Universidade de São Paulo
Faculdade de Saúde Pública

VOLUME 32
NÚMERO 3
JUNHO 1998
p. 267-72

Revista de Saúde Pública

JOURNAL OF PUBLIC HEALTH

Método de simulação e escolha de fatores na análise dos principais componentes*

Method of simulation and choice of factors in the analysis of principal components

Marcelo P.A. Fleck e Marie C. Bourdel

Departamento de Psiquiatria e Medicina Legal da Universidade Federal do Rio Grande do Sul. Porto Alegre, RS - Brasil (M.P.A.F.), Centre Hospitalier Sainte Anne. Service de Santé Mentale et de Thérapeutique. Paris, França (M.C.B.)

Fleck Marcelo P.A., Método de simulação e escolha de fatores na análise dos principais componentes
Rev. Saúde Pública, 32 (3): 267-72, 1998

Método de simulação e escolha de fatores na análise dos principais componentes*

Method of simulation and choice of factors in the analysis of principal components

Marcelo P.A. Fleck e Marie C. Bourdel

Departamento de Psiquiatria e Medicina Legal da Universidade Federal do Rio Grande do Sul. Porto Alegre, RS - Brasil (M.P.A.F.), Centre Hospitalier Sainte Anne. Service de Santé Mentale et de Thérapeutique. Paris, França (M.C.B.)

Resumo

Objetivo Existem vários critérios para a escolha do número de componentes a serem mantidos na análise de componentes principais. Esta escolha pode dar-se por critérios arbitrários (critério de Kaiser p.ex.) ou subjetivos (fatores interpretáveis). Apresenta-se o critério de simulação de Lébart e Dreyfus.

Método É gerada uma matriz de números aleatórios, sendo realizada a análise de componentes principais a partir dessa matriz. Os componentes extraídos de um conjunto de dados como este representam um limite inferior que deve ser ultrapassado para que um componente possa ser selecionado. Utiliza-se como exemplo a análise de componentes principais da escala de Hamilton para a depressão (17 itens) numa amostra de 130 pacientes.

Resultados e Conclusões O método de simulação é comparado com o método de Kaiser. É mostrado que o método de simulação mantém apenas os componentes clinicamente significativos ao contrário do método de Kaiser.

Simulação. Análise fatorial.

Abstract

Objective There are many methods to determine how many components should be retained in principal components analysis. This choice can be made on the basis of arbitrary (Kaiser) or subjective (Interpretable factors) criteria. This work presents the simulation criteria of Lébart e Dreyfus. The method create a matrix of randomized numbers and a principal component analysis is performed on the basis of this matrix. The components extracted from this data represent the cut off values. Those that exceed this cut off value should be retained. As an example, a principal component analysis is performed with the Hamilton depression rating scale (17 items) on a sample of 130 subjects.

* Financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico/CNPq (Processo nº 200578/92.8).
Correspondência para/Correspondence to: Marcelo P. A. Fleck - Rua Ramiro Barcelloa, 2350 - 90035-003 Porto Alegre, RS - Brasil.
E-mail:/mfleck@voyager.com.br
Recebido em 2.12.1996. Aprovado em 24.6.1997.

Results and Conclusion The Simulation method is compared with the Kaiser method and is shown that the Simulation method maintains the components clinically significant.

Simulation. Factor analysis, statistical.

INTRODUÇÃO

A Análise de Componentes Principais (ACP) é uma técnica estatística de análise multivariada que transforma linearmente um conjunto original de variáveis num conjunto substancialmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto original. Esta idéia foi desenvolvida por Hotteling[9], embora Pearson[15] já o tivesse lançado sob uma forma geométrica diferente[1].

O objetivo da ACP é similar ao da análise fatorial no sentido de que ambas as técnicas tentam explicar parte da variação de um conjunto de variáveis a partir de um número menor de variáveis subjacentes. Resumidamente, a principal diferença entre as duas técnicas é a de que a ACP parte da ausência de um modelo estatístico subjacente na divisão das variáveis observadas e focaliza a explicação da variação total das variáveis observadas baseando-se nas propriedades da variância máxima dos componentes principais. A análise fatorial, por outro lado, parte de um modelo estatístico prévio que divide a variância total[1].

