© Revista de Matemática: Teoría y Aplicaciones 2020 **27**(2) : 383–404
cimpa – ucr   issn: 1409-2433 (Print), 2215-3373 (Online)
doi: https://doi.org/10.15517/rmta.v27i2.33677

ROBUSTEZ DE LOS ÍNDICES DE AJUSTE DEL ANÁLISIS FACTORIAL CONFIRMATORIO A LOS VALORES EXTREMOS

ROBUSTNESS OF CONFIRMATORY FACTOR ANALYSIS FIT INDICES TO OUTLIERS

Luis Rojas-Torres *

Received: 26/Jun/2018; Revised: 27/Sep/2019;
Accepted: 6/Feb/2020

*Universidad de Costa Rica, Escuela de Matemática, San José, Costa Rica. E-Mail: luismiguel.rojas@ucr.ac.cr

Resumen

El objetivo de este trabajo es evaluar la robustez de varios índices de ajuste del Análisis Factorial Confirmatorio (SRMR, RMSEA, TLI, CFI y GFI) a la presencia de valores extremos. Para el alcance del objetivo se planteó un estudio de simulaciones con $3\times\ 4 \times 2$ condiciones: tamaño de muestra (100, 200 y 500), porcentaje de valores extremos (0%, 1%, 5% y 10%) y número de variables con valores extremos (1 y 2). Los conjuntos de datos base (0% de valores extremos) por tamaño de muestra, fueron simulados a partir de una distribución que se ajustaba a un AFC con tres factores correlacionados. Los conjuntos de datos con valores extremos se crearon a partir de la sustitución de observaciones en los conjuntos de datos base. Luego, en cada conjunto de datos se estimó un AFC con tres factores correlacionados. Se obtuvo que con los puntos de corte clásicos de los índices de ajuste, todos estos fueron robustos a la presencia de datos extremos en conjuntos de datos de 200 y 500 observaciones. Con 100 observaciones se obtienen índices robustos pero considerando puntos de corte ajustados a la estructura factorial y el tamaño muestral trabajado.

Palabras clave: valores extremos; análisis factorial confirmatorio; robustez; estudio de simulación; error Monte Carlo.

Abstract

The goal of this work is to evaluate the robustness of several Confirmatory Factor Analysis fit indices (SRMR, RMSEA, TLI, CFI and GFI) to the precense of outliers. For this purpose, it was planed a simulation study with $3 \times 4 \times 2$ conditions: sample size (100, 200 and 500), outliers percentage (0%, 1%, 5% and 10%) and number of variables with outliers (1 and 2). The baseline data sets (0% of outliers) by sample size were simulated from a distribution which fit to a CFA with three factors correlated. Data bases with outliers were created from substitution of observations in baseline data sets. Later, in every data base was estimated a CFA with three factors correlated. It was obtained that all indices with classical cutoffs were robust to outliers with sample sizes of 200 and 500. With 100 observations, it was obtained that fit indexes were robust to outliers, but considering cutoffs adjusted by the factorial structure and the sample size.

Keywords: outliers; confirmatory factor analysis; robustness; simulations study; Monte Carlo error.

Mathematics Subject Classification: 62H25.

1 Introducción

Uno de los temas más importantes en la estadística es determinar si un modelo estadístico realmente se ajusta a los datos estudiados. Debido a lo anterior, junto con el desarrollo de un modelo estadístico, también se construyen índices de ajuste, los cuales permiten evaluar el ajuste del modelo a los datos. Estos índices se elaboran a partir de propiedades que deben satisfacer elementos del modelo estimado, como el ajuste a una distribución o la obtención de un mejor ajuste a los datos observados, que la obtenida por un modelo que teóricamente no debería ajustar.

Varios de estos índices tienen una distribución estadística asociada, bajo el supuesto de que el modelo ajusta a los datos, por lo cual el rechazo del modelo depende de que el valor observado del índice sea poco probable bajo la distribución planteada. Otros índices no tienen distribución asociada, por lo que su uso depende del estudio empírico de los valores que toman los índices en distintas situaciones. Este estudio, por lo general, se realiza por medio de simulaciones Monte Carlo, en las cuales se analizan los valores de los índices del modelo correcto en distintas bases de datos, en las que el modelo teóricamente debe ajustar; y también se estudian los valores del índice en otros modelos aplicados a las mismas bases, los cuales teóricamente no deben ajustar. Luego, con base en los valores observados del índice, se establece un punto de corte en el que se minimiza el error tipo I (rechazo de modelos apropiados) y se maximiza la potencia (rechazo de modelos inapropiados). Esta metodología del punto de corte, también se puede utilizar con índices que tenga una distribución asociada, dado que su uso estricto puede ser muy exigente para los modelos bien especificados o muy permisiva para los modelos mal especificados.

En el caso del Análisis Factorial Confirmatorio (AFC) existen una gran cantidad de índices que se han elaborado para este propósito [3]. El índice de ajuste más utilizado en la evaluación del AFC es el índice χ^2, el cual se define a partir de una función F basada en la distribución teorizada de los datos (las fórmulas asociadas al modelo AFC y a los índices señalados se explícitan más adelante). Esta formulación del χ^2 permite evaluar si el modelo planteado se ajusta a la distribución teorizada. No obstante, a pesar de que la definición del índice está abierta a varias distribuciones, la inmensa mayoría de los desarrollos hechos con este índice se basan en la distribución multinormal. Además, como su nombre lo indica, este índice tiene una distribución asociada, la cual es χ^2 con k grados de libertad, los cuales equivalen a la cantidad de parámetros libres del modelo.

Ahora bien, a pesar de la formulación mencionada previamente del estadístico χ^2, se ha observado que este estadístico es robusto a algunas violaciones de la normalidad, es decir, logra discriminar los modelos que cumplen la estructura factorial deseada, aún cuando las variables observadas no están distribuidas multinormalmente. En un estudio realizado con simulaciones Monte Carlo, se observó que un modelo factorial de tres factores con cinco indicadores por variable latente, este índice fue robusto a las violaciones provocadas por la no normalidad de los factores comunes y los errores de las variables, derivadas de las variaciones en la curtosis, pero no fue robusto a la dependencia entre los errores y los factores comunes [4].

