Evaluation of the Effect of Outliers on the GFI Quality Adjustment Index in Structural Equation Models and Proposal of Alternative Quality Indices

Lucia Pereira Barroso[1], and Marcelo Angelo Cirillo[2]
[1]University of São Paulo, São Paulo, SP, Brazil
[2]Federal University of Lavras, Lavras, MG, Brazil
Corresponding author: Lucia Pereira Barroso, email: lbarroso@ime.usp.br

Abstract

When considering a sample characterized by the presence of outliers it is reasonable to assume that the results of the indices of quality of fit of a structural equation model may be affected. The main consequence of the occurrence of this is that the researcher may be influenced to select an inappropriate model. This work is intended to suggest modifications in the construction of the GFI index using robust methods for estimating the unrestricted sample covariance matrix, leading to new indices called $GFI_{(MCD)}$ *and* $GFI_{(MVE)}$*. The validation of this proposal was made using Monte Carlo simulation methods, considering different sample sizes, differences between the unrestricted sample covariance matrix and those imposed by the structural model, and different numbers of outliers generated by distributions with deviations from symmetry and excess kurtosis. It was concluded that for larger samples size (*$n \geq 100$*), given that the outliers are from distributions that are symmetrical, the* $GFI_{(MCD)}$ *and* $GFI_{(MVE)}$ *indices showed similar results, including samples with elevated levels of outliers.*

Keywords: structural equations, outliers, robust estimation, MCD, MVE.

1. Introduction

The goodness of fit of a Structural Equation Model (SEM) is evaluated by various indices. Considering the fact that certain methods of estimation are sensitive to outliers and, that in violation of the assumption of multivariate normality, it is reasonable to assume that the results provided by these indices, regarding the goodness of fit of the model, are suspected of being incorrect (Kirby and Bollen, 2009). For these authors the major problem with these indices is their high degree of sensitivity to sample size. For large samples, these rates generally tend to be significant, so as to favor the rejection of the model. If the sample size is too small, the results of these indices will show a tendency to be not significant. Thus, there is supporting evidence that the researcher may be influenced not to reject one model as compared to other competing models that would provide better fit.

Of course, other factors will be able to contribute such that the results provided by these indices may not be real. Kaplan (2000) mentions that the specification errors are usually characterized by the omission of an indicator variable that must necessarily be inserted into the model. Another situation of the occurrence of these errors is given by the indication of the uni or bi-directional effects of one variable over another being specified incorrectly.

Given the problems mentioned above, the results provided by the goodness-of-fit indices are worth questioning. This suggests that alternative methods, with the use of computing resources, should be applied. Satorra (2000) found that in certain situations the chi-square test is asymptotically robust in violation of the normality assumption.

In dealing with the presence of outliers, the application of re-sampling methods, in conjunction with robust estimation techniques, are suitable for use not only in assessing the behavior of the estimators when certain assumptions inherent in the estimation methods are violated, but also in the construction of new information measures relating to the quality of adjustment of an SEM. In this context, the most used of the robust methods of estimation based on re-sampling of the observations are the Minimum Covariance Determinant Estimator (MCD) (Rousseeuw and Driessen, 1999) and the Minimum Volume Ellipsoid (MVE) (Jackson and Chen, 2004).

In both methods, basically the process of re-sampling is performed on subsets of smaller size in relation to the sample size, but related to the assumed break point, interpreted as the highest percentage of contamination that an estimator would be able to support while providing an accurate estimate.

Thus, among all subsets studied, a subset is selected following the criteria established for each method. In the case of the MCD method the selected subset that will be used in the estimation of a covariance matrix is the one presenting the smallest determinant, when compared with the other matrices obtained from the other subsets. Similarly, the subset selected, using the MVE method, will result in a covariance matrix generated by an ellipsoid of minimum volume presenting a coverage of at least (N/2+1) sample points, N being defined as the sample size. Due to this fact, this estimator is known by producing a maximum break point of 0.5. Because of the above, the application of these methods in the SEM is justified by the fact that the difference between the unrestricted sample covariances and those predicted by the model should be minimized.

Given what has been mentioned above, the purpose of this study is to assess the effect of outlier observations on the GFI as well as suggesting new modifications in the construction of this GFI index, incorporating robust estimates of the sample covariance matrix.

2. Methodology

In line with the goal proposed for Monte Carlo evaluation of the influence of outlier observations on the results of the GFI in its original form, and suggested modifications utilizing the structural equation model, illustrated as follows.

