A global Goodness–of–Fit index for PLS structural equation modelling[1]

Un indice di validazione globale per i modelli ad equazioni strutturali con il metodo PLS

Michel Tenenhaus
Department SIAD
HEC School of Management
tenenhaus@hec.fr

Silvano Amato[2], Vincenzo Esposito Vinzi
Dipartimento di Matematica e Statistica
Università Federico II — Napoli
(silamato, vincenzo.espositovinzi)@unina.it

Riassunto: Scopo di questo lavoro è proporre un indice globale di bontà dell'adattamento per i modelli ad equazioni strutturali stimati mediante il metodo PLS. Al momento gli indici comunemente utilizzati in questo ambito sono la comunalità (bontà di adattamento del modello esterno) e la ridondanaza (qualità del modello interno). Si propone un indice che, alla stregua del test del chi-quadro e dei test ad esso collegati disponibili nel modello classico LISREL stimato mediante la massima verosimiglianza, fornisca una misura globale della bontà di adattamento del modello che tenga conto di entrambi gli aspetti, misurati invece separatamente dalla comunalità e dalla ridondanza. E' inoltre suggerita una procedura di validazione non parametrica per l'indice proposto. Un'applicazione su dati reali è infine presentata.

Keywords: PLS path modeling, Goodness–of–Fit index, bootstrap.

1. Introduction

Here we suggest a global goodness–of–fit index for a structural equation model estimated by PLS, Chatelin *et al.* (2002). Commonly used indexes, communality or redundancy, refer separately to the reconstruction of the measurement model and the structural model. Here we suggest an index that, similarly to the χ^2–based indexes applied in LISREL, yields a measure of the global goodness–of–fit as it is a compromise between communality and redundancy.

2. The proposed index

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J$ be J blocks of manifest variables each made by p_j, $j = 1, 2, \ldots, J$ variables. Within PLS approach we search for J latent variables $\mathbf{y}_j = \mathbf{X}_j \mathbf{a}_j$ so as to be as highly correlated as possible with their own blocks and, for the endogenous ones (J' as a whole), also with their adjacent latent variables. We can define the following index, $\tilde{GF}^2$, for evaluating the global goodness–of–fit of the model:

$$\tilde{GF}^2 = \left[\frac{1}{J}\sum_j \left(\frac{1}{p_j}\sum_i^{p_j} r^2\left(\mathbf{x}_{ij}, \mathbf{y}_j\right)\right)\right] \times \left[\frac{1}{J'}\sum_{j:\mathbf{y}_j endogenous} R^2\left(\mathbf{y}_j; \mathbf{y}_{j1}, \ldots, \mathbf{y}_{jk_j}\right)\right] \quad (1)$$

[1]This paper is financially supported by ESIS (European Satisfaction Index System) European IST Project 2000-31071.

[2]Indirizzo per corrispondenza: silamato@unina.it

where $r^2(\mathbf{x}_{ij}, \mathbf{y}_j)$ and $R^2(\mathbf{y}_j; \mathbf{y}_{j1}, \ldots, \mathbf{y}_{jk_j})$ are squared correlation coefficients.

This index recalls optimization criterion for PLS regression (Tenenhaus (1998)): given two blocks, $\mathbf{X}_j$ and $\mathbf{X}_l$, Tucker criterion searches for two vectors $\mathbf{a}_j$ and $\mathbf{a}_l$ such that:

$$\begin{cases} \max\limits_{\mathbf{a}_j, \mathbf{a}_l} \{\mathbf{a}_j' \mathbf{X}_j' \mathbf{X}_l \mathbf{X}_l' \mathbf{X}_j \mathbf{a}_l\} \\ \mathbf{a}_j' \mathbf{a}_j = 1 \\ \mathbf{a}_l' \mathbf{a}_l = 1 \end{cases} \tag{2}$$

The maximization of (2) is equivalent to the maximization of the squared covariance between $\mathbf{X}_j \mathbf{a}_j$ and $\mathbf{X}_l \mathbf{a}_l$: $cov^2(\mathbf{X}_j \mathbf{a}_j, \mathbf{X}_l \mathbf{a}_l) = r^2(\mathbf{X}_j \mathbf{a}_j, \mathbf{X}_l \mathbf{a}_l) var(\mathbf{X}_j \mathbf{a}_j) var(\mathbf{X}_l \mathbf{a}_l)$. Similarly to the criterion by Tucker —that searches for a compromise between the optimal approximation of each variable block, separately considered, and the optimal reconstruction of the relationship between the two blocks— the $\tilde{GF}^2$ gives a goodness-of-fit index which integrates the two aspects in the PLS approach to structural equation modeling. In the following we motivate this statement. First of all we note that a normalization of $\tilde{GF}^2$ is possible by relating each term in (1) to the corresponding maximum value. We now consider the maximation of the terms separately.

