelagem da Análise Fatorial

- Spearman defendia que a performance de uma criança em um teste qualquer podia ser obtida pela soma de um fator geral F com uma habilidade específica ε_i
- Generalizando a proposta, tem-se o modelo fatorial em que p variáveis observáveis (X_1, X_2, ..., X_p), extraídas de uma população com vetor de média μ e matriz de covariância Σ, são linearmente dependentes de algumas variáveis não observáveis F_1, F_2, ..., F_m, denominadas de fatores comuns e de p fontes adicionais de variação ε_1, ε_2, ..., ε_p, denominadas de erros ou fatores específicos

Modelagem da Análise Fatorial

- O modelo de análise fatores pode ser apresentado como

$$X_1 = \mu_1 + a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1$$
$$X_2 = \mu_2 + a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2$$
$$\dots$$
$$X_p = \mu_p + a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p$$

- Efetuando a padronização de *X*, o modelo fica

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \quad (i = 1, \dots, p)$$

Modelagem da Análise Fatorial

- O modelo anterior assume as seguintes premissas

1. Os fatores comuns (F_k) são independente (ortogonais) e igualmente distribuídos, com média 0 e variância 1 (k = 1, ..., m)
2. Os fatores específicos (ε_i) são independentes e igualmente distribuídos, com média zero e variância ψ_i (i = 1, ..., p)
3. F_k e ε_i são independentes

- O termo ψ_i representa a variância de ε_i, ou seja, $Var(\varepsilon_i) = \psi_i$

Modelagem da Análise Fatorial

- Se as três premissas anteriores forem verificadas, temos um modelo fatorial ortogonal
- Caso contrário, se F_k e ε_i estiverem correlacionados, o modelo fatorial será oblíquo
- Já os fatores podem ser estimados por combinação linear das variáveis

$$F_1 = d_{11}X_1 + d_{12}X_2 + \cdots + d_{1i}X_i$$
$$F_2 = d_{21}X_1 + d_{22}X_2 + \cdots + d_{2i}X_i$$
$$\dots$$
$$F_m = d_{m1}X_1 + d_{m2}X_2 + \cdots + d_{mi}X_i$$

Modelagem da Análise Fatorial

- Sendo F_m os fatores comuns, d_{mi} os coeficientes dos escores fatoriais e X_i as variáveis originais
- O **escore fatorial** resulta da multiplicação dos coeficientes d_{mi} pelo valor das variáveis originais
- Na existência de mais de um fator, o escore fatorial corresponderá às coordenadas da variável em relação aos eixos (fatores)

Modelagem da Análise Fatorial

- A variância será dada por

$$Var(X_i) = Var(a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i) = 1$$
$$= 1$$
$$= a_{i1}{}^2 Var(F_1) + a_{i2}{}^2 Var(F_2) + \cdots + a_{im}{}^2 Var(F_m) + \Psi_i$$
$$= a_{i1}{}^2 + a_{i2}{}^2 + \cdots + a_{im}{}^2 + \Psi_i$$

- Portanto, a variância pode ser decomposta em duas partes

$$Var(X_i) = \underbrace{a_{i1}{}^2 + a_{i2}{}^2 + \cdots + a_{im}{}^2}_{\textbf{comunalidade}} + \underbrace{\Psi_i}_{\textbf{variância específica}}$$

Modelagem da Análise Fatorial

- Sendo

$$h_i{}^2 = a_{i1}{}^2 + a_{i2}{}^2 + \cdots + a_{im}{}^2 + \Psi_i$$

- A comunalidade representa uma estimativa da variância de X_i que é explicada pelos fatores comuns
- ψ_i é chamada de especificidade de X_i, pois não está ligada ao fator comum
- A comunalidade é um índice de variabilidade total explicada por todos os fatores para cada variável

$$Var(X_i) h_i{}^2 + \Psi_i \; (i = 1, 2, \ldots, p)$$

Adequação da Utilização da Análise Fatorial

- Analisar a matriz de correlações

- Verificar a estatística KMO e o teste de esfericidade de Bartlett

- Analisar a matriz anti-imagem

Análise da Matriz de Correlações

- Pressuposto de correlações entre as variáveis
  - Verificar se existem valores significativos para justificar o emprego da técnica
  - Baixa correlação indica uso de outras técnicas
- Variáveis com alta correlação tendem a compartilhar o mesmo fator
- Matriz de correlação de Pearson
  - Se a matriz de correlações não revelar um número substancial de valores superiores a 0,30 há fortes indícios de que a utilização da técnica não é apropriada

KMO e Teste de Esfericidade de Bartlett

- O teste de esfericidade de Bartlett avalia a hipótese de que a matriz de correlações pode ser a matriz identidade com determinante igual a 1

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

- Se a matriz de correlações for igual à matriz identidade, isso significa que não devemos utilizar a análise fatorial
  - H_0: a matriz de correlações é uma matriz identidade

KMO e Teste de Esfericidade de Bartlett

- Uma estatística utilizada é a Kaiser-Meyer-Olkin (KMO) que compara as correlações simples com as correlações parciais