Por otro lado, en [3] se analizó cuáles índices de ajuste eran sensibles a las mismas violaciones a la normalidad presentadas en [4], controlando los factores: tamaño de muestra, mala especificación del modelo y método de estimación. En este estudio se simularon bases que se ajustaban a un AFC de tres factores correlaciones, con cinco indicadores por variable. Se obtuvo que tres índices basados en el χ^2, el CFI, el TLI y el RMSEA, fueron levemente sensibles a las violaciones a la normalidad, es decir que el valor del índice tuvo leves cambios, en la dirección contraria del valor ideal, cuando se consideraron violaciones a la normalidad.

Otro índice basado en el supuesto de la distribución de los datos es el GFI. En el estudio de [3] también se analizó este índice considerando la multinormalidad, se observó que este no era sensible a las violaciones de la normalidad, controlando los demás factores. Además, también se estudió el comportamiento del SRMR, que es un estadístico que depende exclusivamente de los errores observados, por lo cual, no depende del supuesto de la distribución. se obtuvo que este estadístico fue levemente sensible a las violaciones de normalidad.

Luego, otra violación a la multinormalidad de los datos puede ser provocada por la contaminación de los datos, debido a los valores extremos. En [11] se define un valor extremo en un modelo multivariado como aquella observación $\mathbf{x}$ que se encuentre muy lejana del centroide de las observaciones. Con base en lo anterior, si d es la variable aleatoria que representa las distancias de todas las observaciones al centroide, se puede definir que una observación es un valor extremo si su distancia al centroide es mayor que $\bar{d}+3s_d$, donde s_d es la desviación estándar de d.

En [8] se analizó la robustez a la contaminación de los datos de un AFC de nueve indicadores, con dos factores definidos por tres indicadores y uno definido por cuatro indicadores, lo cual implica que un indicador cargaba en dos factores. La contaminación se analizó con datos extremos (outliers) y valores influenciales (bad influential values) y con diferentes proporciones de datos contaminados. Con el método de máxima verosimilitud, se observó que el χ^2, el CFI y el RMSEA presentaron tasas altas de rechazo con los dos tipos de contaminación, las cuales fueron más pronunciadas a mayor nivel de proporción de contaminación y en el caso de valores influenciales. Los índices con mayores tasas de rechazo fueron el χ^2 y el RMSEA, estas tasas llegaron a alcanzar valores mayores que 0.95. Lo anterior indicó que los índices de ajuste con el método de máxima verosimilitud no proporcionan un criterio robusto a la contaminación de los datos.

En el estudio de [8] se analizó un tipo de contaminación que afectaba a todos los valores observados. En este estudio se desea analizar si los índices de ajuste del método de máxima verosimilitud proporcionan un criterio robusto a una contaminación menos drástica de los datos, en específico, a la contaminación de una o dos variables. Este caso es relevante, ya que en algunos conjuntos de datos se pueden encontrar problemas porque algunos sujetos presentan valores extremadamente altos en una única variable.

Esta situación fue observada en [13], en un estudio en el que se esperaba una estructura factorial de tres factores con tres indicadores, se observaron problemas de estimación debido a que un grupo de observaciones presentaban valores particularmente altos en uno de los indicadores.

Ahora bien, el objetivo de este estudio es analizar la robustez de los índices del método de máxima verosimilitud del AFC a la contaminación de los datos en una o dos variables, para este propósito se desarrolla un estudio de simulación de Monte Carlo con distintas proporciones de contaminación y distintos tamaños de muestra. Antes de desarrollar el estudio se presenta una breve descripción de los conceptos del AFC y de los índices de ajuste.

1.1 Análisis factorial confirmatorio

La técnica estadística del Análisis Factorial Confirmatorio fue presentada por primera vez en la literatura científica, en el artículo de [5], titulado *A general approach to confirmatory maximum likelihood factor analysis*. El objetivo de esta técnica es evaluar los modelos que plantean que un conjunto de variables no observadas (latentes) predicen la variabilidad de un conjunto de variables observadas.

Esta técnica supone que las variables observadas (x_i) son explicadas por una combinación lineal de las variables latentes (ξ_j), sometidas a varias restricciones. Por tanto, estas asociaciones pueden representarse por medio de un conjunto de ecuaciones de regresión. Entonces, para n variables observadas y m variables latentes, se plantean la siguientes ecuaciones de regresión:

$$\begin{aligned}
x_1 &= \lambda^x_{11}\xi_1 + \lambda^x_{12}\xi_2 + \ldots + \lambda^x_{1m}\xi_m + \varepsilon_1,\\
x_2 &= \lambda^x_{21}\xi_1 + \lambda^x_{22}\xi_2 + \ldots + \lambda^x_{2m}\xi_m + \varepsilon_2,\\
x_3 &= \lambda^x_{31}\xi_1 + \lambda^x_{32}\xi_2 + \ldots + \lambda^x_{3m}\xi_m + \varepsilon_3,\\
x_4 &= \lambda^x_{41}\xi_1 + \lambda^x_{42}\xi_2 + \ldots + \lambda^x_{4m}\xi_m + \varepsilon_4,\\
&\vdots\\
x_n &= \lambda^x_{nm}\xi_1 + \lambda^x_{n2}\xi_2 + \ldots + \lambda^x_{nm}\xi_m + \varepsilon_n,
\end{aligned} \tag{1}$$

donde λ^x_{ij} representa el coeficiente de regresión de la variable x_i sobre la variable latente ξ_j y, ε_i representa el error asociado a la regresión de x_i.

Luego, las restricciones mencionadas se plantean estableciendo condiciones sobre los coeficientes de regresión u otros parámetros que intervienen en la estimación del modelo, que se presentan más adelante. Una condición muy usual es fijar varios coeficientes de regresión a 0. Por ejemplo, se puede suponer que tres variables observadas (x_1, x_2 y x_3) son explicadas por una variable latente (ξ_1), mientras que otras tres (x_4, x_5 y x_6) son explicadas por otra variable latente (ξ_2), pero ξ_1 no se asocia con las variables explicadas por ξ_2 y viceversa. Las ecuaciones de regresiones asociadas a este ejemplo se presentan en la expresión (2) y su representación gráfica se muestra en la Figura 1.

$$\begin{aligned}
x_1 &= \lambda^x_{11}\xi_1 + 0\cdot\xi_2 + \varepsilon_1,\\
x_2 &= \lambda^x_{21}\xi_1 + 0\cdot\xi_2 + \varepsilon_2,\\
x_3 &= \lambda^x_{31}\xi_1 + 0\cdot\xi_2 + \varepsilon_3,\\
x_4 &= 0\cdot\xi_1 + \lambda^x_{42}\xi_2 + \varepsilon_4,\\
x_5 &= 0\cdot\xi_1 + \lambda^x_{52}\xi_2 + \varepsilon_5,\\
x_6 &= 0\cdot\xi_1 + \lambda^x_{62}\xi_2 + \varepsilon_6.
\end{aligned} \tag{2}$$