A ACP pode ser representada geometricamente sob a forma de uma nuvem de pontos individuais no espaço das variáveis. Os fatores ou eixos principais saídos de uma ACP fornecem imagens aproximadas dessa nuvem de pontos e a ACP propõe-se a medir a qualidade dessa aproximação[2].

Existem alguns conceitos fundamentais em ACP. Numa ACP normalizada, a inércia total da nuvem de pontos é igual ao número de variáveis (k). O *autovalor* associado a cada fator representa a inércia da nuvem segundo a direção desse eixo. O primeiro *autovalor* está sempre compreendido entre l e k; se ele está próximo de um, as variáveis não são correlacionadas duas a duas e a ACP não permite a redução da dimensão de dados. Quanto maior o *autovalor*, maior é sua capacidade de resumir as variáveis e, portanto, mais provável é o fator de ser importante. Um *autovalor* inferior a um indica que o eixo sintetiza menos dados que uma variável isolada. Considerar o número de *autovalores* quase nulos permite calcular a dimensão real dos dados.

A coordenada de uma variável ao longo de um eixo é o coeficiente de correlação entre essa variável e o fator (*saturação*). O quadrado de sua coordenada é igual à qualidade da representação de uma variável e é proporcional à sua contribuição à inércia do eixo. A interpretação do fator funda-se sobre a síntese das variáveis mais ligadas (com maior coeficiente de correlação) ao eixo[10].

Uma das principais questões que se coloca na ACP é o critério de escolha do número de componentes a manter. Existem alguns métodos clássicos para orientar essa escolha. Os mais conhecidos são os seguintes[12]:

Critério de Kaiser[11]

É provavelmente o critério mais usado. Kaiser propõe considerar apenas os *autovalores* superiores a um, demonstrando que esses seriam os valores estatisticamente significativos. No entanto, esta condição não é suficiente. Nem todos os *autovalores* superiores a um correspondem a componentes com significado evidente.

Diagrama de Autovalores

A observação do diagrama de *autovalores* permite conservar aqueles situados acima do ponto de ruptura da queda da curva da função que relaciona a ordem e os *autovalores*. Assim, se dois fatores são associados a *autovalores* quase iguais, eles representam a mesma proporção de variabilidade e não há motivo, a priori, de conservar um e não outro. Inversamente, uma forte diminuição entre dois *autovalores* sucessivos, leva a conservar na interpretação os fatores que a precederam.

Fatores Interpretáveis

Um critério empírico mas não desprovido de sentido é recomendado por diferentes autores, em particular Thurstone[16], que propõem conservar os fatores os quais sabe-se claramente dar um significado. A partir da constatação de que esses métodos em muitos exemplos são excessivamente arbitrários (como o método de Kaiser) ou excessivamente subjetivos (como o método dos fatores interpretáveis), será descrito no presente artigo o método de simulação de Lébart[13] que, embora pouco conhecido, propicia uma abordagem intermediária e por vezes muito útil como guia na escolha do número de componentes a conservar para a interpretação e/ou rotação.

MÉTODO DE SIMULAÇÃO DE LÉBART

O método de simulação de Lébart[13] visa a responder a seguinte questão: " O que acontecerá se se conduzir uma análise de componentes principais sobre um conjunto de dados que na sua construção não continham nenhuma estrutura subjacente". É intuitivamente claro que os fatores extraídos de um conjunto de dados como este representa um limite inferior que deve ser ultrapassado para que um componente possa ser levado em conta.

Para a sua operacionalização, cria-se uma matriz do mesmo tamanho de números aleatórios mas que respeitem a mesma distribuição (desvio-padrão) e a mesma média de cada variável da matriz em estudo. Após, faz-se a análise de componentes principais e obtém-se uma série de *autovalores*. Este processo é repetido "n" vezes. Para cada classificação na série de *autovalores*, conserva-se o *autovalor* máximo observado ao longo das "n" simulações. Esses valores máximos observados representam o limite inferior que deve ser ultrapassado para que um componente possa ser levado em conta. A probabilidade de se observar um *autovalor* maior que o valor máximo obtido durante as n-1 simulações precedentes é de 1/n. Assim, para uma matriz de um tamanho determinado, há um risco de 1/n de se observar um *autovalor* de uma dada matriz que seja superior a este limite de confiança. Vinte simulações são suficientes para atingir um limiar de significância de 5% (1/20).

