

# SOFTWARE, INSTRUMENTACION Y METODOLOGIA

## UNA SIMULACION SOBRE EL COMPORTAMIENTO DE LA REGLA K1 EN LA ESTIMACION DEL NUMERO DE FACTORES

Miguel A. RUIZ y Rafael SAN MARTIN

Facultad de Psicología. Universidad Autónoma de Madrid

A pesar de su controversia teórica, la regla de Kaiser-Guttman sigue siendo utilizada de manera extensiva en la determinación del número de factores en el Análisis Factorial de Factor Común. Mediante una simulación se manipulan el número de factores, el número relativo de variables y el tamaño relativo de las muestras, para determinar el comportamiento de dicha regla y el efecto de la utilización de la misma sobre la estimación de la comunalidad. Aunque la regla sobreestima la comunalidad sistemáticamente, su utilización sólo lleva a un sesgo en la estimación del número de factores cuando la proporción de sujetos por variable es baja.

**Palabras clave:** Análisis Factorial; Número de factores; Estimación de la comunalidad.

*The behavior of the K1 rule estimating the number of factors: A study with simulated data.* Although the theoretical drawbacks of the rule, the Kaiser-Guttman criteria for determining the number of factors is still widely used, when working in Common Factor Analysis. In order to assess the behavior of the rule, the number of factors, the proportion of variables to factors and the proportion of sample size to the number of variables in the sample are controlled in a simulation study. The bias in the estimation of the number of factors and the communalities are gauged. This criteria shows a regular overestimation of the communality. Still, it only fails to determine the right number of factors when the relative sample size is small.

**Key words:** Factor Analysis; Number of factors; Communality estimation.

El Análisis Factorial, siendo una de las técnicas multivariadas más antiguas, sigue siendo una de las herramientas más utilizadas hoy por hoy en la investigación psicológica. A pesar de su antigüedad, algunos de los problemas que plantea todavía no parece que estén resueltos. En concreto, el número de factores idóneo que deben ser rete-

nidos en un análisis sigue sin tener una respuesta adecuada.

Se han desarrollado diversas reglas y estadísticos para la determinación del número de factores en el modelo de Factor Común (Thurstone, 1947), algunas de carácter gráfico como la prueba de sedimentación (Scree-test, Cattell, 1966; Cattell y Vogelmann, 1977), otras de carácter lógico

teórico como la regla de Kaiser-Guttman (Guttman, 1953; Kaiser, 1960) y el índice de interpretabilidad de Crawford (1975; Revelle y Rocklin, 1979); y otras de carácter estadístico como la prueba de significación de Bartlett (1950, 1951), la prueba de bondad de ajuste de Jöreskog (1966) y el MAP de Velicer (Minimum Average Partial Method, 1976). Sin embargo, ninguna de ellas ha demostrado comportarse de manera inequívoca en cualquier tipo de situación.

La mayoría de los paquetes estadísticos estándar utilizados en el análisis de datos ofrecen una o varias de estas reglas. Así, tanto el SPSS como el BMDP utilizan la regla de Kaiser-Guttman como criterio por defecto para la determinación del número de factores a retener, a pesar de su mal comportamiento bajo determinadas circunstancias. Programas más recientes como el SYSTAT no utilizan directamente esta regla, y extraen tantos factores como variables menos 1, a no ser que se especifique lo contrario.

En el presente trabajo se generan muestras correspondientes a poblaciones de estructura factorial conocida, manipulando el número de factores ( $r$ ), la proporción de variables por factor ( $r:n$ ) y la proporción de sujetos por variable ( $n:N$ ), para determinar las condiciones en las que la utilización de la regla puede ser especialmente perniciosa.

## PLANTEAMIENTO

La regla de Kaiser-Guttman, o regla K1, determina el número de factores como aquellos correspondientes a autovalores mayores que la unidad de la autodescomposición de la matriz de correlaciones sin reducir. Inicialmente fue propuesta por Guttman (1953) como un límite para la estimación de la comunalidad, y hallar el rango mínimo de la matriz de correlaciones reducida. Posteriormente, Kaiser (1960) propuso esta regla como un medio para la determinación

del número de factores, basándose en criterios psicométricos.