Ahora bien, para la estimación del AFC, las ecuaciones planteadas en (1) se agrupan en la siguiente ecuación matricial:

$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{3}$$

donde $\mathbf{x}$ es el vector de variables independientes observadas, con dimensión $n \times 1$; $\boldsymbol{\xi}$ es el vector de variables latentes ($m \times 1$); $\boldsymbol{\Lambda}_x$ es una matriz $n \times m$, donde se representan los coeficientes de regresión de las variables observadas sobre las variables latentes y $\boldsymbol{\varepsilon}$ es un vector $n \times 1$ que representa los errores de estimación en las regresiones. A partir de la fórmula (3), se puede concluir que la matriz de variancias de las variables observadas basada en el AFC está dada por la fórmula:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}'_x + \boldsymbol{\Theta}_\varepsilon,$$

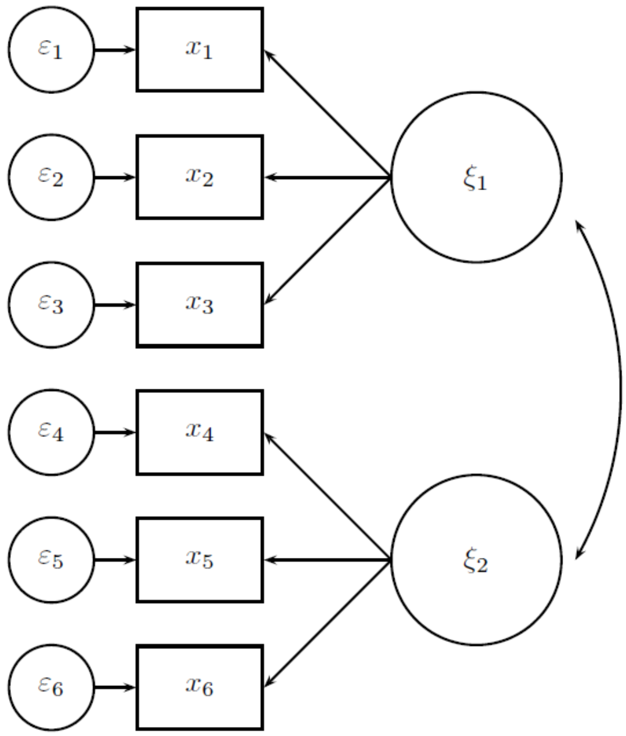

Figura 1: Ejemplo de un AFC.

en la que $\mathbf{\Phi}$ representa la matriz $m \times m$ de variancias y covariancias de las variables latentes y $\mathbf{\Theta}_\varepsilon$, la matriz $n \times n$ de variancias y covariancias de los errores de las regresiones [6]. Luego, la matriz $\mathbf{\Sigma}$ puede ser aproximada por la matriz $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma}(\mathbf{\Omega})$, la cual es el resultado de evaluar la fórmula de $\mathbf{\Sigma}$ en el vector $\mathbf{\Omega}$, compuesto por las estimaciones de los coeficientes de regresión y las varianzas de las variables latentes y los errores.

Con base en lo anterior, se obtiene que una forma de estimar los parámetros del AFC es por medio de la búsqueda del conjunto $\mathbf{\Omega}$ que permita que la matriz de varianzas teórica basada en los coeficientes ($\widehat{\mathbf{\Sigma}}$) sea lo más semejante posible a la matriz de varianzas muestral ($\mathbf{S}$). Para dicho objetivo, se plantea una función $F(\mathbf{\Omega})$ que estime una medida continua positiva de semejanza entre las matrices y se busca su mínimo. Esta función puede ser estimada considerando elementos tales como el tamaño de los residuos o la distribución de los datos observados. En particular, se puede considerar la función de discrepancia de máxima verosimilitud, la cual se basa en que las variables observadas provienen de una distribución multinormal con matriz de varianzas $\mathbf{\Sigma}$, la fórmula de esta función es:

$$F(\mathbf{\Omega}) = \log|\widehat{\mathbf{\Sigma}}| - \log|\mathbf{S}| + \text{tr}(\mathbf{S}\widehat{\mathbf{\Sigma}}^{-1}) - n.$$

1.2 Índices de ajuste

Uno de los elementos que permite evaluar si el modelo de AFC propuesto se ajusta a los datos son los índices de ajuste. Los índices que evalúan el ajuste global del modelo se denominan de ajuste absoluto y los que comparan el ajuste del modelo con el del modelo nulo (modelo que suponen que las variables observadas son independientes), se denominan de ajuste incremental [2]. Entre los índices de ajuste absoluto están la raíz cuadrada de la media de residuos cuadrados (SRMR), el de bondad de ajuste (GFI), la razón de verosimilitud (χ^2) y la raíz del cuadrado medio del error de aproximación (RMSEA). Entre los índices de ajuste incremental están el de Tucker Lewis (TLI) y el de ajuste comparativo (CFI). Las fórmulas de estos índices son:

$$\begin{aligned}
\text{GFI} &= 1 - \text{tr}(\mathbf{S}\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{I})/\text{tr}(\mathbf{S}\widehat{\mathbf{\Sigma}}^{-1}) \quad \text{[La fórmula varía según,} \\
&\quad \text{la función de discrepancia],} \\
\text{SRMR} &= \sqrt{2\sum_{i=1}^{n}\sum_{j=1}^{i}(s_{ij} - \hat{\sigma}_{ij})^2/[n(n+1)]} \\
\chi^2 &= F \cdot N, \\
\text{RMSEA} &= \sqrt{\min\{(\chi^2 - df)/(N-1), 0\}/df}, \\
\text{TLI} &= (\chi_0^2/df_0 - \chi^2/df)/(\chi_0^2/df_0 - 1), \\
\text{CFI} &= 1 - \max(\chi^2 - df, 0)/\max(\chi^2 - df, \chi_0^2 - df_0, 0),
\end{aligned}$$

donde $\mathbf{I}$, es la matriz identidad, s_{ij}, la covarianza observada entre los indicadores j e i y $\hat{\sigma}_{ij}$, la covarianza estimada, n, el número de variables observadas, F, el valor final de la función de discrepancia, N, el número de sujetos de la muestra, df, los grados de libertad del modelo, χ_0^2, la razón de verosimilutd ajustada en el modelo nulo y df_0, los grados de libertad del modelo nulo.

