

Análise Multivariada de Dados - Aula 07

Análise Fatorial

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

1. Conceitos que aprendemos em Aulas anteriores
2. A Ideia Geral
3. Procedimento para uma análise fatorial
4. Análise de fatores de componentes principais
5. Exemplos
6. Comentários Finais
7. Referências

Conceitos que aprendemos em Aulas anteriores

Nas últimas Aulas lidamos em aspectos teóricos e práticos da Análise de Componentes Principais (PCA).

A Ideia Geral

A análise de fatores tem objetivos que são similares àqueles da análise de componentes principais.

A ideia básica é que pode ser possível descrever um conjunto de p variáveis X_1, X_2, \dots, X_p em termos de um número menor de índices ou fatores, e no processo obter uma melhor compreensão do relacionamento destas variáveis.

Há, no entanto, uma diferença importante. A análise de componentes principais não é baseada em um modelo estatístico particular, enquanto que a análise de fatores é baseada em um modelo.

O desenvolvimento inicial de análise de fatores é o resultado do trabalho de Charles Spearman.

Enquanto estudava correlações entre escores de testes de estudantes de vários tipos, ele notou que muitas correlações observadas poderiam estar contidas em um modelo simples (Spearman, 1904).

Por exemplo, em um caso ele obteve a matriz de correlações mostrada na Tabela 7.1 (numeração do livro), para alunos em uma escola preparatória para seus escores em testes em clássicos, francês, inglês, matemática, discriminação de tom e música.

Tabela 7.1 Correlações entre escores de testes para meninos em uma escola preparatória

	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

Fonte: De Spearman, C. (1904), Am. J. Psychol., 15, 201-293.

Ele notou que esta matriz tinha a interessante propriedade de que quaisquer duas linhas eram quase proporcionais se as diagonais fossem ignoradas. Então para as linhas clássicos e inglês na Tabela 7.1 (numeração do livro), há razões:

$$\frac{0.83}{0.67} \approx \frac{0.70}{0.64} \approx \frac{0.66}{0.54} \approx \frac{0.63}{0.51} \approx 1.2$$

Baseado nesta observação, Spearman sugeriu que os seis escores de testes fossem descritos pela equação

$$X_i = a_i F + e_i$$

em que X_i é o i -ésimo escore depois dele ter sido padronizado para ter uma média zero e um desvio-padrão um para todos os alunos.

Aqui a_i é uma constante; F é um valor "fator", o qual tem média zero e desvio-padrão um para todos os alunos; e e_j é a parte de X_i que é específica para o i -ésimo teste somente.

Spearman mostrou que uma razão constante entre as linhas de uma matriz de correlações segue como uma consequência destas suposições, e que, portanto, este é um modelo plausível para os dados.

Além das razões de correlações constantes, segue também que a variância de X_i é dada por

$$\begin{aligned}\text{Var}(X_i) &= \text{Var}(a_i F + e_i) = \text{Var}(a_i F) + \text{Var}(e_i) \\ &= a_i^2 \text{Var}(F) + \text{Var}(e_i) = a_i^2 + \text{Var}(e_i)\end{aligned}$$

porque a_i é uma constante, F e e_j são assumidas independentes, e a variância de F é assumida ser unitária.

Também, porque $\text{Var}(X_i) = 1$,

$$1 = a_i^2 + \text{Var}(e_i).$$

Portanto a constante a_i a qual é chamada de carga do fator, é tal que seu quadrado é a proporção da variância de X_j que está contida no fator.

Com base no seu trabalho, Spearman formulou sua teoria de dois fatores de testes mentais. De acordo com esta teoria, cada resultado do teste é composto de duas partes, uma que é comum a todos os testes (inteligência geral), e outra que é específica para o teste.