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

- r_{ij} = coeficiente de correlação entre variáveis
- a_{ij} = coeficiente de correlação parcial

KMO e Teste de Esfericidade de Bartlett

- A estatística KMO, cujos valores variam de 0 a 1, avalia a adequação da amostra quanto ao grau de correlação parcial entre os valores, que deve ser pequeno
- O valor de KMO próximo de 0 indica que a análise fatorial pode não ser adequada (correlação fraca entre as variáveis)
- Quanto mais próximo de 1 o seu valor, mais adequada é a utilização da técnica

KMO e Teste de Esfericidade de Bartlett

KMO	Análise Fatorial
1 – 0,9	Muito boa
0,8 – 0,9	Boa
0,7 – 0,8	Média
0,6 – 0,7	Razoável
0,5 – 0,6	Má
< 0,5	Inaceitável

Matriz Anti-Imagem

- A matriz de correlações anti-imagem contém os valores negativos das correlações parciais
- É uma forma de obter indícios sobre a necessidade de eliminação de determinada variável do modelo
- A Medida de Adequação da Amostra, ou *Measure of Sampling Adequacy* (MSA), para cada variável, de forma similar ao KMO

Matriz Anti-Imagem

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

- O pesquisador deve analisar os valores de MSA para cada variável individualmente e excluir as que se encontram no domínio inaceitável
- Quanto maiores os valores, melhor
- Se alguma variável apresentar baixo valor na diagonal principal e alto valor fora dela, talvez haja necessidade de excluí-la
- Vale lembrar que, por vezes, a baixa correlação de determinada variável com as demais não necessariamente implica em sua eliminação, podendo representar um fator isoladamente

Extração dos Fatores Iniciais

Método de Extração

- Análise dos Componentes Principais (ACP)
  - Considera a variância total dos dados
- Análise dos Fatores Comuns (AFC)
  - Os fatores são estimados com base na variância comum
- Variância
  - Comum (comunalidade)
  - Específica (variável individual)
  - Erro (fatores aleatórios)

Extração dos Fatores Iniciais

Análise dos Componentes Principais (ACP)

- Combinação linear das variáveis observadas, de maneira a maximizar a variância total explicada
- Se determinadas variáveis forem altamente correlacionadas, elas serão combinadas de modo a formar um fator que explicará a maior quantidade de variância na amostra
- O segundo componente terá a segunda maior quantidade de variância e não será correlacionado com o primeiro e, assim, sucessivamente

Extração dos Fatores Iniciais

- Se o objetivo é reduzir os dados para obtenção do mínimo número de fatores necessários para explicar o máximo da variância representada pelas variáveis originais, optar pela ACP
- Se o objetivo é identificar fatores ou dimensões latentes que reflitam o que as variáveis têm em comum, a AFC é mais apropriada

Extração dos Fatores Iniciais

- Além da ACP e AFC, temos
  - Máxima verossimilhança: indicado quando se trata de uma amostra de indivíduos retirados de uma população normal e se pretende explicar a estrutura latente da matriz de correlações
  - Mínimos quadrados ordinários e generalizados (OLS e GLS): objetivos semelhantes aos do método anterior
  - Alpha: parte do pressuposto de que as variáveis em estudo constituem uma amostra do universo de variáveis existentes e de que os indivíduos compõem toda a população

Extração dos Fatores Iniciais

Escolha do Número de Fatores

- Primeiro extrai a combinação linear que explica a maior parte da variância dos dados, em seguida, uma combinação que explique um montante de variância cada vez menor
- Necessidade de definir quantos fatores:
  - Critério da raiz latente (critério de Kaiser)
  - Critério *a priori*
  - Critério de percentagem da variância
  - Critério do gráfico Scree

