Revista Electrónica de Metodología Aplicada
1996, Vol. 1 nº 1, pp. 1-9
URL:http://www3.uniovi.es/user_html/herrero/REMA/v1n1/a1v1n1.wp5

CROSSFAC: DETERMINACION DEL NUMERO DE FACTORES COMUNES MEDIANTE PROCEDIMIENTOS DE VALIDACION CRUZADA

Pere Joan Ferrando
Urbano Lorenzo
Depto. de Psicología
Universidad Rovira i Virgili
e-mail:uls@astor.urv.es

ABSTRACT.
A library which allows to carry out single sample and split double sample cross validation procedures in the exploratory factor analysis model is presented. Theoretical aspects concerning the utility of the procedures in the assessment of the number of common factors to retain are discussed. An applied example is included to illustrate the use of the library.

Key words: Factor analysis, cross-validation.

1.- Introducción.

El conjunto de procedimientos conocidos genéricamente como 'validación cruzada', han sido utilizados tradicionalmente en diversas aplicaciones del modelo de regresión múltiple. Más en particular, dichos procedimientos han sido utilizados con relativa frecuencia en Psicología para la selección de predictores en estudios de validez referida a criterio. Al lector interesado en esta aplicación, se le recomiendan las lecturas de Mosier (1951); Horst (1966, cap. 23); McNemar (1969, cap. 11) y Lord y Novick (1968, cap. 13).

En épocas más recientes, Cudeck y Browne (1983; Browne y Cudeck, 1989, 1993) han extendido la lógica de la validación cruzada a los modelos de estructuras de covarianza (MEC), en base a la analogía que puede establecerse entre dichos modelos y el modelo no lineal de regresión con regresores fijos (Browne, 1982). La influencia que ha tenido el trabajo de Browne y Cudeck se pone de manifiesto, en primer lugar, en el hecho de que las versiones más recientes de algunos programas comerciales de MEC, (por ejemplo la versión 8 de LISREL) incluyen rutinariamente en el output indices de validación cruzada. Por otra parte, la revisión de trabajos de investigación que usan MEC, muestra que estos indices se incluyen ya con bastante asiduidad en las tablas que indican el ajuste de los modelos.

El trabajo que aquí se presenta consiste en el desarrollo de un paquete informático (CROSSFAC) que permite llevar a cabo diversos procedimientos de validación cruzada en

análisis factorial exploratorio (AFE). Dado que el AFE puede considerarse como un modelo particular de estructuras de covarianza (Browne, 1982) es inmediato que los procedimientos arriba citados resultan perfectamente utilizables en este caso.

El programa CROSSFAC permite obtener una serie de índices que pueden resultar de utilidad en la siempre difícil evaluación del número de factores comunes a retener. En línea con las ideas actuales acerca de la evaluación de la bondad de ajuste en MEC (véase por ej. Tanaka, 1993), los autores no creemos que la decisión acerca del número de factores deba basarse exclusivamente en los resultados de los procedimientos aquí considerados. Sin embargo, sí creemos que dichos resultados pueden aportar una información útil para esta evaluación que, además, es complementaria a la que aportan otras medidas más tradicionales de bondad de ajuste.

Respecto a la justificación de un programa de este tipo, cabe decir, en primer lugar, que los programas de ecuaciones estructurales que incluyen índices de validación cruzada, no suelen permitir (al menos en forma directa) llevar a cabo análisis factoriales exploratorios. Aparte de esto, en todo caso, los índices de validación reportados por estos programas son siempre índices obtenidos en una sola muestra. Los autores no tenemos noticia de ningún programa que permita llevar a cabo los procedimientos de validación cruzada simple y doble basados en la división al azar de la muestra disponible.

2.- Fundamentación y objetivos de los procedimientos utilizados.