El GFI indica la propoción de la información dada por la estimación con respecto a la información dada por la matriz muestral [10]. EL SRMR proporciona un promedio de las desviaciones de los elementos de las matrices $\mathbf{S}$ y $\widehat{\mathbf{\Sigma}}$.

La razón de verosimilitud χ^2 es el índice de ajuste más utilizado, dado que permite evaluar la hipótesis $\mathbf{\Sigma} = \widehat{\mathbf{\Sigma}}$, no obstante, esta prueba de hipótesis es muy exigente, por lo cual se prefieren utilizar correcciones a este índice. De hecho, es conocido que en muchos casos es imposible que la hipótesis anterior pueda ser cierto, debido a que no existe un $\mathbf{\Omega}$ que permita alcanzar la igualdad. Lo anterior implica que hay que considerar el error de aproximación de la mejor estimación posible $\widehat{\mathbf{\Sigma}}$ a $\mathbf{\Sigma}$, el cual puede ser aproximado con la expresión $\chi^2 - df$, que es considerada en el RMSEA. Por tanto, el RMSEA indica que tan cercana está la mejor aproximación de $\widehat{\mathbf{\Sigma}}$, con la verdadera matriz $\mathbf{\Sigma}$.

Por otro lado, el TLI indica la proporción de mejora en los valores esperados de los χ^2 que aporta el modelo estimado con respecto al modelo nulo. Por su parte, el CFI indica la proporción de mejora en el estadístico de ajuste basado en la no centralidad, $\chi^2 - df$, el cual, al igual que el RMSEA, considera que la mayoría de veces, la matriz de varianzas estimada, aún con todos los datos de la población, no puede ser igual a la matriz de varianzas verdadera [9].

2 Metodología

Para el alcance del objetivo se planteó un estudio de simulaciones Monte Carlo, con bases de datos conformadas por nueve variables observadas, que se ajustaban a un diseño factorial de tres factores correlacionados con errores de regresión independientes entre sí. Este modelo es de interés, debido a que representa una de las estructuras factoriales más utilizadas en los estudios sociales, de hecho, es el ejemplo base de [14], también aparece en los estudios de [13] y [8]; por otro lado, también es usado como modelo base en [3] y [4], solo que con 5 indicadores por factor.

Los coeficientes de regresión de las variables sobre los factores del modelo se fijaron a 0.7 y las correlaciones entre los factores se establecieron en 0.5, se escogieron estos valores de λ y ϕ para garantizar que las variables indicadoras definían bien los factores latentes y que estos estaban correlacionados entre sí. Luego, el valor $\lambda = 0.7$ representa una carga factorial alta, bajo el supuesto de que la variable observada está estandarizada, debido a esto las varianzas del error se fijaron en 0.51, ya que la suma de la variabilidad debido a la variable latente (0.7^2) y al error (.51) equivale a la variabilidad de la variable observada. Además, para establecer la métrica de las variables latentes, se fijó su varianza a 1. La representación gráfica del modelo de simulación se presenta en la Figura 2.

Luego, para la simulación de los datos se supuso que las observaciones del modelo provenían de una normal multivariada de nueve entradas (las variables del modelo), cuyas medias eran iguales a 0 y cuya matriz de varianzas era la matriz $\mathbf{\Sigma}$ derivada de las condiciones establecidas en el modelo. De esta manera, para un vector de observaciones $\mathbf{x}$ del conjunto de datos de simulación, se tiene que, $\mathbf{x} \sim \mathbf{N}_9(\mathbf{0}_9, \mathbf{\Sigma})$. Las condiciones del modelo de simulación indican que las

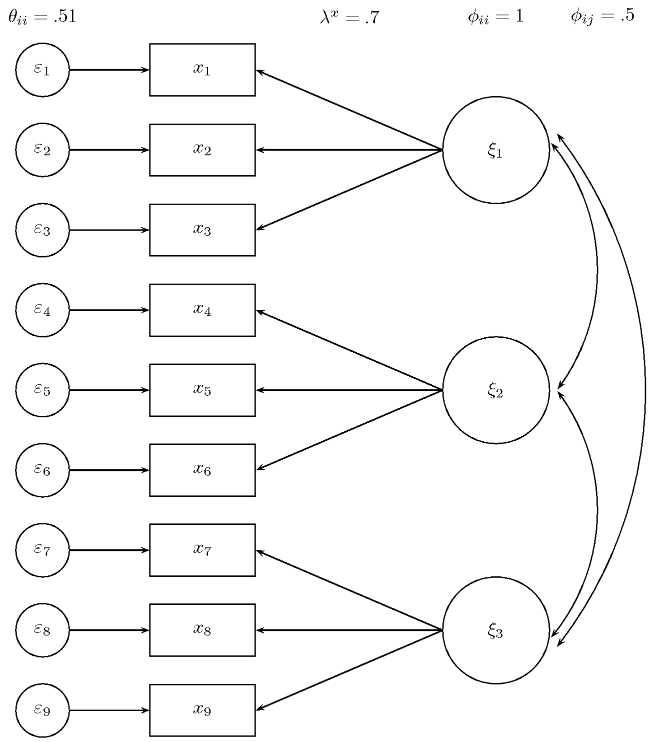

Figura 2: Modelo de simulación.

matrices involucradas en la matriz de varianzas fueron las siguientes:

$$\mathbf{\Lambda}_x = \begin{pmatrix} .7 & 0 & 0 \\ .7 & 0 & 0 \\ .7 & 0 & 0 \\ 0 & .7 & 0 \\ 0 & .7 & 0 \\ 0 & .7 & 0 \\ 0 & 0 & .7 \\ 0 & 0 & .7 \\ 0 & 0 & .7 \end{pmatrix}, \mathbf{\Phi} = \begin{pmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{pmatrix},$$

$$\mathbf{\Theta}_\varepsilon = \begin{pmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ 0 & 0 & 1 & & & & & & \\ 0 & 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Por tanto, la matriz de varianzas verdadera tuvo la siguiente forma

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & & & & & & & & \\ .49 & 1 & & & & & & & \\ .49 & .49 & 1 & & & & & & \\ .245 & .245 & .245 & 1 & & & & & \\ .245 & .245 & .245 & .49 & 1 & & & & \\ .245 & .245 & .245 & .49 & .49 & 1 & & & \\ .245 & .245 & .245 & .245 & .245 & .245 & 1 & & \\ .245 & .245 & .245 & .245 & .245 & .245 & .49 & 1 & \\ .245 & .245 & .245 & .245 & .245 & .245 & .49 & .49 & 1 \end{pmatrix}.$$

Una vez establecida las condiciones del modelo base de simulación, se definieron las variantes de la simulación, las cuales fueron el tamaño de muestra ($n = 100$, $n = 200$ y $n = 500$), el porcentaje de valores extremos de la simulación ($p = 1\%$, $p = 5\%$ o $p = 10\%$) y el número de variables del primer factor en las que se reparten los valores extremos ($m = 1$ o $m = 2$). Por tanto, se establecieron $3 \times 4 \times 2 - 1 = 17$ condiciones de simulación. El caso restado es la variante de $n = 100$, $p = 1\%$ y $m = 2$, ya que en este caso solo aparece un valor extremo y, por ende, no se puede repartir entre dos variables.