Extração dos Fatores Iniciais

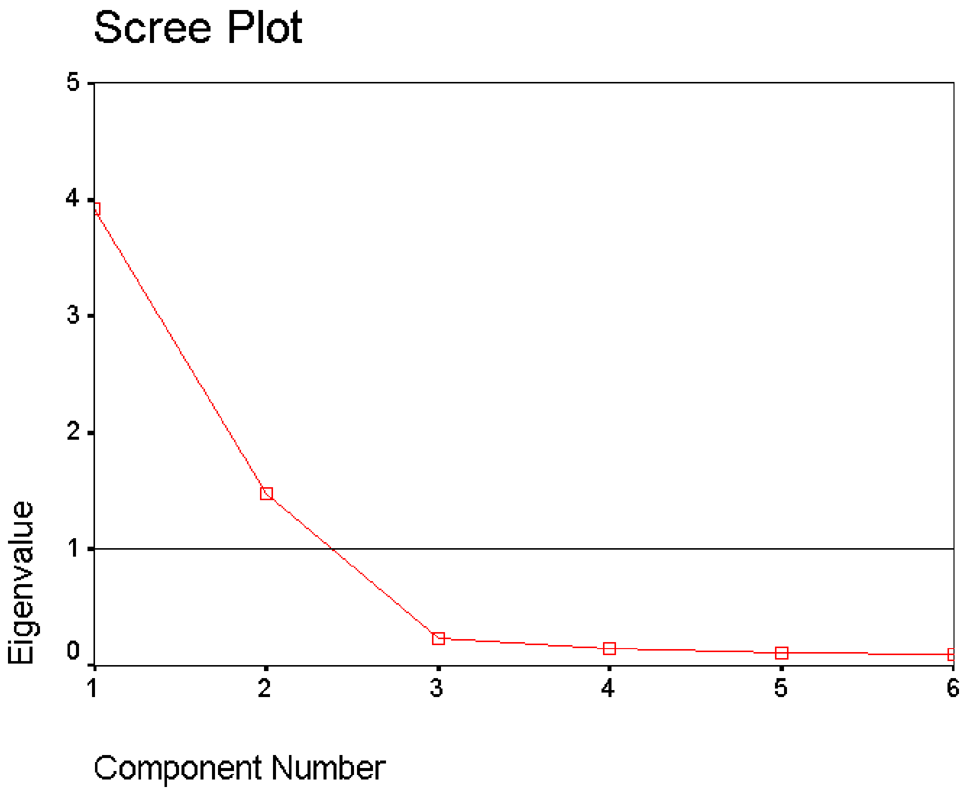
- Critério da raiz latente (critério de Kaiser)
  - Escolhe-se o número de fatores a reter, em função do número de valores próprios acima de 1
  - Os valores próprios, autovalores ou *eigenvalues*, são ordenador por dimensão
  - *Eigenvalues* mostram a variância explicada por cada fator, ou seja, o quanto cada fator explica da variância total
  - No método de extração de componentes principais, a soma dos valores próprios iguala o número de variáveis
  - A escolha dos componentes que apresentam *eigenvalues* maior que 1 decorre do fato de que, no mínimo, o componente deve explicar a variância de uma variável utilizada no modelo, uma vez que são variáveis padronizadas

Extração dos Fatores Iniciais

- Critério *a priori*
  - É o método mais simples, pois, neste caso o pesquisador já sabe quantos fatores extrair
- Critério de percentagem da variância
  - Consiste em escolher, como número de fatores, um número mínimo necessário para que o percentual de variância explicada alcance o nível satisfatório desejado (definido pelo pesquisador)
- Critério do gráfico Scree
  - Utilizado para identificar o número ótimo de fatores que podem ser extraídos antes que a quantia da variância única comece a dominar a estrutura da variância comum

Extração dos Fatores Iniciais

- Critério do gráfico Scree



Rotação dos Fatores

- Nem sempre os fatores produzidos na fase de extração são facilmente interpretados
- O método de rotação tem por objetivo transformar os coeficientes dos componentes principais retidos em uma estrutura simplificada
- Como as cargas fatoriais são pontos entre eixos (fatores), podemos girar os eixos sem alterar a distância entre os pontos (relação entre fator e variável)

Rotação dos Fatores

- Métodos de rotação ortogonais
  - **Varimax**: minimiza o número de variáveis que têm altas cargas em um fator, simplificando a interpretação dos fatores. Privilegia apenas alguns pesos significativos e todos os outros próximos de zero. É o mais utilizado
  - **Quartimax**: busca simplificar as linhas de uma matriz fatorial (número de fatores), tornando os pesos de cada variável elevados para um pequeno número de componentes, e próximos de zero todos os demais (minimiza o número de fatores para explicar uma variável)
  - **Equamax**: congrega características dos outros métodos, com objetivo de simplificar as linhas e colunas simultaneamente (fatores e variáveis)

Rotação dos Fatores

- Métodos de rotação oblíquas
  - No SPSS temos o **Direct Oblim** e o **Promax**
  - As comunalidades são preservadas, porém, os fatores gerados apresentam-se de forma mais fortemente correlacionadas

- Se o objetivo é reduzir o número de variáveis originais, independente da significância dos fatores resultantes, o método ortogonal é preferível

Rotação dos Fatores

- A matriz de componentes, após a rotação ortogonal, visa extremar os valores das cargas fatoriais (*loadings*), de modo que cada variável se associe a apenas um fator
- Variáveis com baixa carga fatorial devem ser eliminadas

Interpretação dos Fatores

- Devemos escolher quais cargas fatoriais devem ser consideradas
- Normalmente, considera-se apenas cargas fatorais acima de 0,30 (nível mínimo), cargas acima de 0,40 são consideradas mais importantes e, se forem maiores que 0,50 são consideradas estatisticamente significativa

Interpretação dos Fatores

- Para identificar cargas fatoriais significativas, ao nível de 5% de significância, com base no tamanho da amostra, temos (próximo slide)

Interpretação dos Fatores

Carga Fatorial	Tamanho da Amosta
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

Análise Fatorial

- Exemplo prático: Fatorial.xls
- Variáveis
  - Cod_Em*:* código da empresa*;*
  - PMRV: prazo médio de recebimento de vendas, em dias;
  - Endividamento: em %;
  - Vendas: em $ x mil;
  - Margem_líquida: margem líquida de vendas em %;

Obrigado pela Atenção!!!

Até a próxima aula

mbotelho@usp.br

www.marcelobotelho.com