NOTAS E COMENTÁRIOS

UMA APLICAÇÃO CORRETA DO MÉTODO DE ANÁLISE FATORIAL[1]

JOSÉ DE JESUS SOUSA LEMOS[2]

O método de análise fatorial, como todos nós sabemos, constitui uma técnica de "rebaixamento", assim como o são a análise discriminante, a análise de correlação canônica, a análise de conglomerados dentre outras. O "rebaixamento" consiste em: dado um número relativamente grande de variáveis, agrupá-las em um número menor, os chamados fatores que serão ortogonais, ou independentes entre si.

Assim, o objetivo fundamental do método é a tentativa de explicação das causas das variações de um certo número de variáveis a partir de fatores diferentes.

Definamos um espaço vetorial R^n, chamado espaço das variáveis. Podemos definir um vetor coluna $X \varepsilon R^n$ tal que o seu transposto X' apresente os seguintes componentes:

$$X' = (X_1\ X_2 \ldots X_n) \tag{1}$$

Definamos um vetor coluna $F \varepsilon R^m$ chamado de vetor dos fatores, tal que o seu transposto seja dado por:

$$F' = (F_1\ F_2 \ldots F_m) \tag{2}$$

Escrevamos adicionalmente um vetor coluna X, também em R^n, tal que

$$\bar{X}' = (\bar{X}_1\ \bar{X}_2 \ldots \bar{X}_n) \tag{3}$$

As componentes do vetor X são as médias das variáveis envolvidas na análise fatorial.

Definindo um outro vetor $x \varepsilon R^n$, cujas componentes são

$$x = X - \bar{X} \tag{4}$$

então, poderemos escrever que:

$$x = AF \tag{5}$$

na qual "A" é conhecida como "matriz de saturação" e tem em "a_{ij}" o seu elemento genérico. "A' é, portanto, uma matriz de dimensão (n x m). A matriz "A" tem a propriedade apresentada abaixo

$$A' = A^{-1} \tag{6}$$

---

[1] Resposta à nota assinada por Rodolfo Hoffman, apresentada na Revista de Economia Rural - vol. 25, número 4, out/dez, 1987.

[2] Professor-Adjunto da UFC, Doutor em Economia.

sendo A' e A^{-1}, respectivamente a transposta e a inversa de A.

Calcula-se a matriz "V" que é a matriz de variância e covariância de x, da seguinte maneira:

$$V = V(x) = E(x\,x') = E(AF\,(AF)') = A\,E(FF')A' \quad (7)$$

sendo E(FF') a matriz de variância e covariância dos fatores que é diagonal, já que os fatores são ortogonais. Chamando de "Λ" esta matriz, poderemos reescrever a equação (7) da seguinte foma:

$$V = A\,\Lambda\,A' \quad (8)$$

Pré-multiplicando (8) por A' e pós-multiplicando por A obtemos a relação:

$$\Lambda \neq A'\,V\,A \quad (9)$$

Os elementos que compõem a diagonal da matriz Λ são os "valores próprios" ou "raízes latentes" da matriz "V".

A equação (8) torna evidente que os resultados obtidos com a análise fatorial por este procedimento, dependeriam das unidades de medida em que estão sendo avaliadas as variáveis, como passaremos a demonstrar a seguir.

Suponhamos que em vez de X, utilizássemos as variáveis representadas pelo vetor Y, tal que:

$$Y = K\,x \quad (10)$$

na qual K é diagnonal de ordem (n x n). Então, a variância de Y pode ser calculada da forma que segue:

$$V(Y) = E(YY') = E(K\,xx'\,K') = K\,E(xx')\,K'$$

ou seja:

$$V(Y) = K\,V\,K' \quad (11)$$

Como K é diagnonal, segue-se que

$$V(Y) = K\,V\,K \quad (12)$$

Desta maneira, a nova formulação para a equação (9) passsa a ser:

$$M = A'\,K\,V\,K\,A \quad (13)$$

Uma vez que "Λ" é geralmente diferente de "M", conclui-se a afirmação acima, e agora demonstrada, de que, se conduzida desta forma, a análise fatorial dependeria das unidades em que estão sendo medidas as variáveis. Por esta razão, substituem-se as variáveis originais pelas "variáveis normalizadas" (z_i) definidas na forma a seguir

$$z_i = \frac{X_i - \bar{X}_i}{\sigma_i}, \quad i = 1, 2, \ldots \tag{14}$$

sendo σ_i o desvio padrão de X_i. Demonstra-se facilmente que

$$V(z) = E(zz') = R(X) = R \tag{15}$$

na qual R é a matriz de correlação de X.

De maneira análoga, poderemos normalizar o vetor F de fatores e obter um vetor Φ de forma que:

$$V(\Phi) = E(\Phi\Phi') = I$$

e a equação (5) se converte em:

$$z = B\Phi \tag{16}$$

em que B é dado por:

$$B = (\text{diag. } V)^{-1/2} A \, \Lambda^{1/2} \tag{17}$$

e a variância de z converte-se em:

$$V(z) = R = BV(\Phi)B' = BB' \tag{18}$$

Em conseqüência dos resultados acima, podemos concluir que o problema se resume na diagonalização de uma matriz R (matriz de correlação), ou simplesmente, na determinação das "direções próprias" associadas à matriz de saturação B, cujo elemento genético é "b_{ij}", de tal forma que:

$$Cov(z_i, \Phi_j) = b_{ij} = r_{ij} \tag{19}$$

O resultado acima é obtido porque z e Φ são variáveis normalizadas. A equação (19) nos informa que a correlação existente entre a variável "z_i" e o fator "Φ_j" é igual à saturação existente entre esta mesma variável e o respectivo fator. Podemos também demonstrar que:

$$\Sigma b_{ij}^2 = \Sigma r_{ij}^2 \tag{20}$$

que é um resultado tão óbvio que dispensa maiores comentários.