Al margen de procedimientos e indicadores que carecen de una justificación clara, la determinación del número de factores comunes en un AFE suele basarse en alguna forma de evaluación de las covarianzas residuales que permanecen tras la extracción de un determinado número de factores. Dicha evaluación puede fundamentarse en indicadores puramente descriptivos (por ejemplo la raíz media cuadrática de los residuales) o bien en pruebas inferenciales de tipo más riguroso. Obviamente la selección en el primer caso puede ser muy arbitraria, dependiendo de lo que el investigador entienda por 'residuales suficientemente pequeños'.

En AFE bajo el método de máxima verosimilitud (ML) o el de mínimos cuadrados generalizados (GLS), es conocido que, bajo determinadas condiciones, el estadístico: $T = (N-1)*F$, donde N es el número de sujetos de la muestra y F el valor mínimo de la función de discrepancia, sigue asintóticamente una distribución Ji-cuadrado central, lo cual permite llevar a cabo una prueba de bondad de ajuste para un determinado modelo. Sin embargo las condiciones para que T se ajuste a dicha distribución teórica son bastante rigurosas. Entre otros supuestos, por una parte se asume que las variables analizadas siguen una distribución normal multivariante; por otra, se asume que el modelo considerado es totalmente correcto en la población.

Aparte de las posibles distorsiones provocadas por el no cumplimiento de los supuestos distribucionales, la asunción de que un modelo está perfectamente especificado es totalmente irreal. Los modelos son (y no pretenden otra cosa) aproximaciones a la realidad. Cuando un

investigador pone a prueba un modelo, en general, lo que pretende es evaluar si este modelo resulta plausible o si explica razonablemente bien las relaciones observadas entre variables. Como consecuencia de la implausibilidad de la hipótesis nula de perfecta especificación, es bien conocido que con una estricta adherencia a la prueba formal de hipótesis, se tenderá a la aceptación de modelos con más factores comunes de los que tienen realmente una interpretación substantiva.

Tal como los plantean Cudeck y Browne, los procedimientos de validación cruzada se proponen como una posible alternativa al problema hasta ahora descrito. Como punto de partida, se considera la idea de que un modelo es sólo una de las posibles aproximaciones que pretenden explicar una matriz de covarianzas (correlaciones). A partir de aquí, cabe probar diferentes modelos alternativos que pretenden explicar la misma matriz (en el contexto AFE modelos con diferente número de factores comunes) y decidir cual de ellos constituye la mejor aproximación de acuerdo con determinados criterios. En el caso particular de la validación cruzada, el 'mejor' modelo es el que da una mejor aproximación, no en la muestra en la que ha sido obtenido, sino en futuras muestras.

Para describir los procedimientos en el caso particular AFE, considérese, de entrada, el conocido modelo factorial para 'p' variables y 'k' factores comunes:

(1) $$\Sigma = \Lambda \Lambda' + \Psi$$

Donde Σ pxp es la matriz de correlación en la población, L pxk es la matriz patrón conteniendo las cargas factoriales y PSI de pxp es una matriz diagonal conteniendo las unicidades.

La estimación de los parámetros contenidos en L y PSI se basa en el criterio de minimizar una determinada función de discrepancia. En particular, aquí se consideran dos de ellas: la que corresponde al criterio de mínimos cuadrados ordinarios:

(2) $$Fuls = tr\left[((S-I)(\Lambda \Lambda' - diag\ (\Lambda \Lambda'))'((S-I)-\Lambda \Lambda' - diag\ \Lambda \Lambda'))\right]$$

y la correspondiente al criterio de máxima verosimilitud:

(3) $$Fml = \ln \det(\hat{\Sigma}) + tr(S^{-1}\hat{\Sigma}) - \ln \det(S) - p$$

Donde S es la matriz de correlación muestral y $\hat{\Sigma}$ la matriz reproducida por el modelo estimado.

La función (2) corresponde en particular al criterio MINRES (Harman y Jones, 1966) e implica la minimización de la suma de cuadrados de los valores residuales no diagonales. La función (3) es la que plantea Jöreskog (1967); minimizar (3) es equivalente a maximizar el logaritmo de la función de verosimilitud en el modelo AFE.