3. Maximization of the two terms in $\tilde{GF}^2$

For clarity sake, let us focus on one of the J manifest variable blocks, say $\mathbf{X}_j$. As well known in principal component analysis, the best rank 1 approximation of $\mathbf{X}_j$ is given by the eigenvector $\mathbf{u}_{j1}$ associated to the largest eigenvalue λ_{j1} of $\mathbf{X}_j' \mathbf{X}_j$; $\mathbf{u}_{j1}$ and $\mathbf{X}_j \mathbf{u}_{j1}$ are respectively the first factor and the first principal component of $\mathbf{X}_j$. Furthermore, $\forall j = 1, 2, \ldots, J$ and $\forall i = 1, 2, \ldots, p_j$: $\sum_i^{p_j} r^2 (\mathbf{x}_{ji} \mathbf{y}_j)$ is a maximum iff $\mathbf{y}_j = \mathbf{X}_j \mathbf{u}_{j1}$.

Then, —if the data are mean centered and with unit variance— the first term in (1) is such that $\sum_i^{p_j} r^2 (\mathbf{x}_{ji}, \mathbf{y}_j) \leq \lambda_{j1}$. If in this expression, the equality sign holds for all j, each component $\mathbf{X}_j \mathbf{a}_j$ is maximally correlated with its own block. Then, two alternative normalized versions of the first term of $\tilde{GF}^2$ are:

$$T_{11}^2 = \left[\frac{1}{J} \sum_j \frac{\sum_i^{p_j} r^2 (\mathbf{x}_{ij}, \mathbf{y}_j)}{\lambda_j} \right] \tag{3}$$

$$T_{12}^2 = \left[\frac{1}{J} \sum_j \frac{\sum_i^{p_j} r^2 (\mathbf{x}_{ij}, \mathbf{y}_j)}{\sum_j \frac{\lambda_j}{p_j}} \right] \tag{4}$$

Let us now consider a partition of the manifest variable set which consists of block $\mathbf{X}_j$, on one side, and the super block made by the manifest variables whose latent variables are explanatory of $\mathbf{y}_j$, say $\mathbf{X}_{\bar{j}}$. We search for two unit vectors $\mathbf{a}_j$ and $\mathbf{a}_{\bar{j}}$ such that the squared correlation coefficient between $\mathbf{X}_j \mathbf{a}_j$ and $\mathbf{X}_{\bar{j}} \mathbf{a}_{\bar{j}}$ is a maximum.

As well known in canonical correlation analysis, $\mathbf{a}_j$ and $\mathbf{a}_{\bar{j}}$ are the eigenvectors respectively of $\mathbf{X}_j' \mathbf{X}_{\bar{j}} (\mathbf{X}_{\bar{j}}' \mathbf{X}_{\bar{j}})^{-1} \mathbf{X}_{\bar{j}}' \mathbf{X}_j$ and $(\mathbf{X}_{\bar{j}}' \mathbf{X}_{\bar{j}})^{-1} \mathbf{X}_{\bar{j}}' \mathbf{X}_j (\mathbf{X}_j' \mathbf{X}_j)^{-1} \mathbf{X}_j' \mathbf{X}_{\bar{j}}$ corresponding to the largest common eigenvalue $\rho_{j\bar{j}}^2$: the squared correlation coefficient between $\mathbf{X}_j \mathbf{a}_j$ and $\mathbf{X}_{\bar{j}} \mathbf{a}_{\bar{j}}$.