Cada coeficiente b_{ij} multiplicado por 100 constitui a percentagem de dependência entre a variável "z_i" e o fator "Φ_j", ou seja, representa a parte da variância da i-ésima variável explicada pelo j-ésimo fator. Por outro lado, a variância total de "z_i" é igual à soma dos quadrados dos valores que estão contidos na i-ésima linha da matriz "B".

Com base nas condições que acabamos de demonstrar para a decomposição em componentes principais da análise fatorial, o número de fatores (m) é sempre inferior ao número de variáveis (n). Os "m" fatores comuns às "n" variáveis são chamados de "fatores comuns" e são simbolizados por:

$$C \varepsilon R^m \tag{21}$$

Por outro lado, existem os "fatores específicos" a cada variável, que explicam o restante da variância não explicada pelos chamados "fatores comuns", e são simbolizados por:

$$S \varepsilon R^n \tag{22}$$

Assim, podemos agora escrever o nosso modelo de análise fatorial de forma mais ampla, incorporando estes resultados de modo que:

$$z = (D\,C + U\,S) \tag{23}$$

na qual D é a matriz de saturação (n x m; $m < n$) e U é uma matriz diagonal (n x n), cujo elemento genético é "μ_i", ou seja, a "unicidade" da variável "z_i".

Reescrevendo o vetor z

$$z = BF \tag{24}$$

fica claro que:

$$B = (DU) \text{ e } F = \left(\frac{C}{S}\right)$$

Para a estimação dos fatores, partimos da redefinição apresentada na equação (24), pré-multiplicando ambos os lados pela matriz $(BB')^{-1}$, e o resultado obtido sendo pré-multiplicado pela matriz (B') e lembrando que (BB') = R, chegamos finalmente ao resultado que queríamos, ou seja:

$$F = B'R^{-1}\,z. \tag{25}$$

Condonier *et al.*, referindo-se ao resultado encontrado acima, no último parágrafo da página 155, dizem textualmente: "os cálculos para se chegar a esta equação são bastante prolixos uma vez que implicam a inversão da matriz R". Como se percebe, ao contrário do que imaginavam, a inversão da matriz de correlação constitui etapa importante na estimação dos fatores.

Quanto à retirada de variáveis do modelo final, sem saber o que estava fazendo, tenho a informar que, como o estudo era exploratório[3], utilizei critérios puramente estatísticos para selecionar as nove variáveis finais. As onze demais apresentaram coeficientes de correlação praticamente unitários em valores absolutos entre si e com as nove variáveis restantes. Como qualquer pesquisador iniciante sabe, quando isto ocorre, nada impede que apenas parte das variáveis seja selecionada. Este é um procedimento elementar para contornar problemas de multicolinearidade, quando não é possível ampliar o tamanho da amostra, como foi o caso. Além disso, se não era esse o problema, como então conseguiria estimar os fatores depois que retirei as variáveis "problemas"? Qualquer pessoa que haja trabalhado com análise fatorial saberá que, se existir este problema de perfeita colinearidade entre variáveis, não poderá ser possível a estimação dos fatores.

Por outro lado, gostaria de lembrar que as 20 variáveis postas inicialmente, foram selecionadas por mim, portanto, todas elas estavam dentro dos meus planos de um provável entendimento de algumas características da economia coreana. Assim sendo, ao selecionar as nove variáveis finais, eu sabia o que estava fazendo. Desta forma, fica caracterizada a forma tendenciosa e desprovida de caráter científico com que foram feitas as críticas ao meu trabalho. Quanto ao número de observações que o referido autor alega serem seis ao invés de cinco, conforme disse no texto, tenho a esclarecer que os dados de que dispunha estavam dispostos da seguinte forma: 1974/75, 1975/76, 1976/77, 1977/78 e 1978/79. Assim, com observações cobrindo o período de 1974 a 1979, eu dispus de apenas cinco observações.

REFERÊNCIAS

CORDONIER, P. *et alii.* **Economia de la empresa agrária.** Madrid, Mundi-Prensa, 1973. 506p.

HARMAN, H. H. **Modern factor analysis.** Chicago, University of Chicago Press, 1960. 474p.

LAWLEY, M. **Factor analysis as a statistical method.** London, Butterworts, 1963. 118p.

LEMOS, J. J. S. Análise de desenvolvimento rural da República da Coréia: um estudo exploratório[3]. **Revista de Economia Rural, Brasília, 25** (2): 251-62, 1987.

NIE, N. H. *et alii.* **Statistical package for the social sciences.** Chicago, Mc FGraw-Hill. 1975. p. 468-514.

THURSTONE, L. L. **Multiple factor analysis.** Chicago, University of Chicago Press, 1961. 535p.

[3] Exploratório, segundo o dicionário do Aurélio, é palavra derivada de explorar que significa: **procurar, descobrir, especular.**