In matrix form, the theoretical model was represented by equations (1) – (3).

$$\boldsymbol{\eta} = \mathrm{B}\boldsymbol{\eta} + \Gamma\xi + \zeta \tag{1}$$

$$\mathbf{Y} = \Lambda_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

$$\mathbf{X} = \Lambda_x \boldsymbol{\xi} + \boldsymbol{\delta}. \tag{3}$$

In (1) the terms were defined as follows: $\boldsymbol{\eta}_{\mathrm{m}\times 1}$ represents the vector of endogenous latent variables; $\xi_{\mathrm{r}\times 1}$ the vector of latent exogenous variables; $\zeta_{\mathrm{m}\times 1}$ the vector of structural errors; $\mathrm{B}_{\mathrm{m}\times\mathrm{m}}$ the matrix of coefficients relating the endogenous latent variables and finally, $\Gamma_{\mathrm{m}\times\mathrm{r}}$ the matrix of coefficients that relates the latent exogenous variables to the latent endogenous variables.

Specifically the measurement model **Y**, conforms to equation (2), the relationship of the latent endogenous variables with observed variables y_j (j=1,,...,p) was determined by the matrix of coefficients, $\Lambda_{y(\mathrm{p}\times\mathrm{m})}$, given the measurement error represented by $\boldsymbol{\varepsilon}_{(\mathrm{p}\times 1)}$.

Equation (3) refers to the X measurement model used to relate the latent exogenous variable with the observed variables x_k (k=1,..,q). The relationship between the latent exogenous variable and the observed variables was specified by the matrix of coefficients $\Lambda_{x(\mathrm{q}\times\mathrm{r})}$ and finally the measurement error represented by the vector $\boldsymbol{\delta}_{(\mathrm{q}\times 1)}$. Keeping the assumptions of the structural model in which the expectations of the error vectors and the latent variables are equal to zero, $\zeta_{\mathrm{m}\times 1}$ and $\xi_{\mathrm{r}\times 1}$ are not correlated, $\boldsymbol{\varepsilon}_{(p\times1)}$ is not correlated with $\boldsymbol{\eta}_{m\times1}$, $\boldsymbol{\xi}_{r\times1}$ and $\boldsymbol{\delta}_{(q\times1)}$, are not correlated with, $\boldsymbol{\xi}_{r\times1}$, $\boldsymbol{\eta}_{m\times1}$ and $\boldsymbol{\varepsilon}_{(p\times1)}$. The structural equation model was simulated with respect to the relationship $\mathbf{v} = A\mathbf{v} + \mathbf{u}$, in matrix form in (4)

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ x_1 \\ x_2 \\ \eta_1 \\ \eta_2 \\ \xi_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.3 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ x_1 \\ x_2 \\ \eta_1 \\ \eta_2 \\ \xi_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \delta_1 \\ \delta_2 \\ \zeta_1 \\ \zeta_2 \\ \xi_1 \end{bmatrix}, \qquad (4)$$

where **v** corresponded to the vector formed by the latent and observed variables. The average error vectors formed by the structural errors provided by equation (1) and the measurement errors provided by the sub-models (2) and (3) formed the vector **u**.

Assuming a as the number of variables defined in **v** and k=p+q the number of observed variables, for which the covariance matrix imposed by the structural model was completely specified, the following matrices were defined:

$$\mathbf{P}_{(a\times a)} = \begin{bmatrix} 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -0.7 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.7 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -0.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} ; \mathbf{J}_{(k\times a)} = \left[\mathrm{I}_k \vdots 0 \right], \qquad (5)$$

such that **P** represents the covariance matrix of the elements of vector **u** (4) assuming the parametric values used in the Monte Carlo simulation, **J** is a matrix whose first k columns form the identity matrix I_k. With these specifications, the parameter relative to the covariance matrix imposed by the structural model was built using equation (6)

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \mathbf{J}\left(\mathbf{I}_a - \mathbf{A}\right)^{-1} \mathbf{P}\left[\left(\mathbf{I}_a - \mathbf{A}\right)^{-1}\right]^{\mathbf{t}} \mathbf{J}^{\mathrm{t}} . \qquad (6)$$

After obtaining the covariance matrix (6), the multivariate samples were generated with respect to the structural model, in such a way that the data matrix was generated according to, $\mathrm{G} \sim \mathrm{N}_{(k)}\left(\underset{\sim}{0}, \Sigma_\theta\right)$. In this way, $\hat{\Sigma}_\theta$ was obtained for each Monte Carlo simulation performed. In obtaining the estimates of the sample covariance matrix, the parametric value used in the generation of the samples was specified in $\Sigma = \sqrt[k]{\mu}\,\Sigma_\theta$. for μ=5.