Isto dá o modelo de análise de fatores geral, o qual estabelece que

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{im}F_m + e_i$$

em que X_i é o i -ésimo score do teste com média zero e variância unitária; a_{ij} são as cargas dos fatores para o i -ésimo teste; F_1, \dots, F_m são m fatores comuns não correlacionados, cada um com média zero e variância unitária; e e_i é um fator específico somente para o i -ésimo teste que é não correlacionado com qualquer dos fatores comuns e tem média zero.

Com este modelo,

$$\begin{aligned}\text{Var}(X_i) = 1 &= a_{i1}^2 \text{Var}(F_1) + a_{i2}^2 \text{Var}(F_2) + \dots + a_{im}^2 \text{Var}(F_m) \\ &= a_{i1}^2 + a_{i2}^2 + \dots + a_{im}^2 + \text{Var}(e_i)\end{aligned}$$

em que $a_{i1}^2 + a_{i2}^2 + \dots + a_{im}^2$ é chamado a **comunalidade** de X (a parte de sua variância que é relacionada aos fatores comuns), e $\text{Var}(e_i)$ é chamada **especificidade** de X (a parte de sua variância que não é relacionada aos fatores comuns).

Pode também ser mostrado que a correlação entre X_i e X_j é

$$\text{Correl}(X_i, X_j) = a_{i1}a_{j1} + a_{i2}a_{j2} + \dots + a_{im}a_{jm}.$$

Portanto dois escores de teste podem somente ser altamente correlacionados se eles têm altas cargas nos mesmos fatores.

Além disso, como a comunalidade não pode exceder um, é preciso que $-1 < a_{ij} < 1$.

Procedimento para uma análise fatorial

Os dados para uma análise de fatores têm a mesma forma como para uma análise de componentes principais.

Procedimento para uma análise fatorial

Caso	X_1	X_2	...	X_p
1	X_{11}	X_{12}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2p}
.
.
.
n	X_{n1}	X_{n2}	...	X_{np}

Procedimento para uma análise fatorial

Há três estágios para uma análise de fatores. Para começar, cargas de fatores provisórias a_{ij} são determinadas.

Procedimento para uma análise fatorial

Uma abordagem começa com uma análise de componentes principais e negligencia os componentes principais após os primeiros m , os quais são então tomados como sendo os m fatores.

Os fatores encontrados desta maneira são não correlacionados entre si, e são também não correlacionados com os fatores específicos. Isto pode não ser um problema desde que as comunalidades sejam altas.

Qualquer que seja a maneira como as cargas de fatores provisórios são determinadas, é possível mostrar que eles não são únicos.

Procedimento para uma análise fatorial

Se F_1, F_2, \dots, F_m são os fatores provisórios, então as combinações lineares deles da forma

$$F_1^* = d_{11}F_1 + d_{12}F_2 + \dots + d_{1m}F_m$$

$$F_2^* = d_{21}F_1 + d_{22}F_2 + \dots + d_{2m}F_m$$

.

.

.

$$F_m^* = d_{m1}F_1 + d_{m2}F_2 + \dots + d_{mm}F_m$$

podem ser construídos de modo a serem não correlacionados e explicar os dados tão bem quanto os fatores provisórios.

Procedimento para uma análise fatorial

De fato, há uma infinidade de soluções alternativas para o modelo de análise de fatores. Isto leva ao segundo estágio na análise, o qual é chamado de rotação de fator.

Procedimento para uma análise fatorial

Neste estágio, os fatores provisórios são transformados a fim de encontrar novos fatores que sejam mais fáceis de interpretar.

Girar ou transformar neste contexto significa essencialmente escolher os valores d_{ij} nas equações já vistas.

Procedimento para uma análise fatorial

O último estágio de uma análise envolve calcular os escores dos fatores. Estes são os valores dos fatores rotacionados $F_1^*, F_2^*, \dots, F_m^*$ para cada um dos n indivíduos para os quais os dados estão disponíveis.

Procedimento para uma análise fatorial

Geralmente, o número de fatores (m) depende do analista, apesar de algumas vezes poder ser sugerido pela natureza dos dados.