Otros procedimientos han sido desarrollados para la estimación de los límites inferior y superior de la comunalidad (Yanai e Ichikawa, 1990), y para la estimación del rango mínimo de una matriz de covarianzas (Ten Berge y Kiers, 1991), pero dudamos que estos procedimientos sean fáciles de implementar y por ello, resultan poco accesibles para la mayoría de los investigadores.

A parte de las críticas teóricas que ha recibido esta regla (Cliff, 1988; Velicer y Jackson, 1990; Schönemann, 1990), los resultados de los trabajos de simulación que hemos podido encontrar indican que la regla K1 sobreestima generalmente el número de factores, y su comportamiento empeora cuando el tamaño de la muestra aumenta, la cuantía de las comunalidades disminuye, y la proporción de sujetos por variable es baja ( $n:N=1:2$ ; Zwick y Velicer, 1986), Linn (1968) encontró sobreestimaciones cuando el tamaño de la muestra era pequeño y la fiabilidad de las variables era baja. Cattell y Vogelmann (1977); Hakslian, Rogers y Cattell (1982) encontraron un empeoramiento a medida que se aumentaba el número de variables, al igual que los resultados de Hakslian y Muller (1979) en el análisis de datos reales.

Sin embargo, Mote (1970) y Humphreys (1964), utilizando criterios de interpretabilidad para su discusión, encontraron una infraestimación del número de factores en muestras reales de puntuaciones psicométricas.

A pesar del aparente mal comportamiento de esta regla, su utilización sigue siendo extensiva, dada su simplicidad de cálculo y su disponibilidad en los paquetes estadísticos más extendidos. Nos proponemos verificar el comportamiento de dicha regla y hacer una llamada de atención para una utilización más cauta de criterios como

éste. La retención de un número excesivo de factores infla artificialmente las comunalidades, incita a interpretar factores meramente aleatorios, afecta artificiosamente las rotaciones, y resta importancia a los factores correspondientes a los mayores autovaleores, que en definitiva, son los que mayor variabilidad deberán explicar.

### PROCEDIMIENTO

Para generar las muestras a analizar se utilizó la fórmula general del modelo de Factor común del Análisis Factorial, donde las puntuaciones de N sujetos en n variables, pueden ser reproducidas según la fórmula

$$Z = CF' + UD \quad (1)$$

donde **Z** es la matriz Nx<sub>n</sub> de puntuaciones observadas, **F** es la matriz de estructura de orden rx<sub>n</sub>, determinada por nosotros con las características deseadas, y que representa la estructura común de las variables, **C** es la matriz de orden Nx<sub>r</sub> de puntuaciones de los sujetos en los factores comunes, **D** es la matriz diagonal de dimensiones nx<sub>n</sub>, que contienen las unicidades y representa la parte específica de las variables, y la matriz **U** de puntuaciones específicas y de dimensiones Nx<sub>n</sub>.

Las estructuras se determinaron a priori como se especifica más abajo, con lo que las matrices **F** y **D**, son definidas por nosotros, y reproducen las características de interés.

Las matrices **C** y **U** fueron generadas por el algoritmo

$$Z_i = [-2 \log (X_i)]^{1/2} \cos (2 \pi X_i) \quad (2)$$

propuesto por Lewis y Orav (1989) y el generador de números pseudoaleatorios del BASIC versión 2.0 del IBM 3090, produ-

ciendo n variables independientes distribuidas normalmente y en escala típica. Los tamaños de la muestras vienen determinados a priori por la combinación de las condiciones manipuladas.

### DISEÑO

Las variables manipuladas fueron el tamaño de la muestra (N), el número de variables (n) y el número de factores (r) de la estructura factorial. La estructura factorial era una estructura de Factor Común simple, con saturaciones de las variables elevadas en el factor correspondiente (.77 a .72, según el número de factores) y saturaciones despreciables (.01 a .089) en los otros factores.