Supóngase que un investigador desea poner a prueba una serie de modelos con distinto número de factores comunes y que dispone de una muestra relativamente numerosa. En una primera etapa la muestra se divide al azar en dos mitades del mismo tamaño A y B. Seguidamente, se elige un criterio de minimización, y se estiman los parámetros correspondientes a los distintos modelos puestos a prueba utilizando la matriz de correlaciones obtenida en la muestra A. Una vez estimados los parámetros, se obtiene la matriz de correlaciones reproducida por el modelo mediante la expresión (1). Hecho esto, para cada uno de los modelos probados, puede obtenerse el valor de la función de discrepancia entre la matriz observada y la reproducida por el modelo.

Adviértase que, siguiendo el procedimiento hasta ahora descrito, la función irá dando valores cada vez menores cuantos más factores tenga el modelo. En el extremo, un modelo con $p\,(p+1)/2$ factores (modelo saturado) daría un valor 0 de discrepancia.

Considérese ahora la matriz de correlaciones obtenida en la muestra B. El valor de la función de discrepancia entre la matriz de correlaciones observada en B y la matriz reproducida desde el modelo estimado en A es el 'índice de validación cruzada'(IVC). En este caso, no necesariamente los modelos con más factores deberán tener un menor IVC. Como señalan Cudeck y Browne (1983), el modelo saturado produce un valor de 0 en la función de discrepancia en la misma muestra en que es estimado, pero, en cambio, su IVC puede ser bastante grande en relación al de otros modelos. En conjunto, la lógica del procedimiento es la de que, si en una muestra se extraen factores espurios, estos contribuirán a minimizar la función de discrepancia en la misma muestra en que se han obtenido, pero no se replicarán en otras muestras, pudiendo entonces empeorar el IVC.

El procedimiento que se acaba de describir es una validación cruzada simple. Fácilmente puede llevarse a cabo después una doble validación cruzada, estimando ahora los parámetros de los modelos en la muestra B y obteniendo los IVC respecto a la matriz de correlaciones observada en A. Intuitivamente se aprecia que si el mismo modelo es el que muestra el menor IVC en ambas direcciones, es bastante razonable suponer que este será el modelo con más estabilidad a través de distintas muestras.

Los procedimientos de validación simple y doble, dan indicadores puramente descriptivos. No se asumen supuestos distribucionales y, en consecuencia, puede elegirse libremente cualquier función de discrepancia.

En ocasiones, un investigador no dispondrá de una muestra suficientemente numerosa como para obtener estimaciones estables al dividirla en dos mitades. En estos casos, es posible obtener una estimación del valor esperado en el IVC a partir de una sola muestra, pero esto requerirá ya asumir ciertos supuestos distribucionales. Como es de suponer, la correcta interpretación de este valor observado dependerá de la medida en que los supuestos se cumplan.

Los dos supuestos principales son los que siguen.

En primer lugar, se asume que la función de discrepancia elegida es correcta para la

distribución de los datos. En el presente trabajo, la función a considerar es la que se deriva del criterio de máxima verosimilitud; esto significa asumir que los datos siguen una distribución normal multivariante.

En segundo lugar, se considera que el modelo no se ajusta exactamente en la población, pero si que se aproxima bastante. De ser así, si se cumplen también los supuestos distribucionales, (y otras condiciones no descritas que suelen cumplirse habitualmente), entonces el estadístico $T_{ML} = (N-1)*F_{ML}$ seguirá asintóticamente una distribución Ji-cuadrado no central, y el parámetro de no centralidad será el valor de F_{ML} en la población; es decir, el valor que se obtendría en la ecuación (3) si se utilizasen la matriz de covarianza obtenida en la población y la matriz reproducida por el modelo también en la población (que, por supuesto, son desconocidas). Adviértase pues que el parámetro de no centralidad es aquí un indicador del error de especificación del modelo.