In the special case where $p_j = 1$, that is $\mathbf{X}_j$ reduces to a column vector, $\mathbf{x}_j$, we have $\mathbf{a}_{\bar{j}} = c \mathbf{x}_j \mathbf{X}_{\bar{j}} (\mathbf{X}_{\bar{j}}' \mathbf{X}_{\bar{j}})^{-1} \mathbf{X}_{\bar{j}}' \mathbf{x}_j$ and $\rho_{j\bar{j}}^2 = \mathbf{x}_j \mathbf{X}_{\bar{j}} (\mathbf{X}_{\bar{j}}' \mathbf{X}_{\bar{j}})^{-1} \mathbf{X}_{\bar{j}}' \mathbf{x}_j (\mathbf{x}_j' \mathbf{x}_j)^{-1}$, where c is a proper

constant. Then $\rho^2_{j\bar{j}}$ is the multiple correlation coefficient between the variable $\mathbf{x}_j$ and the manifest variables of the $\bar{j}^{th}$ (exogenous) block.

In a more general situation $R^2\left(\mathbf{y}_j;\mathbf{y}_{j1},\mathbf{y}_{j2},\ldots,\mathbf{y}_{jk_j}\right) \leq \rho_j^2$, where ρ_j is the first canonical correlation of the canonical analysis of matrix $\mathbf{X}_j$ and $\left[\mathbf{X}_{j1},\mathbf{X}_{j2},\ldots,\mathbf{X}_{jk_j}\right]$. When in the above expression the equality sign holds for all j, endogenous latent variables $\mathbf{y}_j$ are maximally correlated to their adjacent latent variables. Then, two alternative normalized versions of the second term of $\tilde{GF}^2$ are:

$$T_{21}^2 = \left[\frac{1}{J'}\sum_{j:\mathbf{y}_j endogenous}\frac{R^2\left(\mathbf{y}_j;\mathbf{y}_{j1},\ldots,\mathbf{y}_{jk_j}\right)}{\rho_j^2}\right] \tag{5}$$

$$T_{22}^2 = \left[\frac{1}{J'}\sum_{j:\mathbf{y}_j endogenous}\frac{R^2\left(\mathbf{y}_j;\mathbf{y}_{j1},\ldots,\mathbf{y}_{jk_j}\right)}{\sum_{j:\mathbf{y}_j endogenous}\rho_j^2}\right] \tag{6}$$

By combining equations (3) and (5) or equations (4) and (6) we obtain two different normalized goodness-of-fit indexes yielding the same result and ranging between 0 and 1: $GF^2 = T_{11}^2 T_{21}^2 = T_{12}^2 T_{22}^2$. The geometric mean of T_{11} and T_{21}, named GF (i.e. Goodness-of-Fit index), is used because it can be interpreted similarly to the R^2 index in a regression framework. In the following, we refer to $\hat{GF}$ and to $\hat{GF}^2$ as the sample estimates of, respectively, GF and GF^2.

4. A non–parametric procedure for validating the index

If we are interested in the interval estimate of GF, or GF^2, we can apply a non–parametric procedure. Let $\hat{F}_{\mathbf{X}}$ be the empirical cumulative distribution function (cdf) of $\mathbf{X} = [\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_J]$. In the following we directly consider GF but all statements can be well referred to GF^2.

1. Draw B random samples from $\hat{F}_{\mathbf{X}}$ and for $b = 1,2,\ldots,B$ compute $\hat{GF}^{(b)}$, i.e. for each bootstrap sample, perform a PLS estimation of the same structural model that yielded $\hat{GF}$ and compute the corresponding index.
2. The cdf of the Monte Carlo approximation $\Phi_{\hat{GF}}^{(B)}$ of the bootstrap distribution of $\hat{GF}$ is yielded by the bootstrap estimates $\hat{GF}^{(b)}$. The distribution is estimated under the hypothesis $GF = \hat{GF}$ and approximated by means of B bootstrap samples.

A confidence interval with nominal confidence level $1-2\alpha$ is $[\Phi_{\hat{GF}}^{(B)}(\alpha), \Phi_{\hat{GF}}^{(B)}(1-\alpha)]$ i.e. the 100α% and the $100(1-\alpha)$% percentiles of $\Phi_{\hat{GF}}^{(B)}$. In order to heuristically verify whether the observed $\hat{GF}$ is significantly greater than 0, we can check if the above interval does not include 0.