La comunalidad fue del 60% de varianza común para todas las variables y condiciones. Esta saturación representa más del 50% de la varianza de cada variable, lo que coincide con la conceptualización del modelo de factor común, no es tan elevada como para llevar a pensar que se trata de estructuras demasiado ideales y por tanto separadas de las encontradas en la realidad en el análisis de datos psicológicos, y coincide con la utilizada en otros estudios similares como el de Hakstian et al. (1982).

La comunalidad se impuso mediante la normalización de Kaiser de los vectores fila de las matrices de estructura.

Las condiciones manipuladas fueron:

- a) El número de factores de la estructura factorial. Sus valores fueron fijados en 2, 6 y 12. Que representan tres situaciones características de los modelos de factor común. Estos tres niveles de la variable dependiente pueden considerarse como un número bajo de factores, medio y elevado. Aunque existen estudios que han encontrado un número de factores mayores, en Psicología es infrecuente considerar

un número de dimensiones mayor que éste, ya que dicha técnica se utiliza fundamentalmente para la síntesis de información.

- b) Las proporciones de variables por factor fueron fijadas en 1/3, 1/6 y 1/9 de variables por factor es el mínimo recomendado para este modelo, de lo contrario los factores no se considerarían suficientemente definidos. 6 variables por factor sobreidentifica los factores, y es el valor recomendado por la mayoría de los analistas. 9 variables por factor es una condición que supera ampliamente la tendencia de la regla K1 comunicada por Gorsuch (1974) a sobreestimar el número de factores de 1/3 a 1/5 de las variables utilizadas.
- c) Las proporciones de sujetos por variable fueron fijadas en 2, 6 y 12 sujetos por variable. Un tamaño relativo de 2 sujetos por variable es un valor que, siendo bajo, puede reproducir situaciones relativamente frecuentes en psicología, además es una de las ratios utilizadas en estudios similares como el de Zwick y Velicer (1986). 6 sujetos por variable es el mínimo recomendado por los analistas, y 12 sujetos por variable es una proporción que puede encontrarse en estudios de validación psicométricos, siendo además una situación que puede ser considerada perfecta para este tipo de análisis.

Las variables dependientes utilizadas para medir el comportamiento de la regla son:

- a. La diferencia entre el número de factores extraído en cada momento y el número de factores teórico de la estructura subyacente, todo ello dividido por el número de factores teó-

rico, para que el índice sea comparable a través de los distintas condiciones. Este índice representa el sesgo relativo en la estimación del número de factores.

- b. El sesgo en la estimación de la comunalidad, definido como la diferencia entre la comunalidad reproducida por el número de factores determinado por la regla y la comunalidad teórica. Con ella pretendemos valor hasta que punto se infla la comunalidad.

La extracción utilizada fue Componentes Principales, procedimiento para el que fue inicialmente propuesta la regla K1.

Para cada combinación de variables manipuladas de generaron 5 muestras aleatorias independientes. Los resultados fueron analizados mediante un Análisis de Varianza  $3 \times 3 \times 3$ .

## RESULTADOS

En la tabla 1 se recogen los resultados en la estimación del número de factores en las diversas condiciones. Sólo en la condición de menor número de sujetos (2 sujetos por variable) se apreciaron sesgos en la estimación. A la vista de las medias de cada condición y de sus desviaciones típicas, podemos observar un empeoramiento de la predicción a medida que la proporción de variables por factor aumenta, así como el número de factores de la estructura factorial teórica. Las únicas infraestimaciones fueron debidas a la combinación de 12 factores teóricos, 3 variables por factor y 2 sujetos por variable.