Para este trabajo se simularon $N = 500$ conjuntos de datos del modelo base de simulación, por cada tamaño de muestra, la cantidad N de replicaciones se determinó por medio del análisis de estabilidad del error de Monte Carlo. Luego, para generar los conjuntos de datos asociados a un tamaño de muestra n y a $p\%$ de valores extremos en una variable, en cada conjunto de datos de tamaño n se seleccionaron $p \cdot n$ datos de x_1 y, se sustituyeron por datos provenientes de una variable uniforme con parámetros $a = 10$ y $b = 11$. Los datos provenientes de esta uniforme son valores extremos en x_1, ya que dicha variable es una normal estándar y aquellos valores se ubican a más de 10 desviaciones estándar del promedio.

Por otro lado, los conjuntos de datos extremos asociados a un tamaño de muestra n con $p * \%$ de valores extremos en dos variables se realizaron de forma similar a los basados en una variable, solo que primeramente se seleccionaron $n \cdot p/2$ valores de x_1 y luego, $n \cdot p/2$ de x_2, con la condición de que no hubiesen valores x_1 y x_2 que pertenecieran a un mismo vector de observaciones $\mathbf{x}$ (* indica que la sustitución por valores extremos se repartieron entre dos variables).

Es importante indicar, que en un conjunto de datos de $n = 1000$ simulado con las condiciones del modelo base de simulación, se obtuvo que las observaciones presentaron una distancia d de Mahalanobis[1] promedio de 3.56 (1.62). En cambio, las distancias de Mahalanobis para las observaciones simuladas como valores extremos presentaron valores promedio de 12.36, 22.32, 31.02, 10.19, 17.72 y 24.42, para m =1%, 5%, 10%, 1*%, 5*%, 10*%, respectivamente. Este aumento de la distancia indicó que las nuevas observaciones, efectivamente representaban valores extremos, ya que eran mayores que $\bar{d} + 3s_d$.

Posteriormente, en los conjuntos de datos simulados para cada condición se estimó el modelo factorial que representa adecuadamente a los datos: un modelo de tres factores correlacionados, definidos por la varianza común de tres variables observadas. Luego, en los conjuntos de datos de cada condición se calcularon las tasas de rechazo del modelo adecuado, con base en cinco distintos índices: tres de ajuste absoluto (SRMR, RMSEA y GFI) y dos índices de ajuste incremental (TLI y CFI). Los puntos de corte utilizados para evaluar el ajuste fueron los puntos de corte clásicos (0.08 para el SRMR, 0.05 para el RMSEA y 0.95 para el resto, [1]) y unos puntos de corte ajustados, los cuales representan el valor del índice donde se obtiene una tasa de rechazo de 5% en 1000 bases de datos simuladas con la estructura factorial deseada y sin valores atípicos.

Finalmente, las simulaciones, la inclusión de los valores extremos y la estimación de los modelos AFC, se realizó con el paquete lavaan, en su versión 0.6-1 [14], disponible en la plataforma R en su versión 3.1.0 [12].

3 Resultados

Primeramente, para decidir la cantidad de replicaciones que se iban a utilizar en las simulaciones, se estudió la estabilidad de la estimación bootstrap del error de Monte Carlo (i.e. la desviación estándar) de los índices de ajuste en algunas condiciones de análisis, según cantidad de réplicas. En un conjunto de N réplicas de una simulación Monte Carlo, se define el error bootstrap de Monte Carlo del estadístico X en k réplicas, con orden b, como

$$MCE_{boot}(X_k, b) = DE\left[X_k\right],$$

donde b es el número de muestras de k réplicas del conjunto inicial, X_k es la variable aleatoria que indica el promedio del estadístico X en cada una de las b muestras y $DE[\cdot]$ representa la desviación estándar [7].

[1]La distancia de Mahalanobis indica la distancia de una observación $\boldsymbol{y_i}$ al centroide de los datos $\bar{\boldsymbol{y}}$, ponderada por la matriz de varianzas observada $\boldsymbol{S}$, se define como $\sqrt{(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'\boldsymbol{S}(\boldsymbol{y_i} - \bar{\boldsymbol{y}})}$.

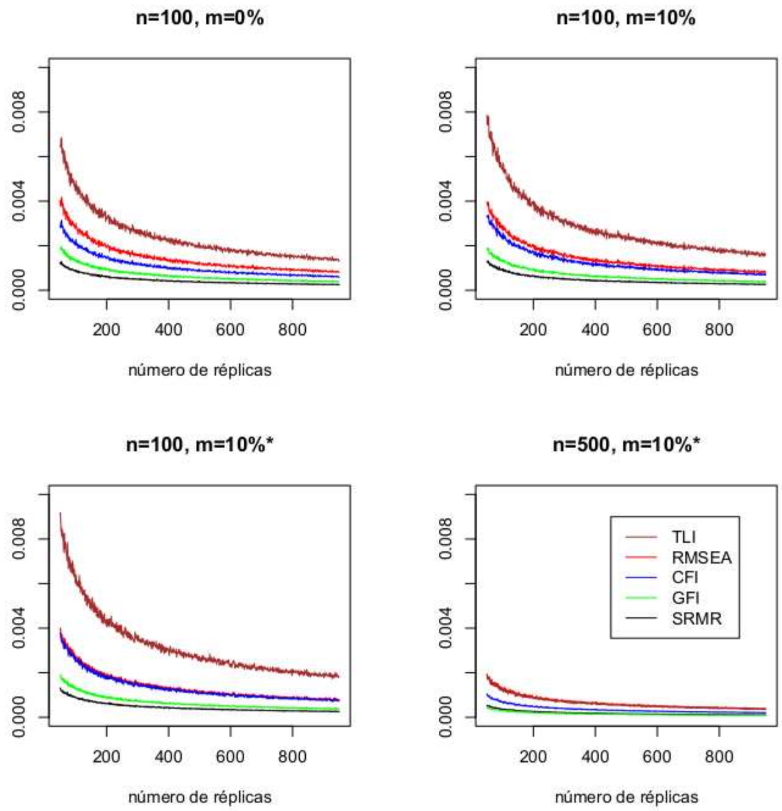

Figura 3: Error bootstrap de Monte Carlo, de orden 500, de los índices de ajuste según número de réplicas, para cada condición analizada.