In order to perform a non–parametric hypothesis testing on GF we need to set hypotheses on one or more coefficients of the structural model:

$$\mathrm{H}_0 : \beta_{ij} = 0 \quad vs. \quad \mathrm{H}_0 : \beta_{ij} \neq 0 \tag{7}$$

where β_{ij} is the PLS coefficient in the structural equation linking $\mathbf{y}_i$ to $\mathbf{y}_j$.

In order to perform a non–parametric hypothesis testing we need to properly transform $\mathbf{X}$ and define an empirical cdf $\hat{F}_{\mathbf{X}^*}$ such that the null hypothesis H_0 holds. Let $\mathbf{X}_j^* = \mathbf{X}_j - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{X}_j$ be the part of $\mathbf{X}_j$ not due to $\mathbf{X}_i$; finally $\mathbf{X}^* = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_j^*, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_J]$.

Similarly to the cdf $\Phi_{\hat{GF}}^{(B)}$, a Monte Carlo, $\Phi_{\mathrm{H}_0}^{(B)}$, approximation of the bootstrap cdf of $\hat{GF}$ under the null hypothesis in (7), can be estimated. According to a nominal significance level $100\alpha\%$ H_0 can not be rejected if $\Phi_{\mathrm{H}_0}^{(B)}(1-\alpha) \geq \hat{GF}$. Furthermore, the achieved significance level (or p-value) of the test can be computed as the proportion of B bootstrap estimates $\hat{GF}^{(b)}$ lower than the observed $\hat{GF}$. As a rule of thumb, a p-value less than or equal to 0.025 gives a strong evidence against H_0.

Notice that if the density distribution functions corresponding to $\Phi_{\hat{GF}}^{(B)}$ or $\Phi_{\mathrm{H}_0}^{(B)}$ are not symmetric, bias–corrected and accelerated (BC_a) bootstrap percentiles can be computed, Efron and Tibshirani (1993).

By generalizing the hypothesis in (7) to all coefficients of all structural equations, the hypothesis $GF = 0$ against $GF > 0$ can be tested.

5. Application to real data and conclusions

The suggested goodness–of–fit index has been applied to Russet data by referring to the structural model shown in Tenenhaus (1999). The analyzed data consist of 3 manifest variable blocks. $\mathbf{X}_1$: GINI, FARM and Ln(RENT+1); $\mathbf{X}_2$: Ln(GNPR) and Ln(LABO); $\mathbf{X}_3$: Exp(INST-16.3), Ln(ECKS+1), Ln(DEAT+1), DEMOSTAB, DEMOINSTAB and DICTATUR. The measurement model is defined as:

$$\begin{aligned}
\mathbf{x}_{1h} &= \pi_{1h}\xi_1 + \varepsilon_{1h}, \quad h = 1, 2, \ldots, 3 \\
\mathbf{x}_{2h} &= \pi_{2h}\xi_2 + \varepsilon_{2h}, \quad h = 1, 2 \\
\mathbf{x}_{3h} &= \pi_{3h}\xi_3 + \varepsilon_{3h}, \quad h = 1, 2, \ldots, 6
\end{aligned}$$

while the structural model is given by: $\xi_3 = \beta_{31}\xi_1 + \beta_{32}\xi_2 + \epsilon_3$. Finally, mode A for external estimation and factorial scheme for internal estimation are chosen.

The global index is $\hat{GF}^2 = 0.989 \times 0.610 = 0.603$ and, consequently, $\hat{GF} = 0.778$ meaning that the model is able to take into account 78% of the achievable fit.

The obtained results are shown to be statistically significant by the non–parametric procedure suggested above.

The proposed normalized index can be well applied also to compare performances yielded by PLS and LISREL when both feasible on the same model.

References

Chatelin Y., Esposito Vinzi V. and Tenenhaus M. (2002) State–of–art on PLS path modeling throug the aviable software, *HEC Research paper series CR 764.*

Efron B. and Tibshirani R.J. (1993) *An Introduction to the Bootstrap*, Chapman and Hall.

Tenenhaus M. (1998) *La Régression PLS, Theorie et pratique*, Éditions Technip, Paris.

Tenenhaus M. (1999) The PLS approach, *Revue de Statistique Appliquée*, 47.