El Análisis de Varianza de la proporción de factores infra o sobreestimados fue significativo en la interacción de la proporción de variables por factor y la proporción de sujetos por variable ( $p = 0.000$ ) y la interacción del número de factores por la proporción de sujetos por variable ( $p < 0.05$ ), así

r		2			6			12		
n/N	r/n	1/3	1/6	1/9	1/3	1/6	1/9	1/3	1/6	1/9
1/2	M	2	2.6	2.6	6	6.2	7	11	12.2	13.8
	DT	.0	.55	.89	.0	.45	1.3	.71	.45	.84
1/6	M	2	2	2	6	6	6	12	12	12
	DT	.0	.0	.0	.0	.0	.0	.0	.0	.0
1/12	12	2	2	2	6	6	6	12	12	12
	DT	.0	.0	.0	.0	.0	.0	.0	.0	.0

Tabla 1: Número medio (M) de factores estimados y desviaciones típicas (DT) por combinación de número de factores (r), proporción de variables por factor (r/n) y proporción de sujetos por variable (n/N).

como los efectos simples de cada condición manipulada.

Respecto a la estimación de la comunalidad los resultados fueron analizados mediante un Análisis de Varianza 3 x 3 x 3, los cuales aparecen resumidos en la tabla 2. Como puede observarse en la tabla 3, todas la estimaciones sobreestimaron la comunalidad, lo que se refleja en la significación del término constante del ANOVA (p =

0.000). Ninguna de las interacciones de tratamientos fue estadísticamente significativa. Sólo los efectos de la proporción de variables por factor (p = 0.000) y la proporción de sujetos por variable (p = 0.000) alcanzaron la significación.

A la vista de la tabla 3 podemos concluir que la estimación empeora a medida que la proporción de variables por factor disminuye. Cuando la proporción de sujetos

Fuente de variación	SC	GL	MC	F	Sig de F
Intragrupo	.05	108	.00		
Constante	1.38	1	1.38	2954.21	.000
R	.00	2	.00	1.01	.367
n	.21	2	.10	221.84	.000
N	.03	2	.01	29.89	.000
R x n	.00	4	.00	.88	.481
R x N	.00	4	.00	.84	.502
n x N	.00	4	.00	.89	.474
R x n x N	.01	8	.00	1.35	.227

Tabla 2: Tabla resumen del ANOVA para la estimación de la comunalidad.

r		2			6			12		
n/N	r/n	1/3	1/6	1/9	1/3	1/6	1/9	1/3	1/6	1/9
1/2	M	.155	.108	.105	.195	.105	.081	.163	.098	.080
	DT	.041	.052	.042	.024	.018	.011	.012	.012	.006
1/6	M	.152	.091	.050	.151	.070	.048	.153	.076	.052
	Dt	.033	.027	.020	.022	.007	.009	.006	.007	.002
1/12	M	.153	.069	.058	.136	.069	.047	.138	.073	.05
	DT	.035	.018	.017	.007	.006	.012	.005	.004	.004

Tabla 3: Sesgo medio (M) de las comunalidades estimadas y desviaciones típicas (DT) por combinación de número de factores (r), proporción de variables por factor (r/n) y proporción de sujetos por variable (n/N).

por variable aumenta, la estimación de la comunalidad mejora sensiblemente.

## DISCUSION

En la línea de los resultados de Zwick y Velicer (1986), la regla K1 se comporta de manera equívoca cuando la proporción de sujetos por variable es baja y el número de variables aumenta. También en la línea de las ideas planteadas por Gorusch (1974) parece que esta regla tiende a retener un número de factores en torno a 1/3 y 1/5 de las variables introducidas en el análisis, si además la proporción de sujetos es baja. Sin embargo, en este estudio sólo se utilizaron estructuras simples con saturaciones elevadas y comunalidades considerablemente altas. Cabe suponer que bajo estructuras menos favorables, como el modelo intermedio (Thurstone, 1947) el comportamiento de la regla empeore.