Procedimento para uma análise fatorial

Quando uma análise de componentes principais é usada para encontrar uma solução provisória, uma regra rústica envolve escolher m como sendo o número de autovalores maiores do que a unidade na matriz de correlações dos escores do teste.

Procedimento para uma análise fatorial

A lógica aqui é a mesma que foi explicada no capítulo anterior sobre análise de componentes principais. Um fator associado com um autovalor menor que a unidade responde por menos variação nos dados do que os escores de teste originais.

Em geral, aumentando m aumenta as comunalidades das variáveis. Entretanto, comunalidades não são alteradas por rotação de fator.

Procedimento para uma análise fatorial

Rotação de fatores pode ser ortogonal ou oblíqua. Com rotação ortogonal, os novos fatores são não correlacionados, como os fatores provisórios. Com rotação oblíqua, os novos fatores são correlacionados.

Qualquer que seja o tipo de rotação usada, é desejável que as cargas de fator para os novos fatores sejam ou próximas de zero ou muito diferentes de zero.

Procedimento para uma análise fatorial

Um a_{ij} próximo de zero significa que X_i não é fortemente relacionado com o fator F_j . Um grande valor positivo ou negativo de a_{ij} significa que X_i , é determinado em grande parte por F_j .

Procedimento para uma análise fatorial

Se cada escore de teste é fortemente relacionado com alguns fatores, mas não relacionado com outros, então isso torna os fatores mais fáceis de serem identificados do que o seria em outro caso.

Procedimento para uma análise fatorial

Um método de rotação de fatores ortogonal que é muitas vezes usado é chamado de rotação varimax.

Procedimento para uma análise fatorial

Este é baseado na suposição de que a interpretabilidade do fator j pode ser medida pela variância dos quadrados de suas cargas de fator, i.e., a variância de $a_{1j}^2, a_{j2}^2, \dots, a_{mj}^2$.

Se esta variância é grande, então os valores a_{ij} tendem a ser ou próximos de zero ou próximos da unidade. A rotação varimax, portanto, maximiza a soma destas variâncias para todos os fatores.

Inúmeros outros métodos de rotação ortogonal têm sido propostos. Entretanto, rotação varimax parece ser uma boa abordagem padrão.

Procedimento para uma análise fatorial

Algumas vezes analistas de fatores são preparados para desistir da ideia dos fatores serem não correlacionados a fim de tomar as cargas de fator tão simples quanto possível.

Uma rotação oblíqua pode então dar uma melhor solução do que uma ortogonal. Novamente, há numerosos métodos disponíveis para fazer a rotação oblíqua.

Procedimento para uma análise fatorial

Um método para calcular os escores de fator para indivíduos, baseado nos componentes principais, é descrito na próxima seção. Existem outros métodos disponíveis, de modo que aquele escolhido para uso dependerá do pacote computacional que está sendo usado na análise.

Análise de fatores de componentes principais

Análise de fatores de componentes principais

Foi observado anteriormente que uma maneira de fazer uma análise de fatores é começar com uma análise de componentes principais e usar os primeiros componentes principais como fatores não rotacionados.

Isto tem a virtude da simplicidade, apesar que, devido aos fatores específicos e_1, e_2, \dots, e_p , serem correlacionados, o modelo de análise de fatores não é muito correto.

Algumas vezes analistas de fatores fazem primeiro uma análise de fatores de componentes principais e então, após isto, tentam uma outra abordagem.

O método para encontrar os fatores não rotacionados é como segue.
Com p variáveis, haverá o mesmo número de componentes principais.

Análise de fatores de componentes principais

Estes são combinações lineares das variáveis originais

$$Z_1 = b_{11}X_1 + b_{12}X_2 + \dots + b_{1p}X_p$$

$$Z_2 = b_{21}X_1 + b_{22}X_2 + \dots + b_{2p}X_p$$

.