Aplicação a um Exemplo

Cento e trinta pacientes internados por uma síndrome depressiva em serviços universitários de dois países (França e Brasil) foram incluídos no presente estudo.

Os critérios de inclusão foram os seguintes: a) início de um episódio depressivo atual definido como os primeiros três dias após a internação; b) presença de ao menos dois dos sintomas seguintes: humor depressivo, idéias de suicídio, desesperança, sentimento de inutilidade, hipocondria e/ou ansiedade, sentimento de capacidade diminuída, auto-acusações ou culpabilidade, incapacidade de sentir ou de desfrutar e lentificação ídeo-motora; c) ser considerado um indivíduo da nacionalidade do país em estudo (francês na França/ brasileiro no Brasil) definida pelo seguinte: ter a língua oficial do país em estudo como primeira língua; ter sido educado no país em estudo durante pelo menos os dez primeiros anos de vida; ter vivido a maior parte da vida no país em estudo.

Os critérios de exclusão (ou de não-inclusão) foram os seguintes: a) doença física real em evolução; b) transtornos mentais orgânicos segundo o "Diagnostic and Statistical Manual of Mental Disorders (DSM III-R)"; c) transtornos ligados ao uso de substâncias ilegais ou álcool (dependência) segundo o DSM III-R; d) esquizofrenia segundo o DSM III-R; e) presença de dificuldades de linguagem ou de audição e f) retardo mental (Q.I. = 70 ou menos).

Todos os pacientes foram avaliados do ponto de vista diagnóstico com a entrevista CIDI (Composite International Diagnostic Interview) nas versões francesa[14] e brasileira. Este instrumento permite a obtenção dos dados necessários para fazer o diagnóstico das principais categorias diagnósticas da Classificação Internacional de Doenças (CID-10) e do DSM III-R de forma completamente padronizada. Além da avaliação diagnóstica, foram aplicadas quatro escalas de depressão: Escala de Depressão de Hamilton com 17 itens (HDRS-17); Escala de Montgomery-Asberg (MADRS); Escala de Lentificação Depressiva (ELD) e Escala Complementar (EC) composta por itens não avaliados pelas outras escalas.

As escalas foram aplicadas a partir de um guia para uma entrevista semi-estruturada adaptado às quatro escalas de depressão utilizadas no estudo. A versão original foi feita em francês[3], sendo realizada uma versão em português, seguida de retro-tradução comparada com o original em francês[3].

As características da amostra, bem como os resultados do presente estudo foram objeto de publicações específicas[4,5].

Para o presente trabalho se utilizará como ilustração o processo de escolha do número de fatores para análise de componentes principais da escala de Hamilton para a depressão (17 itens)[7, 8] numa amostra de 130 pacientes.

Os dados foram analisados através do programa BMDP-90 versão 7.0. Foi usada a ACP com rotação Varimax. Para o método de simulação foram realizadas 20 simulações utilizando a geração de número aleatórios presente no programa BMDP-90.

RESULTADOS

Observa-se que o número de fatores mantidos utilizando os diferentes critérios variou (Tabela 1).

Ao se utilizar o *Critério de Kaiser*, observa-se que os seis primeiros *autovalores* são superiores a um, o que levaria a que estes 6 fossem mantidos (Tabela 2). No entanto, a interpretação clínica desses fatores é possível para apenas alguns deles. O fator 1 ("depressão") agrupa os itens "nucleares" da sintomatologia depressiva: "*humor depressivo*", "*retardo*", "*trabalho e atividades*" e "*culpabilidade*". O fator 2 ("insônia") reúne os três itens referentes a insônia ("*insônia terminal*", "*insônia intermediária*" *e "insônia inicial").* O fator 3 ("ansiedade") agrupa os itens que descrevem sintomas de ansiedade tanto psíquica como somática ("*hipocondria*", "*ansieda-*

Tabela 1- Número de fatores retidos segundo os diferentes critérios na Análise de Componentes Principais da escala de Hamilton-17.

Método	Nº de fatores
Diagrama de autovalores	1
Método de simulação	3
Método dos fatores interpretáveis	4
Método de Kaiser	6

Tabela 2- Relação dos autovalores na amostra e no método de simulação.