De los 20 escenarios estudiados (considerando los casos con 0% de valores perdidos), se eligieron cuatro condiciones para analizar el error de Monte Carlo: a) $n = 100, m = 0$, b) $n = 100, m = 10$, c) $n = 100, m = 10*$ y b) $n = 500, m = 10*$. Se consideraron varios casos con tamaño de muestra baja, debido a que se espera que en estos conjuntos se obtenga la mayor variabilidad de los índices. Además, se seleccionaron casos con altos porcentajes de valores extremos, por la misma razón: expectativa de alta variabilidad. El caso con $m = 0$ se utiliza como un caso extremo para los tamaños de muestra con 0% de valores extremos. Finalmente, se seleccionó un caso con $n = 500$, para tener un caso de análisis con tamaño de muestra grande, pero considerando la situación de valores extremos que puede provocar mayor variabilidad.

En la Figura 3 se presenta el error bootstrap de Monte Carlo, de orden 500, de cada índice en k réplicas, para cada condición analizada (las muestras fueron tomadas de una simulación de 10000 réplicas). Se puede observar que en las cuatro condiciones examinadas, todos los índices presentan errores bootstrap de Monte Carlo que decrecen en función del número de réplicas. Además, estos errores presentan un cambio considerable en la tasa de decrecimiento instantánea alrededor de las 200 réplicas, es decir, que la "velocidad" de decrecimiento es menor después de esta cantidad de réplicas. En este punto todos los errores bootstrap de Monte Carlo son menores a 0.005 (lo cual es 0.5% del rango de variabilidad de los índices).

Con base en lo anterior, se concluyó que con 200 o más réplicas se obtenía una estimación adecuada de los índices, ya que con estas cantidades de réplicas se obtenían errores bootstrap de Monte Carlo para los índices que variaban mínimamente con el aumento de la cantidad de réplicas. El número seleccionado de réplicas fue de 500, este valor también permitió mitigar el efecto de la no convergencia del modelo, ya que en la condición c), se observó que un 25% de las estimaciones del modelo especificado de análisis no convergió (en las demás condiciones el porcentaje de no convergencia fue mucho menor).

En segundo lugar, antes de analizar si los índices son robustos a los valores extremos, se estimaron los puntos de corte de los índices de ajuste utilizados según tamaño de muestra. Para establecer los puntos de corte que indican una bondad de ajuste aceptable, se realizó una simulación de 5000 replicaciones con cada tamaño de muestra, generadas a partir del modelo de simulación y luego se les estimó el modelo de análisis. En teoría, el modelo de análisis debe presentar un ajuste ideal en estas replicaciones, por lo cual se establecen los puntos de corte, basándose en el rango donde se ubica el 95% de los valores más apropiados del índice en cuestión (el 95% es utilizado tradicionalmente para no considerar a los valores extremos, presentes en la distribución de un índice). En el caso de los índices en que el buen ajuste se asocia a valores bajos (RMSEA y SRMR), se utilizó como punto de corte al percentil 95, en el caso contrario (CFI, TLI y GFI), se utilizó al percentil 5. Los puntos de corte generados se muestran en la Tabla 1.

Tabla 1: Puntos de corte según índice y tamaño de muestra.

n	SRMR	RMSEA	CFI	TLI	GFI
100	.0662	.0764	.9428	.9141	.9257
200	.0466	.0529	.9719	.9578	.9609
500	.0287	.0320	.9895	.9843	.9842

Tabla 2: Tasas de rechazo del modelo especificado según índice de ajuste y condiciones de los datos.

n	m	Puntos de corte clásicos					Puntos de corte ajustados					No Conv.
		SRMR	RMSEA	CFI	TLI	GFI	SRMR	RMSEA	CFI	TLI	GFI	
100	0%	.00	.25	.07	.17	.49	.05	.04	.04	.04	.05	.00
	1%	.00	.24	.08	.18	.48	.07	.05	.06	.06	.04	.01
	5%	.01	.20	.08	.17	.49	.08	.04	.06	.06	.04	.02
	5%*	.00	.19	.10	.19	.45	.08	.03	.08	.08	.03	.12
	10%	.01	.22	.10	.19	.48	.08	.05	.07	.07	.05	.03
	10%*	.01	.20	.10	.20	.44	.10	.05	.09	.09	.03	.23
200	0%	.00	.05	.00	.03	.00	.03	.04	.04	.04	.03	.00
	1%	.00	.05	.00	.03	.00	.05	.03	.05	.05	.03	.00
	1%*	.00	.04	.01	.03	.00	.06	.03	.05	.05	.03	.00
	5%	.00	.06	.00	.04	.00	.06	.05	.06	.06	.04	.00
	5%*	.00	.05	.01	.06	.00	.09	.04	.08	.08	.04	.02
	10%	.00	.06	.01	.05	.00	.09	.04	.08	.07	.04	.00
	10%*	.00	.05	.01	.06	.00	.09	.03	.07	.07	.03	.09
500	0%	.00	.00	.00	.00	.00	.05	.05	.06	.06	.05	.00
	1%	.00	.01	.00	.00	.00	.09	.07	.09	.09	.06	.00
	1%*	.00	.00	.00	.00	.00	.08	.07	.09	.09	.07	.00
	5%	.00	.00	.00	.00	.00	.12	.08	.10	.10	.07	.00
	5%*	.00	.00	.00	.00	.00	.14	.06	.11	.11	.06	.00
	10%	.00	.00	.00	.00	.00	.12	.08	.10	.10	.07	.00
	10%*	.00	.00	.00	.00	.00	.18	.07	.12	.12	.07	.01

Se tiene n=tamaño de muestra, m=porcentaje de valores perdidos en una variable, *=los valores extremos se reparten en dos variables, No conv.= proporción de modelos no convergentes.
Las tasas de rechazo consideran únicamente los modelos convergentes.