.

.

$$Z_p = b_{p1}X_1 + b_{p2}X_2 + \dots + b_{pp}X_p$$

em que os valores b_{ij} são dados pelos auto vetores da matriz de correlações.

Esta transformação dos valores X para valores Z é ortogonal ($A^{-1} = A^t$), de modo que o relacionamento inverso é simplesmente

$$X_1 = b_{11}Z_1 + b_{12}Z_2 + \dots + b_{p1}Z_p$$

$$X_2 = b_{12}Z_1 + b_{22}Z_2 + \dots + b_{p2}Z_p$$

.

.

.

$$X_p = b_{1p}Z_1 + b_{2p}Z_2 + \dots + b_{pp}Z_p$$

Análise de fatores de componentes principais

Para uma análise de fatores, somente m das componentes principais são retidas, assim as últimas equações se tornam

$$X_1 = b_{11}Z_1 + b_{12}Z_2 + \dots + b_{m1}Z_m + e_1$$

$$X_2 = b_{12}Z_1 + b_{22}Z_2 + \dots + b_{m2}Z_m + e_2$$

.

.

.

$$X_p = b_{1p}Z_1 + b_{2p}Z_2 + \dots + b_{mp}Z_m + e_p$$

em que e_i é uma combinação linear dos componentes principais Z_{m+1} a Z_p .

Tudo que é preciso ser feito agora é escalonar os componentes principais Z_1, Z_2, \dots, Z_m para terem variâncias unitárias, como requerido pelos fatores.

Para fazer isto, Z_i precisa ser dividido pelo seu desvio-padrão, o qual é $\sqrt{\lambda_i}$, a raiz quadrada do correspondente autovalor na matriz de correlações.

As equações então se tornam

$$X_1 = \sqrt{\lambda_1}b_{11}F_1 + \sqrt{\lambda_2}b_{12}F_2 + \dots + \sqrt{\lambda_m}b_{m1}F_m + e_1$$

$$X_2 = \sqrt{\lambda_1}b_{12}F_1 + \sqrt{\lambda_2}b_{22}F_2 + \dots + \sqrt{\lambda_m}b_{m2}F_m + e_2$$

.

.

.

$$X_p = \sqrt{\lambda_1}b_{1p}F_1 + \sqrt{\lambda_2}b_{2p}F_2 + \dots + \sqrt{\lambda_m}b_{mp}F_m + e_p$$

em que $F_i = Z_i/\sqrt{\lambda_i}$.

O modelo de fatores não rotacionado é então

$$X_1 = a_{11}F_1 + a_{12}F_2 + \dots + a_{m1}F_m + e_1$$

$$X_2 = a_{12}F_1 + a_{22}F_2 + \dots + a_{m2}F_m + e_2$$

.

.

.

$$X_p = a_{1p}F_1 + a_{2p}F_2 + \dots + a_{mp}F_m + e_p$$

onde $a_{ij} = \sqrt{\lambda_j}b_{ji}$.

Após uma rotação varimax ou outro tipo de rotação, uma nova solução tem a forma

$$X_1 = g_{11}F_1^* + g_{12}F_2^* + \dots + g_{m1}F_m^* + e_1$$

$$X_2 = g_{21}F_1^* + g_{22}F_2^* + \dots + g_{m2}F_m^* + e_2$$

.

.

.

$$X_p = g_{p1}F_1^* + g_{p2}F_2^* + \dots + g_{mp}F_m^* + e_p$$

em que F_i^* representa o novo i -ésimo fator.

Os valores do i -ésimo fator não rotacionado são justamente os valores do i -ésimo componente principal após eles terem sido escalonados para terem uma variância um.

