MBA USP ESALQ

UNSUPERVISED MACHINE LEARNING: Análise Fatorial e PCA

Prof. Dr. Wilson Tarantin Junior

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

MBA USP ESALQ

Contextualização

- Quando aplicar a análise fatorial?

  - Quando as variáveis forem **métricas**: depende das correlações entre variáveis

  - Trata-se do **agrupamento das variáveis** em fatores. Os objetivos podem ser:

    - Obter o comportamento conjunto de variáveis, combinando-as para redução estrutural
    - Análise da validade de construtos pela identificação das variáveis alocadas aos fatores
    - Elaboração de rankings para classificação de desempenho por meio dos fatores
    - Criação de fatores ortogonais entre eles e posterior uso em modelos supervisionados

Contextualização

- Análise fatorial por componentes principais
  - **Componentes principais**: método de determinação dos fatores que se baseia na criação de fatores não correlacionados a partir da combinação linear das variáveis originais
- Análise fatorial PCA: modelo não supervisionado de *machine learning*
  - Portanto, a técnica não tem um caráter preditivo para observações que não estejam presentes na amostra. Se surgirem novas observações, novos fatores atualizados devem ser gerados

Implementação

Matriz de correlações

- Procedimento inicial
  - A PCA fundamenta-se na existência de correlações entre variáveis originais para a criação dos fatores
    - **Coeficiente de correlação de Pearson**: relação linear entre duas variáveis métricas
    - Coeficientes de correlação mais próximos dos valores extremos (-1; +1) propiciam a extração de um único fator → indicam existência de relação entre as variáveis
    - Coeficientes de correlação mais próximos de zero propiciam a extração de diferentes fatores → indicam que a relação entre as variáveis é (praticamente) inexistente

Matriz de correlações

- Procedimento inicial
  - A seguir, tem-se a representação da matriz de correlações para K variáveis e a expressão de cálculo do coeficiente de correlação de Pearson

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix}$$

$$\rho_{12} = \frac{\sum_{i=1}^{n} (X_{1i} - \overline{X}_1) \cdot (X_{2i} - \overline{X}_2)}{\sqrt{\sum_{i=1}^{n} (X_{1i} - \overline{X}_1)^2} \cdot \sqrt{\sum_{i=1}^{n} (X_{2i} - \overline{X}_2)^2}}$$

Fonte das equações: Fávero & Belfiore (2017, Capítulo 10)

Adequação global

- A extração de fatores é adequada?
    - Para que que a análise fatorial seja adequada, devem existir valores elevados (-1; +1) e estatisticamente significantes na matriz de correlações
    - Para investigar a adequação global da análise fatorial, vamos utilizar o **teste de esfericidade de Bartlett**
        - Os coeficientes de correlação de Pearson são estatisticamente diferentes de zero?

Adequação global

- Teste de esfericidade de Bartlett
  - Compara a matriz de correlações com a matriz identidade de mesma dimensão e espera-se que tais matrizes sejam diferentes para que a análise seja aplicável

$$H_0: \rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} = I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad H_1: \rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \neq I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$\chi^2_{\text{Bartlett}} = -\left[(n-1) - \left(\frac{2 \cdot k + 5}{6}\right)\right] \cdot \ln|D| \quad \text{com} \quad \frac{k \cdot (k-1)}{2} \quad \text{graus de liberdade}$$

Fonte das equações: Fávero & Belfiore (2017, Capítulo 10)

Autovalores e autovetores

- Autovalores
  - A matriz de correlações de dimensão K x K possui K autovalores (λ^2) e podem ser obtidos da seguinte forma:
    - Solução de $\det(\lambda^2 \cdot \mathrm{I} - \rho) = 0$ equivalente a $\begin{vmatrix} \lambda^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda^2 - 1 \end{vmatrix} = 0$
  - Os autovalores indicam o percentual da variância compartilhada pelas variáveis originais para a formação de cada fator

Fonte das equações: Fávero & Belfiore (2017, Capítulo 10)

Autovalores e autovetores

- Autovetores

  - Os autovetores da matriz de correlações são obtidos com base em cada um dos autovalores

  - $v_{1k}, v_{2k}, \dots, v_{kk}$ são os autovetores para o K-ésimo autovalor (λ^2) em análise

    - Solução de $\begin{pmatrix} \lambda_k^2-1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_k^2-1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_k^2-1 \end{pmatrix} \cdot \begin{pmatrix} v_{1k} \\ v_{2k} \\ \vdots \\ v_{kk} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ ou $\begin{cases} (\lambda_k^2-1)\cdot v_{1k} - \rho_{12}\cdot v_{2k} \ldots - \rho_{1k}\cdot v_{kk} = 0 \\ -\rho_{21}\cdot v_{1k} + (\lambda_2^2-1)\cdot v_{2k} \ldots - \rho_{2k}\cdot v_{kk} = 0 \\ \vdots \\ -\rho_{k1}\cdot v_{1k} - \rho_{k2}\cdot v_{2k} \ldots + (\lambda_k^2-1)\cdot v_{kk} = 0 \end{cases}$

Fonte das equações: Fávero & Belfiore (2017, Capítulo 10)