Bajo los supuestos arriba descritos, Browne y Cudeck (1989, 1993), derivan cual sería el valor esperado en el IVC (EIVC) del siguiente modo. En primer lugar, se estima cual sería el valor esperado en el IVC si se tomase como muestra de calibración fija a la muestra observada y se fuesen tomando al azar de la población muestras de validación del mismo tamaño. En segundo lugar, se estima la esperanza del valor estimado en la etapa anterior, ahora a través de distintas muestras de calibración obtenidas en las mismas condiciones. El valor esperado en el IVC permite definir dos estimadores insesgados, de los que en este trabajo se ha elegido el más simple:

(4) $\hat{E}IVC = Fml + 2*q/ N-1$

Donde 'q' es el número de parámetros efectivos que debe estimar el modelo y que, en el caso concreto del AFE ML, viene dado por: $q= p*k + p - 1/2\ k(k-1)$ (Jöreskog, 1967).

Es posible, por último, suplementar el estimador puntual de (4) con unos intérvalos de confianza. Esto requerirá obtener los valores superior e inferior del parámetro de no-centralidad en una distribución Ji-cuadrado, con unos grados de libertad determinados y para un nivel de confianza especificado (Browne y Cudeck recomiendan el del 90%).

3.- Nota sobre los procedimientos utilizados.

Los procedimientos implementados en este trabajo que resultan más relevantes desde el punto de vista técnico y que pueden requerir justificación, son tres: el algoritmo para la solución factorial MINRES, el algoritmo para la solución factorial de máxima verosimilitud y el procedimiento para obtener los valores extremos en la distribución Ji-cuadrado no central.

Para la solución factorial MINRES se ha elegido la modificación propuesta por Zegers y ten Berge (1983) del algoritmo original de Harman y Jones (1966). La elección viene motivada por el hecho de que dicha modificación es más simple y rápida que el procedimiento original.

El programa elaborado para el análisis factorial de máxima verosimilitud se basa en el

procedimiento Newton-Raphson desarrollado inicialmente por Jennrich y Robinson (1969) y modificado más adelante por Jöreskog (1977). Se eligió este procedimiento a sugerencia de Browne (comunicación personal). Aparte de que mostró un comportamiento totalmente satisfactorio en diversos ensayos con soluciones conocidas, los resultados obtenidos en diversas pruebas coincidían con los proporcionados por los paquetes estadísticos al uso (SPSS y SAS).

Por último, para la estimación de los valores del parámetro de no-centralidad, se utilizó la aproximación a la distribución normal de la distribución Ji-cuadrado no central (véase Kendall y Stuart, 1967, vol II, p.229). Con este método cada extremo se obtenia como la raíz positiva de una ecuación no lineal que se resolvió mediante el método de Newton-Raphson. Al compararlo con los resultados obtenidos en programas standard, se vió que el procedimiento producía estimaciones satisfactorias a partir de los 10 grados de libertad, lo cual, en principio, parece suficiente desde un punto de vista práctico.

4.- Descripción de la libreria CROSSFAC.

Actualmente no se dispone de un programa que realize todos los cálculos de los procedimientos propuestos. No obstante, se ha desarrollado una libreria de funciones en el lenguaje matricial MATLAB que sí lo permite. Si bien MATLAB puede ser un instrumento inconveniente por ciertas circunstancias técnicas (como el trabajar muy lentamente con valores escalares), resulta, en cambio, muy apropiado por la precisión de su rutina de diagonalización en valores y vectores propios, un aspecto de vital importancia en la estimación ML (Jöreskog, 1977). Por otra parte, permite que el mismo algoritmo sea directamente aplicado indiferentemente del entorno de trabajo (DOS, Mac, Windows 95, UNIX, etc).