Exemplos

Tabela 7.2 Autovalores e autovetores para dados de emprego europeu da Tabela 1.5

Autovalores	Autovetores								
	X ₁ AGR	X ₂ MIN	X ₃ FAB	X ₄ FEA	X ₅ CON	X ₆ SER	X ₇ FIN	X ₈ SSP	X ₉ TC
3,111	0,512	0,375	-0,246	-0,315	-0,222	-0,382	-0,131	-0,428	-0,205
1,809	-0,024	0,000	0,432	0,109	-0,242	-0,408	-0,553	0,055	0,516
1,495	-0,278	0,516	-0,503	-0,292	0,071	0,064	-0,096	0,360	0,413
1,063	0,016	0,113	0,058	0,023	0,783	0,169	-0,489	-0,317	-0,042
0,705	0,025	-0,345	0,231	-0,854	-0,064	0,269	-0,133	0,046	0,023
0,311	-0,045	0,203	-0,028	0,208	-0,503	0,674	-0,399	-0,167	-0,136
0,293	0,166	-0,212	-0,238	0,065	0,014	-0,165	-0,463	0,619	-0,492
0,203	0,539	-0,447	-0,431	0,157	0,030	0,203	-0,026	-0,045	0,504
0,000	-0,582	-0,419	-0,447	-0,030	-0,129	-0,245	-0,191	-0,410	-0,061

Exemplos

$X_1 = + \underline{0,90}$	$F_1 - 0,03$	$F_2 - 0,34$	$F_3 + 0,02$	$F_4 + e_1 (0,93)$
$X_2 = + \underline{0,66}$	$F_1 + 0,00$	$F_2 + \underline{0,63}$	$F_3 + 0,12$	$F_4 + e_2 (0,85)$
$X_3 = - 0,43$	$F_1 + \underline{0,58}$	$F_2 - \underline{0,61}$	$F_3 + 0,06$	$F_4 + e_3 (0,91)$
$X_4 = - \underline{0,56}$	$F_1 + 0,15$	$F_2 - 0,36$	$F_3 + 0,02$	$F_4 + e_4 (0,46)$
$X_5 = - 0,39$	$F_1 - 0,33$	$F_2 + 0,09$	$F_3 + \underline{0,81}$	$F_4 + e_5 (0,92)$
$X_6 = - \underline{0,67}$	$F_1 - \underline{0,55}$	$F_2 + 0,08$	$F_3 + 0,17$	$F_4 + e_6 (0,79)$
$X_7 = - 0,23$	$F_1 - \underline{0,74}$	$F_2 - 0,12$	$F_3 - \underline{0,50}$	$F_4 + e_7 (0,87)$
$X_8 = - \underline{0,76}$	$F_1 + 0,07$	$F_2 + 0,44$	$F_3 - 0,33$	$F_4 + e_8 (0,88)$
$X_9 = - 0,36$	$F_1 + \underline{0,69}$	$F_2 + \underline{0,50}$	$F_3 - 0,04$	$F_4 + e_9 (0,87)$

Aqui, os valores entre parênteses são as comunalidades. Por exemplo, a comunalidade para a variável X_1 é $(0,90)^2 + (-0,03)^2 + (-0,34)^2 + (0,02)^2 = 0,93$.

As comunalidades são bastante altas para todas as variáveis exceto X_4 (FEA, fornecimento de energia e água). Grande parte da variância para as outras oito variáveis originais está, portanto, contida nos quatro fatores comuns.

Cargas de fator que são 0,50 ou mais (ignorando o sinal) estão sublinhadas nas equações acima. Estas cargas grandes ou moderadas indicam como as variáveis estão relacionadas com os fatores.

Uma indesejável propriedade desta escolha de fatores é que cinco das nove variáveis X são fortemente relacionadas a dois dos fatores. Isto sugere que uma rotação de fatores pode fornecer um modelo mais simples para os dados.