En tercer lugar, se estimó el modelo especificado en cada una de las bases simuladas y se analizaron las tasas de rechazo de ajuste del modelo, según cada índice considerado. Los resultados de este análisis se presentan en la tabla 2. Se consideró que para que un índice sea robusto a la contaminación de los datos debía presentar una tasa de rechazo inferior a 0.10.

Con los puntos de corte clásicos se obtuvo que en los tamaños de muestra de 200 y 500, todos los índices fueron robustos a todas las condiciones de contaminación de los datos propuestas; de hecho, las tasas de rechazo fueron menores o iguales 0.06 en todos los casos y con SRMR, TLI y GFI fue menor o igual que 0.01. Con $n = 100$, únicamente el SRMR y el CFI resultaron robustos. El índice menos robusto a esta violación fue el GFI, luego el RMSEA y el TLI.

Luego, el uso de punto de cortes ajustados permitió obtener tasas de rechazo menores a 0.10 para los índices en prácticamente todas las condiciones. Las únicas tasas de rechazo mayores o iguales a .10 fueron obtenidas en los índices SRMR, TLI y CFI con $n = 500$ y porcentajes de valores extremos mayores o iguales a 5%, pero el valor máximo fue de 0.18, seguido por 0.14.

Por último, también se observó que la presencia de valores extremos en dos variables perjudicaba la convergencia del modelo. Con $n = 100$ se obtuvieron porcentajes de 13% y 23% de estimaciones no convergentes, en el caso con $m = 5*$ y $m = 10*$, respectivamente. Los porcentajes de no convergencia redujeron conforme aumentó el tamaño de muestra, con $n = 200$ se redujo a 2% y 9%, respectivamente.

4 Conclusiones

En primer lugar, se observó que el índice que tuvo menor error bootstrap de Monte Carlo fue el SRMR. Esto coincide con que este fue el único índice utilizado que consideraba la distibución multinormal de los datos. En cambio, el TLI fue el índice que presentó la mayor variabilidad, lo cual coincide con que este el índice que más utiliza el supuesto de la multinormalidad en su fórmula. Lo anterior señala que los índices más estables en las condiciones estudiadas en este artículo son aquellos que no dependen de la multinormalidad. A pesar de lo anterior, la variabilidad de los índices en los casos estudiados fue muy baja y se estabilizo rápidamente con el aumento de réplicas. Esto brindó una primera evidencia de que los índices clásicos de AFC eran apropiados para el análisis de conjuntos de datos con la contaminación de datos estudiada.

La mayor conclusión de este trabajo fue que los índices de ajuste del AFC, considerados, con los puntos de corte clásicos, son robustos al tipo de contaminación de los datos presentado en este estudio, tanto en muestras de 200 como de 500 observaciones. Esta conclusión es opuesta al resultado encontrado por [8], en el cual se observó que el CFI y el RMSEA, bajo el método de máxima verosimilitud, no eran robustos a la contaminación de datos considerada. La razón de esta divergencia se debe a que la contaminación de [8] alcanzaba a todos los indicadores del modelo de un conjunto de observaciones; en cambio, en este trabajo la contaminación solo alcanzaba a una variable por observación (el caso de dos variables se repartía entre observaciones). Lo anterior evidencia que la violación a la multinormalidad presentado en el artículo presente, no era tan drástica como la considerada en el estudio de comparación. Por otro lado, para muestras pequeñas ($n = 100$) solo se recomienda el uso de los puntos de corte clásicos en dos índices considerados: el SRMR y el CFI.

Luego, dado el buen desempeño de los índices de ajuste con los puntos de corte clásicos en muestras de 200 y 500 observaciones, no se considera necesario incentivar el uso de puntos de corte ajustados para obtener robustez a los datos extremos en estos tamaños de muestra. Aunque se pudo observar que en general con los puntos de corte ajustados todos los índices fueron robustos a la contaminación de datos propuesta. En el caso de $n = 100$, es preferible usar los puntos de corte ajustados debido a que proporcionar tasas de rechazo menores (con excepción del SRMR). Los puntos de corte clásicos proporcionaron criterios muy severos de aceptación, que más bien inflaban el error tipo I.

Por último, hay que mencionar que este trabajo consideró una única estructura factorial, sería importante estudiar si en otras estructuras factoriales se dan tasas de rechazo igualmente bajas. Esto permitiría generalizar los resultados a familias de estructuras factoriales. Además, para complementar la pertinencia del uso de los puntos de corte presentados en este estudio se puede realizar un análisis de potencia de los índices, mediante el análisis de tasas de rechazo de modelos mal especificados. En este último caso, es probable que los puntos de corte ajustados sean más apropiados.

Agradecimientos

Se le agradece a la Dra. Eliana Montero-Rojas el apoyo brindado como directora de la tesis de Maestría en Estadística que dio origen al artículo presente.

Fuente de financiamiento

Este artículo fue producto de la tesis de Maestría en Estadística de la Universidad de Costa Rica realizada por el autor, fue parcialmente financiada por él y por dicha Universidad.

Referencias

[1] T.A. Brown, *Confirmatory Factor Analysis for Applied Research*, The Guilford Press, New York, 2006.

[2] M. Cea, *Análisis Multivariable Teoría y Práctica en la Investigación Social*, Editorial Síntesis, Madrid, 2002. Doi: `10.5944/empiria.6.2003.941`

[3] L.T. Hu, P.M. Bentler, *Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification*, Psychological Methods **3**(1998), no. 4, 424–453. Doi: `10.1037/1082-989X.3.4.424`

[4] L.T. Hu, P.M. Bentler, Y. Kano, *Can test statistics in covariance structure analysis be trusted?*, Psychological Bulletin **112**(1992), no. 2, 351–362. Doi: `10.1037/0033-2909.112.2.351`

[5] K.G. Jöreskog, *A general approach to confirmatory maximum likelihood factor analysis*, Psychometrika **34**(1969), no. 2, 183–202. Doi: `10.1007/BF02289343`

[6] D. Kaplan, *Structural Equation Modeling: Foundations and Extensions*, 2nd. Ed., Sage Publications, Thousand Oaks CA, 2009. Doi: `10.4135/9781452226576`

[7] E. Koehler, E. Brown, S.J. Haneuse, *On the assessment of Monte Carlo error in simulation-based statistical analyses*, The American Statistician **63**(2009), no. 2, 155–162. Doi: `10.1198/tast.2009.0030`

[8] M. Lai, J. Zhang, *Evaluating fit indices for multivariate t-based: Structural equation modeling with data contamination*, Frontiers in Psychology **8**(2017), 1286–1298. Doi: `10.3389/fpsyg.2017.01286`

[9] S.A. Mulaik, *Linear Causal Modeling with Structural Equations*, Chapman and Hall/CRC, Boca Raton FL, 2009.