Las funciones implementadas en Matlab son las siguientes:

HALFS: Esta función divide la muetra inicial en dos muestras, siendo cada una de las nuevas muestras una mitad aleatoria de la muestra inicial. Nótese que Matlab ejecuta esta función con un gran coste computacional en términos de tiempo. Así pues, en al caso de que la muestra inicial se considere lo suficientemente aleatorizada, la función HALFS puede ser sustituida por las siguientes lineas de comando:

```
A = X(1:N/2,:);
B = X((N/2)+1:N,:);
```

donde X es la matriz que contiene la muestra original, N el número de casos en la muestra, A la primera mitad de la muestra aleatoria y B la segunda mitad de la muestra aleatoria.

EIGENC: calcula la descomposición de valores y vectores propios tal que los valores y los vectores se hallan ordenados en orden creciente (d1 < d2 < … < dk).

PCA: obtiene una solución en componentes principales estandarizados para el número de componentes especificado (Meredith y Millsap, 1985). Se utiliza cuando se necesita una estimación inicial del patrón factorial.

MINRES: obtiene el patrón factorial mediante el criterio de mínimos cuadrados a partir de la matriz de correlaciones para el número de factores espeficicado (Zegers y ten Berge, 1983).

ML77: obtiene el mínimo de la función de máxima verosimilitud en el análisis factorial de la matriz de correlaciones para un número determinado de factores comunes (Jöreskog, 1977). Esta función puede ser fácilmente ampliada para obtener el patrón factorial resultante de la aplicación de dicho método.

LAMBDA: calcula los intérvalos de confianza para el parámetro de no centralidad en la distribución Ji-cuadrado.

LEDERMAN: estima el máximo de factores que podrían ser extraidos de la muestra de acuerdo con la desigualdad de Lederman. En el caso que de se pretendan extraer más factores de los que pueden ser determinados, el número de factores máximo a extraer se corrige.

CROSSFAC: es la función que realiza la validación cruzada y total propiamente. La ejecución del mismo requiere ejecutar la siguiente orden desde la línea de comandos de MATLAB,

crossfac (X,i,s)

donde X es la matriz que contine la muestra total, 'i' y 's' son el mínimo y el máximo respectivamente del intervalo de factores que va a ser testado. Nótese que esta línea de comandos ha de ser ejecutada desde el directorio donde todas las funciones anteriores han sido previamente grabadas.

5.- Un ejemplo de la aplicación de Crossfac.

Para estudiar el comportamiento de los procedimientos descritos, se llevó a cabo un estudio de simulación. De este modo, se tenía un conocimiento previo del número de factores comunes de la solución y lo que se pretendía era evaluar hasta que punto los diversos procedimientos implementados en la librería permitían recuperar correctamente dicho número.

Inicialmente se diseño un patrón de 9 variables por tres factores totalmente ajustado al criterio de estructura simple en el que todas las variables eran canónicas (es decir, tenían una saturación de 1 en un factor y de 0 en los restantes). Las tres primeras variables definían al primer factor, las tres siguientes al segundo y las tres últimas al tercero.

Seguidamente se generó una matriz de 1000 observaciones por 9 variables, de tal forma que la matriz de correlación coincidiera exactamente con la matriz reproducida desde el patrón que se acaba de describir. Por último, se generaron 9 variables aleatorias normales que se sumaron ponderadamente a cada una de las variables anteriores. De esta forma, cada variable se ajustaba en parte al modelo y, en parte, contenía error aleatorio. En consecuencia, al factorizar la matriz de correlación, debía obtenerse una clara solución en tres factores y un patrón muy similar (aunque no exactamente igual) al diseñado inicialmente.

Para ilustrar las posibilidades de Crossfac, se llevaron a cabo los procedimientos de doble validación cruzada utilizando el algoritmo MINRES y los procedimientos por máxima verosimilitud en una sola muestra. En ambos casos, se evaluaron modelos desde 0 hasta 4 factores comunes. Los resultados se presentan a continuación.