Exemplos

$X_1 = + \underline{0,85}$	$F_1 + 0,10$	$F_2 + 0,27$	$F_3 - 0,36$	$F_4 + e_1$
$X_2 = + 0,11$	$F_1 + 0,30$	$F_2 + \underline{0,86}$	$F_3 - 0,10$	$F_4 + e_2$
$X_3 = - 0,03$	$F_1 + 0,32$	$F_2 - \underline{0,89}$	$F_3 - 0,09$	$F_4 + e_3$
$X_4 = - 0,19$	$F_1 - 0,04$	$F_2 - \underline{0,64}$	$F_3 + 0,14$	$F_4 + e_4$
$X_5 = - 0,02$	$F_1 + 0,08$	$F_2 - 0,04$	$F_3 + \underline{0,95}$	$F_4 + e_5$
$X_6 = - 0,35$	$F_1 - 0,48$	$F_2 - 0,15$	$F_3 + \underline{0,65}$	$F_4 + e_6$
$X_7 = - 0,08$	$F_1 - \underline{0,93}$	$F_2 + 0,00$	$F_3 - 0,01$	$F_4 + e_7$
$X_8 = - \underline{0,91}$	$F_1 - 0,17$	$F_2 - 0,12$	$F_3 + 0,04$	$F_4 + e_8$
$X_9 = - \underline{0,73}$	$F_1 + \underline{0,57}$	$F_2 - 0,03$	$F_3 - 0,14$	$F_4 + e_9$

As comunalidades não mudaram e os fatores são ainda não correlacionados. No entanto, esta é uma solução um pouco melhor do que a anterior, pois somente X_9 é apreciavelmente dependente de mais do que um fator.

Neste estágio, é usual tentar colocar rótulas aos fatores. É honesto dizer que isto muitas vezes requer um grau de criatividade e imaginação!

Tabela 7.3 Escores de fatores rotacionados para 30 países europeus

País	Fator 1	Fator 2	Fator 3	Fator 4
Bélgica	-0,97	-0,56	-0,10	-0,48
Dinamarca	-0,89	-0,47	-0,03	-0,67
França	-0,56	-0,78	-0,15	-0,25
Alemanha	0,05	-0,57	-0,47	0,58
Grécia	0,48	0,19	-0,23	0,02
Irlanda	0,28	-0,60	-0,36	0,03
Itália	0,25	-0,13	0,17	1,00
Luxemburgo	-0,46	-0,36	0,02	0,92
Países Baixos	-1,36	-1,56	-0,03	-2,09
Portugal	0,66	-0,45	-0,37	0,64
Espanha	0,23	-0,11	-0,09	0,93
Reino Unido	-0,50	-1,14	-0,35	-0,04
Áustria	0,18	0,05	-0,71	0,56
Finlândia	-0,78	-0,20	-0,21	-0,52
Islândia	-0,18	-0,04	-0,06	0,46
Noruega	-1,36	-0,17	0,20	-0,42
Suécia	-1,20	-0,52	0,04	-0,74
Suíça	0,12	-0,67	0,01	0,65
Albânia	3,16	-1,82	1,76	-1,78
Bulgária	0,47	1,56	-0,57	-0,65
Repúblicas Tcheca/Eslováquia	-0,26	1,45	3,12	0,44
Hungria	-1,05	1,70	2,82	-0,15
Polónia	0,97	0,71	-0,37	-0,42
Romênia	1,11	1,73	-1,69	-0,81
USSR (antiga)	0,08	2,09	-0,11	0,14
Iugoslávia (antiga)	0,13	1,48	-1,70	0,17
Cingapura	0,46	-0,32	0,03	1,08
Gibraltar	-0,05	-1,05	0,08	3,26
Malta	-1,17	0,49	-0,79	-1,31
Turquia	2,15	0,07	0,15	-0,56

Comentários Finais

Em resumo, na aula de hoje nós:

- aprendemos o que é a Análise Fatorial;
- executamos no Excel um exemplo proveniente do livro-texto.

Na próxima aula nós lidaremos com a Análise de Variância (ANOVA).

Referências