Aunque los resultados de la aplicación de la regla K1 no hayan resultado tan desastrosos como otros autores mantienen, eso no quiere decir que la extracción de Compo-

nentes Principales parezca la más apropiada para analizar estructuras que puedan ser enmarcadas en el modelo de Factor Común. La extracción de Componentes Principales, al trabajar sobre la matriz de correlaciones, sin reducir los elementos diagonales hacia las comunalidades, sobreestima las comunalidades finales, llegando a inflarlas de forma artificiosa hasta un incremento del 20% de la varianza explicada. Creemos que, si bien la regla K1 puede ser utilizada como una primera aproximación al número de factores subyacentes, es mejor utilizar extracciones como Ejes Principales o Máxima Verosimilitud, que analizan la matriz reducida, con las comunalidades estimadas en la diagonal (Mulaik, 1972; Kaiser y Derflinger, 1990). Autores como Velicer y Jackson (1990) recomiendan, sin embargo, la utilización del modelo de Componentes Principales frente al de Factor común (sobre esta discusión ver el volumen 25 de *Multivariate Behavioral Research*, 1990).

En cualquier caso, recomendamos utilizar métodos alternativos de estimación como el test de sedimentación, o criterios

de interpretabilidad de la estructura, además de esta regla, siendo conscientes de que la estructura rotada puede estar afectada por factores que han capitalizado el azar, por lo que tal vez sea necesario reanalizar nuestras muestras descartando alguno de los factores, con lo que la interpretabilidad puede mejorar.

## REFERENCIAS

- Bartlett, M. S. (1951). A further note on tests of significance in factor analysis. *British Journal of Psychology*, 4, 1-2.
- Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, 3, 77-85.
- Cattell, R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Cattell, R. B. y Vogelmann, S. A. (1977). A comprehensive trial of the Scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, 12, 289-325.
- Cliff, N. (1988). The eigenvalues-greater-than-1 rule and the reliability of components. *Psychological Bulletin*, 103, 276-279.
- Crawford, C. B. (1975). Determining the number of interpretable factors. *Psychological Bulletin*, 82, 226-237.
- Gorsuch, R. L. (1974). *Factor analysis*. Philadelphia: W.B. Saunders Co.
- Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, 18, 277-296.
- Hakstian, A. R., Rogers, W. T. y Cattell, R. B. (1982). The behavior of number of factor rules with simulated data. *Multivariate Behavioral Research*, 17, 193-219.
- Hakstian, A. R. y Muller, V. J. (1973). Some notes on the number of factors problem. *Multivariate Behavioral Research*, 461-475.
- Humphreys, L. G. (1964). Number of cases and number of factors: An example where N is very large. *Educational and Psychological Measurement*, 24, 457-466.
- Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika*, 31, 165-178.
- Kaiser, H. F. y Derflinger, G. (1990). Some contrasts between maximum likelihood Factor Analysis and Alpha Analysis. *Applied Psychological Measurement*, 14, 29-32.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.
- Lewis, P. A. y Orav, E. J. (1989). *Simulation Methodology for Statisticians, Operations Analysts and Engineers*. (Vol. 1). Belmont, C.A.: Wadsworth & Brooks /Cole.
- Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika*, 33, 37-71.
- Mote, T. A. (1970). An artifact of the rotation of too few factors: Study orientation vs. trait anxiety. *Revista Internacional de Psicología*, 37, 61-91.
- Mulaik, S. A. (1972). *The Foundations of Factor Analysis*. NY: McGraw-Hill.
- Revelle, W. y Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14, 403-414.
- Schöneman, P. H. (1990). Facts, fictions and common sense about factors and components. *Multivariate Behavioral Research*, 25, 47-51.
- Ten Berge, J. M. F. y Kiers, H. A. L. (1991). A numerical approach to the appropriate and exact minimum rank of a covariance matrix. *Psychometrika*, 56, 309-315.
- Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago Press.
- Velicer, W. F. y Jackson, D. N. (1990). Component Analysis versus Common Factor Analysis: Some issues in selecting an

- appropriate procedure. *Multivariate Behavioral Research*, 25, 1-28.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321-327.
- Yanai, H. D. e Ichikawa, N. (1990). New Lower and upper bounds for communality in factor analysis. *Psychometrika*, 55, 405-410.
- Zwick, W. R. y Velicer W. F. (1986). A comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432-442.