Autovalores	Amostra	Simulação
1	3,76	1,93
2	1,90	1,62
3	1,52	1,51
4	1,29	1,37
5	1,15	1,31
6	1,07	1,22
7	0,97	...
8	0,88	...
9	0,84	...
10	0,71	...
11	0,59	...
12	0,55	...
13	0,48	...
14	0,44	...
15	0,32	...
16	0,28	...
17	0,26	...

de somática" e "ansiedade psíquica") enquanto o fator 4 agrupa os itens exclusivamente somáticos (*"sintomas somáticos gastrintestinais" e "perda de peso").* Os fatores 5 e 6 agrupam itens sem coerência clínica evidente (Tabela 3).

Quando se utiliza o critério do *Diagrama de autovalores* observa-se que existe um ponto de inflexão na curva entre o fator 1 e 2 (Figura). Desta forma, segundo este critério, apenas o fator 1 deve ser retido e os demais fatores suprimidos.

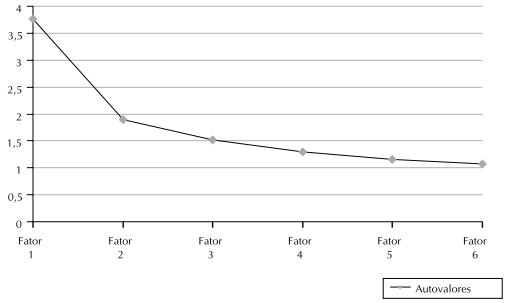

Figura - Representação dos escores dos autovalores de cada fator (método do Diagrama de autovalores).

Utilizando o método dos *Fatores interpretáveis* observou-se que esses são em número de 4, conforme visto anteriormente: um primeiro fator representando o "núcleo depressivo"; o segundo fator representando o fator "insônia"; o terceiro sendo um fator de "ansiedade" e o quarto agrupando os itens "perda de peso" e "sintomas gastrintestinais". Os fatores seguintes parecem itens sem ligação clínica evidente.

Através do método de *Simulação* três fatores foram mantidos. Os valores máximos obtidos para os 6 primeiros *autovalores,* alcançados depois de 20 simulações, foram 1,93; 1,62; 1,51; 1,37; 1,31; 1,22 (Tabela 2). Como apenas os três primeiros autova-

Tabela 3- Coeficientes de saturação dos itens da escala de Hamilton-17 em relação a 6 fatores (Critério de Kaiser: autovalor >1).

	1	2	3	4	5	6
	Depressão	Insônia	Ansiedade	Sint. somáticos		
Sintomatologia	3,76 (a)	1,9	1,52	1,29	1,15	1,07
	(21%) (b)	(14%)	(11%)	(9%)	(7%)	(6%)
1. Humor depressivo	0,84*					
8. Retardo	0,81*					
7. Trabalho e atividades	0,73*					
2. Culpa	0,56*			0,27		0,29
6. Insônia (terminal)		0,82*				
5. Insônia (intermed)		0,77*		0,26		
4. Insônia (inicial)		0,76*	0,27			
15. Hipocondria			0,82*			
11. Ansiedade somática			0,72*			
10. Ansiedade psíquica	0,32		0,51*		- 0,37	
12. Sint. somáticos gastrintestinais				0,84 *		0,43
16. Perda de peso		0,33		0,70 *		
3. Suicídio					0,61*	
9. Agitação			0,27	0,37	0,57*	
17. "Insight"						0,74 *
14. S. Genitais	0,28				0,47	0,57 *
13. Sint. somáticos gerais	0,39		0,31	0,27	0,36	

(a) Autovalor
(b) Percentagem da variância explicada pelo fator.
* Coeficiente de saturação > 0,5
Os coeficientes de saturação < 0,25 foram suprimidos.

lores de nossa análise ultrapassam esses valores (que são os limiares para 5,0%), conservam-se portanto três fatores. Após a rotação Varimax sobre três fatores, os três eixos principais são interpretados como um núcleo depressivo, um componente de insônia e outro de ansiedade (Tabela 4).

Tabela 4- Coeficientes de saturação dos itens da escala de Hamilton-17 em relação a 3 fatores (critério de Lébart).