Proposed modification of the Goodness of Fit Index (GFI)

The proposed change in the GFI, while keeping the structural model is derived from the residual matrices (10) estimated in conventional form and considering the robust covariance matrix estimates, obtained by the MCD methods and with the same specifications as the MVE method assuming a breaking point of 50%. The notation used for the representation of these matrices is defined in (10).

$$D_{(MCD)} = S_{(MCD)} - \hat{\Sigma}_{\theta} \;;\; D_{(MVE)} = S_{(MVE)} - \hat{\Sigma}_{\theta} \text{ and } D_{(S)} = S - \hat{\Sigma}_{\theta}, \tag{10}$$

where $\hat{\Sigma}_{\theta}$ is the estimator of the structural covariance matrix; S_{MCD} and S_{MVE}, the estimators of the robust sample covariance matrices, obtained by the MCD and MVE methods respectively; and finally S is the conventional sample covariance matrix. Based on these matrices, the suggested modification in the GFI adjustment is given by the use of the residual matrices defined in (11)

$$GFI_{(MCD)} = 1 - \frac{tr\left[\left(D^2_{(MCD)}\right)\right]}{tr\left(S^2_{(MCD)}\right)} \;;\; GFI_{(MVE)} = 1 - \frac{tr\left[\left(D^2_{(MVE)}\right)\right]}{tr\left(S^2_{(MVE)}\right)}, \tag{11}$$

where $tr(M)$ denotes the trace of the matrix M.

Then, the modification suggested in (11) could be evaluated against the original form of the GFI, in order to verify the occurrence of an under or over estimation of the value related to the actual quality of the fit, respectively, the percentages of variation were calculated as in expression (12)

$$P_{(MCD)} = \left(\frac{GFI_{(MCD)} - GFI_{(S)}}{GFI_{(S)}}\right) \times 100 \text{ and } P_{(MVE)} = \left(\frac{GFI_{(MVE)} - GFI_{(S)}}{GFI_{(S)}}\right) \times 100, \tag{12}$$

where $GFI_{(S)}$ is the index of quality of conventional tuning.

To validate the proposed methodology, a program was developed in software version 2.11.1 R. For each case 1000 samples were generated, and in each sample 10,000 re-samplings were considered to obtain the robust estimates of the sample covariance matrix.

3. Results and Discussion

Preliminary to the discussion of the results regarding the modified indices of goodness of fit, it should be noted that in simulation studies Anderson and Gerbing (1984) observed that when adjusting a structural equation model, using the maximum likelihood method, the mean of the GFI index in its original form was increased due to increased sample size. Given the occurrence of these results it is reasonable to assume that in addition to the sample size, the number of outliers in the sample leads one to suspect the results provided by the various measures of adjustment. The discussion of the results described in Table 1 is given considering only the extreme situations evaluated in the simulation, represented by the lowest ($\alpha = 5\%$) and highest ($\alpha = 30\%$) proportion of outliers for all the sample sizes and multivariate distributions.

Considering the smallest proportion of outliers contained in the sample (α=0.05) the values of the GFI indices, in the conventional form, for all sample sizes, have remained close to the unity, except in mixtures with the log-normal distribution. This fact was apparent for the distributions in which the outliers were generated by symmetric distributions (t-Student and Uniform). In dealing with the percentage of the variation, in the proposal to compare the results of the quality of fit of the model, as measured by conventional and modified GFI indices, it was observed that in situations of smaller samples ($n \leq 50$) the use of the $GFI_{(MCD)}$ index provided results with more discrepancies compared to the results provided by the $GFI_{(MVE)}$ index. Thus, due to the low proportion of outliers, there is evidence to state that the GFI$_{(MCD)}$ index is more sensitive to heterogeneity between the unrestricted sample covariance matrix and the covariance matrix estimated by the hypothetical model. For large sample sizes, this effect was corrected.

Table 1 - Median value and the percentage of variation P$_{(MCD)}$ and P$_{(MVE)}$ as a function of the probabilities of the mixture (α=0.05 and 0.30.

			n=25		n=50	
Distribution	α	GFI	P$_{(MCD)}$ (%)	P$_{(MVE)}$ (%)	P$_{(MCD)}$ (%)	P$_{(MVE)}$ (%)
t-Student	0.05	0.922	-29.8	3.8	-9.2	3.7
	0.30	0.893	-25.1	6.9	-8.1	7.6
Uniform	0.05	0.956	-27.4	2.2	-7.1	2.5
	0.30	0.981	-5.8	-26.9	-10.2	-33.8
Log-normal	0.05	0.871	-24.7	7.8	-5.1	11.6
	0.30	0.291	89.9	188.8	149.9	200.4
			n=100		n=200	
Distribution	α	GFI	P$_{(MCD)}$ (%)	P$_{(MVE)}$ (%)	P$_{(MCD)}$ (%)	P$_{(MVE)}$ (%)
t-Student	0.05	0.934	-2.7	3.0	-0.5	3.6
	0.30	0.841	3.8	12.7	5.3	12.1
Uniform	0.05	0.969	-2.4	2.2	-0.9	2.0
	0.30	0.996	-18.5	-21.1	-19.7	-14.9
Log-normal	0.05	0.741	13.2	21.7	26.2	31.3
	0.30	0.195	145.6	177.4	307.3	385.1