[10] S.A. Mulaik, L.R. James, J. Van Alstine, N. Bennett, S. Lind, C.D. Stilwell, *Evaluation of goodness-of-fit indices for structural equation models*, Psychological Bulletin, **105**(1989), 430–445. Doi: `10.1037/0033-2909.105.3.430`

[11] J. Pek, R.C. MacCallum, *Sensitivity analysis in structural equation models: Cases and their influence*, Multivariate Behavioral Research **46**(2011), no. 2, 202–228. Doi: `10.1080/00273171.2011.561068`

[12] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2007. `https://www.r-project.org/`

[13] L. Rojas-Torres, *Construcción y validación del índice de calidad de la gestión en la educación superior mediante el análisis factorial confirmatorio*, Tesis de Maestría, Universidad de Costa Rica, 2014.

[14] Y. Rossel, *Lavaan: An R package for structural equation modeling*, Journal of Statistical Software **48**(2012), no. 2, 1–35. Doi: `10.18637/jss.v048.i02`

Anexo: Sintaxis de las simulaciones

```
#######################
# Matrices del modelo #
#######################

Lambda=matrix(c(.7,.7,.7,rep(0,9),.7,.7,.7,
rep(0,9),.7,.7,.7),9,3)
Phi=matrix( c(1,0.5,0.5,0.5,1,0.5,0.5,0.5,1),3,3)
Theta=diag(rep(0.51,9))
S= Lambda %*% Phi %*% t(Lambda) + Theta

#Vector de medias
mu<-rep(0,9)

##############################
# Simular bases de 100 datos #
##############################

library(MASS)
BASES100 <- list()
for(i in 1:500){
BASES100[[i]]<-mvrnorm(100,mu,S);
colnames(BASES100[[i]])<-c("x1","x2","x3","x4","x5",
"x6","x7","x8","x9")}

# BASESn.m.k; n indica tama\~no de muestra,
# m indica porcentaje de valores extremos
# (no se indica el 0)
# k indica en cu\'antas variables se reparten
# los valores extremos(no se indica el 1)

BASES100.1=BASES100.5=BASES100.5.2<-BASES100
BASES100.10=BASES100.10.2<-BASES100
for(i in 1:500){
BASES100.1[[i]][1,1]<-runif(1,10,11);
BASES100.5[[i]][1:5,1]<-runif(5,10,11);
BASES100.5.2[[i]][1:2,1]<-runif(2,10,11);
BASES100.5.2[[i]][3:5,2]<-runif(3,10,11);
```

```
BASES100.10[[i]][1:10,1]<-runif(10,10,11);
BASES100.10.2[[i]][1:5,1]<-runif(5,10,11);
BASES100.10.2[[i]][6:10,2]<-runif(5,10,11)
}

##############################
# Simular bases de 200 datos #
##############################

BASES200 <- list()
for(i in 1:500){
BASES200[[i]]<-mvrnorm(200,mu,S);
colnames(BASES200[[i]])<-c("x1","x2","x3","x4","x5",
"x6","x7","x8","x9")}

BASES200.1=BASES200.1.2=BASES200.5=BASES200.5.2
<-BASES200
BASES200.10=BASES200.10.2<-BASES200
for(i in 1:500){
BASES200.1[[i]][1:2,1]<-runif(2,10,11);
BASES200.1.2[[i]][1,1]<-runif(1,10,11);
BASES200.1.2[[i]][2,2]<-runif(1,10,11);
BASES200.5[[i]][1:10,1]<-runif(10,10,11);
BASES200.5.2[[i]][1:5,1]<-runif(5,10,11);
BASES200.5.2[[i]][6:10,2]<-runif(5,10,11);
BASES200.10[[i]][1:20,1]<-runif(20,10,11);
BASES200.10.2[[i]][1:10,1]<-runif(10,10,11);
BASES200.10.2[[i]][11:20,2]<-runif(10,10,11)
}


##############################
# Simular bases de 500 datos #
##############################

BASES500 <- list()
for(i in 1:500){
BASES500[[i]]<-mvrnorm(500,mu,S);
colnames(BASES500[[i]])<-c("x1","x2","x3","x4","x5",
```

```
"x6","x7","x8","x9")}

BASES500.1=BASES500.1.2=BASES500.5=BASES500.5.2
<-BASES500
BASES500.10=BASES500.10.2<-BASES500
for(i in 1:500){
BASES500.1[[i]][1:5,1]<-runif(5,10,11);
BASES500.1.2[[i]][1:2,1]<-runif(2,10,11);
BASES500.1.2[[i]][3:5,2]<-runif(3,10,11);
BASES500.5[[i]][1:25,1]<-runif(25,10,11);
BASES500.5.2[[i]][1:13,1]<-runif(13,10,11);
BASES500.5.2[[i]][14:25,2]<-runif(12,10,11);
BASES500.10[[i]][1:50,1]<-runif(50,10,11);
BASES500.10.2[[i]][1:25,1]<-runif(25,10,11);
BASES500.10.2[[i]][26:50,2]<-runif(25,10,11)
}


########################################
# Estimar el modelo especificado en la #
# simulaci\'on de una condici\'on ######
########################################

library(simsem)
library(lavaan)

modAn<- "F1=~x1+x2+x3;F2=~x4+x5+x6;F3=~x7+x8+x9"

Out100<- sim(modAn, std.lv=TRUE, std.ov=TRUE, "cfa",
BASES100, "ml")


###############################################
# Bases de datos con los \'indices de ajuste  #
###############################################

Ind100<-
data.frame(attributes(Out100)$fit[,c("srmr","rmsea",
"cfi","tli","gfi")],
```

```
attributes(Out100)$converged)

###########################
### Tasas de rechazo      #
###########################


X<-subset(Ind100,Ind100[,6]==0)
#base con estimaciones convergentes


c(1-colSums(data.frame(X[,1]<.08,
X[,2]<.05,X[,3]>.95,X[,4]>.95,
X[,5]>.95))/dim(X)[1],
#puntos de corte cl\'asicos
1-colSums(data.frame(X[,1]<.0662,
X[,2]<.0764,X[,3]>.9428,
X[,4]>.9141,X[,5]>.9257))/dim(X)[1],
#puntos de corte ajustados
(500-dim(X)[1])/500 )
# modelos convergentes
```