Obtenção dos fatores

- Identificação dos *scores* fatoriais
  - Após a análise fatorial ser considerada adequada pelos testes anteriores, será necessário criar os *scores* que geram os fatores propriamente ditos
  - ***Scores* fatoriais**: são os parâmetros que relacionam o fator com as variáveis originais, representados em um modelo linear
  - Para K variáveis originais, existem, no máximo, K fatores (F_1, F_2, ... , F_k)
  - Os *scores* vêm a partir dos autovalores e autovetores da matriz de correlações

Scores fatoriais

- Definindo os *scores*

  - A partir dos autovalores e autovetores, obtém-se os *scores* fatoriais s_1, s_2, ... , s_k
    São gerados K grupos de *scores* (é o limite máximo de K fatores possíveis)

$$\mathbf{S}_1 = \begin{pmatrix} s_{11} \\ s_{21} \\ \vdots \\ s_{k1} \end{pmatrix} = \begin{pmatrix} \frac{v_{11}}{\sqrt{\lambda_1^2}} \\ \frac{v_{21}}{\sqrt{\lambda_1^2}} \\ \vdots \\ \frac{v_{k1}}{\sqrt{\lambda_1^2}} \end{pmatrix} \quad \mathbf{S}_2 = \begin{pmatrix} s_{12} \\ s_{22} \\ \vdots \\ s_{k2} \end{pmatrix} = \begin{pmatrix} \frac{v_{12}}{\sqrt{\lambda_2^2}} \\ \frac{v_{22}}{\sqrt{\lambda_2^2}} \\ \vdots \\ \frac{v_{k2}}{\sqrt{\lambda_2^2}} \end{pmatrix} \quad \mathbf{S}_k = \begin{pmatrix} s_{1k} \\ s_{2k} \\ \vdots \\ s_{kk} \end{pmatrix} = \begin{pmatrix} \frac{v_{1k}}{\sqrt{\lambda_k^2}} \\ \frac{v_{2k}}{\sqrt{\lambda_k^2}} \\ \vdots \\ \frac{v_{kk}}{\sqrt{\lambda_k^2}} \end{pmatrix}$$

Fonte das equações: Fávero & Belfiore (2017, Capítulo 10)

Fatores

- Definindo os K fatores
  - O valor do fator F é obtido com as variáveis X transformadas pelo Z-Score (ZX)
  - Tais fatores são ortogonais entre si, ou seja, não são correlacionados

$$F_{1i} = \frac{v_{11}}{\sqrt{\lambda_1^2}} \cdot ZX_{1i} + \frac{v_{21}}{\sqrt{\lambda_1^2}} \cdot ZX_{2i} + \ldots + \frac{v_{k1}}{\sqrt{\lambda_1^2}} \cdot ZX_{ki}$$

$$F_{2i} = \frac{v_{12}}{\sqrt{\lambda_2^2}} \cdot ZX_{1i} + \frac{v_{22}}{\sqrt{\lambda_2^2}} \cdot ZX_{2i} + \ldots + \frac{v_{k2}}{\sqrt{\lambda_2^2}} \cdot ZX_{ki}$$

$$F_{ki} = \frac{v_{1k}}{\sqrt{\lambda_k^2}} \cdot ZX_{1i} + \frac{v_{2k}}{\sqrt{\lambda_k^2}} \cdot ZX_{2i} + \ldots + \frac{v_{kk}}{\sqrt{\lambda_k^2}} \cdot ZX_{ki}$$

Fonte das equações: Fávero & Belfiore (2017, Capítulo 10)

Escolha dos fatores

- Todos os K fatores serão utilizados?
  - Embora seja possível estabelecer a priori quantos fatores são desejados, é de fundamental importância realizar uma análise por meio dos autovalores
  - Lembrando: os autovalores indicam o percentual da variância compartilhada pelas variáveis originais para a formação de cada fator
  - Neste sentido, fatores formados a partir de autovalores menores do que 1 podem não ter representatividade. **O critério de Kaiser (ou critério da raiz latente) indica que sejam considerados apenas fatores correspondentes a autovalores > 1**

Cargas fatoriais

- Análise da composição dos fatores
  - As cargas fatoriais representam as correlações de Pearson entre os fatores e as variáveis originais
  - Pode ser interpretada como a importância de cada variável na constituição daquele fator em particular
    - Quanto maior a carga fatorial, mais aquele fator é influenciado pela variável

Comunalidades

- Composição dos fatores selecionados
  - Ao utilizar o critério da raiz latente, somente os fatores que são derivados de autovalores maiores que 1 serão considerados
  - Portanto, as comunalidades mostram a variância total compartilhada, para cada variável, em todos os fatores extraídos e selecionados com base no critério da raiz latente
  - É possível analisar se houve perda de variância, por variável, após a exclusão de fatores por meio do critério da raiz latente

Criação de rankings

- Soma ponderada e ordenamento

  - Para criar rankings a partir dos fatores obtidos utilizando o critério da soma ponderada e ordenamento, para cada observação da amostra, calcula-se:

    - $Resultado_i = (F_{1i} * \%\ var.\ comp.\ F_1) + (F_{2i} * \%\ var.\ comp.\ F_2) + ... + (F_{ki} * \%\ var.\ comp.\ F_k)$

  - Em resumo, multiplica-se o resultado obtido de cada fator por seu percentual de variância compartilhada e depois é realizado o ordenamento do resultado

Referência

Fávero, Luiz Paulo; Belfiore, Patrícia. (2017). Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Elsevier