Considering the smallest proportion of outliers contained in the sample (α=0.05) the values of the GFI indices, in the conventional form, for all sample sizes, have remained close to the unity, except in mixtures with the log-normal distribution. This fact was apparent for the distributions in which the outliers were generated by symmetric distributions (t-Student and Uniform). In dealing with the percentage of the variation, in the proposal to compare the results of the quality of fit of the model, as measured by conventional and modified GFI indices, it was observed that in situations of smaller samples ($n \leq 50$) the use of the GFI$_{(MCD)}$ index provided results with more discrepancies compared to the results provided by the GFI$_{(MVE)}$ index. Thus, due to the low proportion of outliers, there is evidence to state that the GFI$_{(MCD)}$ index is more sensitive to heterogeneity between the unrestricted sample covariances matrix and the covariance matrix estimated by the hypothetical model. For large sample sizes, this effect was corrected.

Increasing the proportion of outliers (α = 0.30) in relation to the symmetric distributions it was observed that the GFI remained close to the unity, this being due to the high number of outliers in the sample. This is in accordance with Yuan and Bentler (2001), who asserted that in non-normal, multivariate samples the estimates of the covariance matrices are distorted. Hoyle and Panter (1995) and Shah and Goldstein (2006) showed that a sample presenting multivariate normality for $n \geq 650$ does not show effects on the normalized goodness-of-fit indices and comparative indices. In the situations simulated in this work, the largest sample size considered was n = 200. Therefore, due to the arguments imposed by these authors, the hypothesis should be suspicious of the values of quality of fit, provided by these modified GFI indices. In this context, comparing the effect of sample size, it was observed for smaller samples ($n \leq 50$) that there were discrepancies between the GFI$_{(MCD)}$ and GFI$_{(MVE)}$ results, with a tendency to underestimate the quality of fit provided by the conventional GFI, while, in large samples ($n \geq 100$) the results for these indices were more similar.

Faced with this situation where the modified indices present robust results in the presence of outliers, it is again recommended that the FAST-MCD algorithm be used, increasing the number of re-samplings, resulting in a robust new estimate for the sample covariance matrix. In situations where the results of the modified indices were not similar to the results of the conventional index, there is evidence to state that for n≥100, the proposed changes are worth being used by the researcher, according to the simulated conditions.

In dealing with the log-normal distribution in the case of low numbers of outliers (α = 0.05) the use of modified goodness-of-fit indices are plausible for sample sizes ≥ 100.

However, it should be emphasized that due to the results expressed by the percentages of variation, it was observed that with increasing sample size there is evidence that the values may be inflated in relation to goodness of fit provided by the conventional GFI index. As a result of the above, it is recommended that the researcher seek agreement between the results provided by other comparative indices, such as support for decision making, in selecting the specific model.

For larger samples considering the log-normal distribution, given a significant number of outliers (α = 0.30) the use of the modified goodness-of-fit indices is not recommended due to the elevated percent values, indicating a significant inflation of the estimates in relation to the quality of fit provided by the conventional GFI. This can be corroborated by the excess kurtosis caused to a great degree, by the high number of outliers added to the sample.

4. Acknowledgements

The authors would like to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for financial support (grant 2013/08738-9).

References

Anderson, J., Gerbing, D.W. (1984). The effects of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis, Psychometrics, 49, 155-173.

Hoyle, R.H., Panter, A.T. (1995). Writing about structural equation models, in: R. H. Hoyle (ed.), Structural Equation Modeling: Concepts, Issues and Applications, Sage Publications, Calif., 159–176.

Jackson, D.A., Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data, Environmetrics, 15, 129–139.

Kaplan, D. (2000). Structural Equation Modeling: Foundations and Extensions, second ed., Sage: Thousand Oaks, Wisconsin.

Kirby, J., Bollen, K.A. (2009). Using instrumental variable (IV) tests to evaluate model specification in latent variable structural equation models, Sociological Methodology, 39, 327-355.

R Development Core Team (2010). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Austria.

Rousseeuw, P.J., Driessen, K. (1999). Fast algorithm for the minimum covariance determinant estimator, Technometrics, 41, 212–223.

Satorra, A. (2000). Robustness issues in structural equation modeling: a review of recent developments. Quality & Quantity, 24, 367-386.

Shah, R., Goldstein, S.M. (2006). Use of structural equation modeling in operations management research: looking back and forward, Journal of Operations Management, 24, 148–169.

Yuan, K.H., Bentler, P.M. (2001). Effects of “outliers” on estimators and tests in covariance structure analysis, British Journal of Mathematical and Statistical Psychology, 54, 161–175.