Sintomatologia	1 Depressão 3,76 (22%)	2 Ansiedade 1,9 (11%)	3 Insônia 1,52 (9%)
1. Humor depressivo	0,83 *		
8. Retardo	0,79 *		
7. Trabalho e atividades	0,74 *		
2. Culpa	0,54 *		
15. Hipocondria		0,68 *	
9. Agitação		0,63 *	
11. Ansiedade somática		0,57 *	
12. Sintomas somáticos gastro-intestinais		0,55 *	
10. Ansiedade psíquica	0,26	0,55 *	
5. Insônia (intermediária)			0,80 *
6. Insônia (terminal)			0,78 *
4. Insônia (precoce)			0,76 *
16. Perda de peso		0,48	0,34
3. Suicídio			0,26
17. "Insight"	-0,25	0,31	
13. Sintomas somáticos gerais	0,47		
14. Sintomas genitais	0,31		

* Coeficiente de saturação > 0,5
Os coeficientes de saturação < 0,25 foram suprimidos.

DISCUSSÃO

Os quatro métodos conduzem a escolhas muito diferentes indo de um (método de Diagrama de *autovalores*) a seis fatores (método de Kaiser). A simulação que permite conservar os três fatores que mais claramente são interpretados (o quarto é composto por apenas dois itens) é satisfatório. Esse método permite reduzir o número de fatores eliminando todos aqueles que poderiam ser observados pelo acaso com uma probabilidade dada; essa probabilidade pode ser reduzida tanto quanto desejada, aumentando o número de simulações. No entanto, as simulações demandam tempo de cálculo e devem ser refeitas para cada estudo, pois a tabela de simulação depende do número de variáveis, de sujeitos e, sobretudo, da distribuição das variáveis.

Embora seja um método pouco conhecido e utilizado, ele foi descrito há aproximadamente 25 anos (Lébart e Dreyfus[13], 1972). Apresenta como características principais o fato de não ser arbitrário como o método de Kaiser, sendo adaptado para cada amostra de dados já que as simulações devem ser individualizadas, tomando como base a média e desvio-padrão da amostra em estudo. Sua principal desvantagem é que sua aplicação é mais demorada, necessitando realizar 20 ACP a partir de dados gerados ao acaso para escolher os autovalores limites.

REFERÊNCIAS

1. DUNTEMAN, G. *Principal components analysis.* London, Sage, 1989.

2. ESCOFIER, B. & PAGES, J. *Analyses factorielles simples et multiples*. Paris, Dunod, 1990.

3. FLECK, M.P.A. et al. Application d'un guide pour l'entretien structuré adapté à quatre échelles de dépression. *Encephale,* **20**: 479-86, 1994.

4. FLECK, M.P.A. et al. Factorial structure of the 17-item Hamilton depression rating scale. *Acta Psychiatr. Scand.*, **92**:168-72, 1995.

5. FLECK, M.P.A. et al. Avaliação diagnóstica de paciente com síndrome depressiva: um estudo comparativo entre entrevista clínica e estruturada. *Rev. HCPA & Fac. Med. Univ. Fed. Rio Gd. do Sul,* **15**:32-4, 1995.

6. FLECK, M.P.A. et al. Aplicação da versão em português de um guia para entrevista semi-estrutrada adaptada a quatro escalas de depressão. *J. Bras. Psiquiatr.*, **46**: 339-45,1997.

7. HAMILTON, M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry,* **26:** 56-62, 1960.

8. HAMILTON, M. Development of a rating scale for primary depressive illness. *Br. J. Soc. Clin. Psychol.*, **6**:278-96, 1967.

9. HOTTELING, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, **24**: 417-41, 498-520, 1933 .

10. JOLLIFE, I.T. *Principal components analysis*. New York, Springer-Verlag, 1986.

11. KAISER, H.F. The varimax criteria for analytical rotation in factor analysis. *Psychometrica*, **23**: 141-51, 1958.

12. KIM, J-O. & MUELLER, C. *Factor analysis: statistical methods and practical issues.* London, Sage, 1978.

13. LÉBART, L. & DREYFUS, J. F. Comment limiter de façon non arbitraire le nombre de facteurs dans une analyse en composantes principales. *Rev. Rech. Fond. Barth*, **2**:7-9, 1972.

14. ORGANIZAÇÃO MUNDIAL DA SAÚDE. *Composite International Diagnostic Interview* (CIDI). Geneva, 1991.

15. PEARSON, K. On lines and planes of closest fitto systems of points in space. *Philos. Mag.*, **2**: 559-72, 1901.

16. THURSTONE, L.L. *Multiple factor analysis.* Chicago, University of Chicago Press, 1947.