Tabla I. Output de Crossfac en el estudio de simulación

Model	A over B		B over A	
FAC n	F	CVI	F	CVI
0	8.8030	9.0998	9.0998	8.8030
1.0000	0.4210	0.3541	0.3422	0.4299
2.0000	0.1863	0.1690	0.1297	0.2099
3.0000	0.0066	0.0331	0.0058	0.0396
4.0000	0.0037	0.0401	0.0023	0.0430

Model	Total				
FAC n	F	Gl	ECVI	90% conf.	interval
0	5.2314	36.0000	5.2494	5.0147	5.4915
1.0000	1.0808	27.0000	1.1168	1.0127	1.2284
2.0000	0.3918	19.0000	0.4438	0.3830	0.5122
3.0000	0.0121	12.0000	0.0782	0.0737	0.0910
4.0000	0.0049	6.0000	0.0830	0.0841	0.0925

La solución es muy clara. Nótese que en todos los casos el índice de validación cruzada alcanza su valor mínimo en el número correcto de factores, es decir, tres; y luego vuelve a subir. Por otra parte, el hecho de que no se mejora a partir de tres factores, se evidencia también en el solapamiento de los intérvalos de confianza correspondientes a las soluciones en tres y cuatro factores.

6.- Disponibilidad de Crossfac.

La librería CROSSFAC sobre MATLAB está disponible para todo lector que la solicite. Actualmente estamos desarrollando un programa para entorno DOS que realice los procedimientos propuestos, junto a un breve manual. Si el lector está interesado en esta segunda opción, póngase en contacto con los autores.

7.- Referencias.

Browne, M.W. (1982) Covariance structures. En D.M. Hawkins (ed.), **Topics in applied multivariate analysis**. Cambridge. Cambridge univ. press.

Browne, M.W y Cudeck, R. (1989) Single sample cross-validation indices for covariance structures. **Multivariate Behavioral Research**, **24**, 445-455.

Browne, M.W y Cudeck, R. (1993) Alternative ways of assessing model fit. En K.A. Bollen y J.S. Long (Eds.), **Testing structural equation models** . Newbury Park: Sage.

Cudeck, R. y Browne, M.W. (1983) Cross-validation of covariance structures. **Multivariate Behavioral Research**, **18**, 147-167.

Harman, H.H. y Jones, W.H. (1966) Factor analysis by minimizing residuals (Minres). **Psychometrika**, **31**, 351-369.

Horst, P. (1966) **Psychological measurement and prediction**. Belmont. Wadsworth.

Jennrich, R.I. y Robinson, S.M. (1969) A Newton-Raphson algorithm for maximum likelihood factor analysis. **Psychometrika**, **34**, 111-123.

Jöreskog, K.G. (1967) Some contributions to Maximum Likelihood factor analysis. **Psychometrika**, **32**, 443-482.

Jöreskog, K.G. (1977) Factor analysis by Least-squares and Maximum-Likelihood methods. En K. Enslein, A. Ralston y H.S. Wilf (Eds.), **Statistical methods for digital computers**, Vol. 3. New York. Wiley.

Kendall, M.G. y Stuart, A. (1967) **The advanced theory of statistics**, Vol II. London. Griffin.

Lord, F.M. y Novick, M.R. (1968) **Statistical theories of mental tests scores**. Massachusetts. Addison-Wesley.

McNemar, Q. (1969) **Psychological statistics**. New York. Wiley.

Meredith, W. y Millsap, R.E. (1985). On component analysis. **Psychometrika**, **50**, 495-507.

Mosier, C.I. (1951) Problems and designs of cross-validation. **Educational and Psychological Measurement**, **11**, 5-11.

Tanaka, J.S. (1993) Multifaceted conceptions of fit in structural equation models. En K.A. Bollen y J.S. Long (Eds.), **Testing structural equation models**. Newbury Park: Sage.

Zegers, F.E. y ten Berge, J.M.F. (1983) A fast and simple computational method of minimum residual factor analysis. **Multivariate Behavioral Research**, **18**, 331-340.