

Dynamic Structural Equation Models and Pseudo-panel data: a case study

Modelli strutturali dinamici e dati pseudo-panel: analisi di un caso reale

Giuseppe Boari, Gabriele Cantaluppi

Istituto di Statistica, Università Cattolica del S. Cuore, Milano
e-mail: giuseppe.boari@unicatt.it, gabriele.cantaluppi@unicatt.it

Piergiuseppe Merlo, Daniele Gunetti

SANPAOLO IMI, Torino
e-mail: piergiuseppe.merlo@sanpaoloimi.com, daniele.gunetti@sanpaoloimi.com

Riassunto: Attraverso i dati di un'indagine reale viene dapprima verificata la validità di una particolare versione di modello strutturale dinamico, quindi vengono saggiati gli effetti della presenza di dati pseudo-panel in confronto al caso di dati panel.

Keywords: Dynamic Structural Equation Models, Latent Variables, Partial Least Squares, Pseudo-panel data.

1. The static reference model

In Bayol *et al.* (2000), Zanella *et al.* (2002) and Boari, Cantaluppi (2004) the application of structural equation models (SEM) with latent variables in a micro-economic framework is presented, which recalls the approach used to define the well-known national Customer Satisfaction (CS) indices ACSI and ECSI (American and European CS Index).

In Ronsisvalle (2005) these models have been focused and adapted to the analysis of the data collected through a questionnaire, given to a random sample of customers of the branch network of a leading commercial bank: the application of SEMs has focused the fundamental role played by the branch office personnel in the defining process of Customer opinions about the company.

In particular, the final model refers to the following unobservable conceptual constructs (latent variables):

Y_1	Branch personnel relationships
Y_2	Office work organization
Y_3	Value of products/services
Y_4	Company image
Y_5	Overall Customer Satisfaction
Y_6	Customer Loyalty

measured by appropriately related manifest variables collected in the survey.

As known, a structural equation model may be formulated by two components: the inner model, which shows the linear relationships among q exogenous and $p-q$ endogenous latent variables, and the measurement model, relating each latent variable Y_j to its corresponding manifest indicators $X_{jh}, j=1, \dots, p, h=1, \dots, p_j$.

We consider a “recursive” structural model of the form

$$[Y_{q+1} \cdots Y_p]' = [\Gamma|\mathbf{B}][Y_1 \cdots Y_q Y_{q+1} \cdots Y_p]' + [\zeta_{q+1} \cdots \zeta_p]', \quad (1)$$

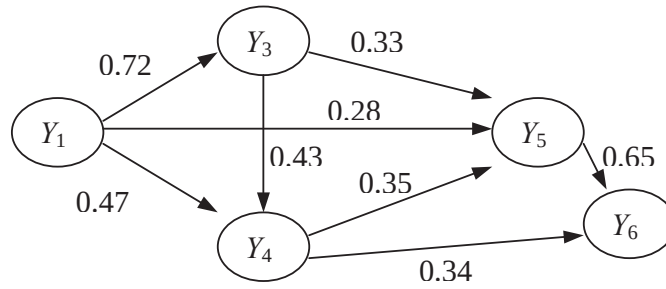
where the elements of the matrix $[\Gamma|\mathbf{B}]$, of dimension $(p-q) \times p$, are the regression coefficients β_{jk} ($j=q+1, \dots, p$, $k=1, \dots, j-1$), the generic endogenous Y_j may depend only on the previous latent variables and the covariance matrix of the equation errors ζ_j is assumed to be diagonal. We also consider a measurement model of the so-called “reflexive” type, where the manifest variables are function of the latent, that is

$$(X_{jh} - \bar{x}_{jh}) = \lambda_{jh} Y_j + \varepsilon_{jh}, \quad j = 1, \dots, p, \quad h = 1, \dots, p_j, \quad (2)$$

where $\bar{x}_{jh}$ are the average reference values and ε_{jh} the measurement errors.

Distributional hypotheses concern null-average and no mutual correlation of the random variables Y_j ($j=1, \dots, p$), ζ_j ($j=q+1, \dots, p$) and ε_{jh} ($j=1, \dots, p$, $h=1, \dots, p_j$).

The analysis of Customer Satisfaction, performed during 2004 via the Partial Least Squares (PLS) methodology, was referred to a sample of 600 interviews and gave rise to the SEM represented by the path diagram in the following picture; it also reports the estimates of the β_{jk} coefficients, all being highly significant.



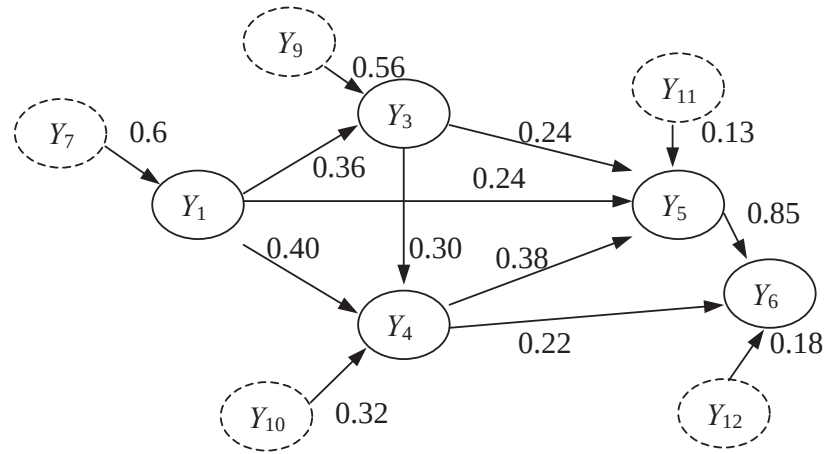
The latent exogenous variable Y_1 (Personnel) plays the main role of driver of all the other endogenous variables; the effect of the dimension Y_2 (Work organization) does not result significant. The λ_{jh} estimates, recalled in (2), do not appear for the sake of brevity.

2 The dynamic model

The availability of the interview data, coming from a survey repeated in 2005 over the same group of customers (panel), allows us to explore the adaptability of the dynamic version of SEM, which is discussed, for example, in Boari, Sala (2004). These models take into account, for instance, the memory effect of some latent variables in the definition of the current level of Customer Satisfaction and Customer Loyalty, which represent the leading indicators to measure the performance of a business organization.

As mentioned in Sala (2005), a simple model, also useful in the present context, may be the dynamic first order model, in the structural component, completed by a simple static scheme, regarding the so-called measurement component. In other words, we add, to the static structural model examined in the preceding paragraph, a linear relationship between every latent variable, there considered, and the corresponding lagged one. Therefore, the analysis over time is performed using a SEM model, where each of the considered 5 latent variables is joined with a lagged variable of the exogenous type, so interpreting the dynamic relationships between every latent variable Y_j , $j=1,3,4,5,6$,

defined at time t , and the corresponding $Y_{j(t-1)} = Y_{j+6}$ defined at the previous time ($t-1$); the latter being represented with dashed ovals in the next picture. Conventionally, the static model is adopted for the analysis of the beginning survey (here referring to 2004). Furthermore, remembering that, when evaluating latent scores, the PLS algorithm involves all adjacent latent variables, the scores calculated at time ($t-1$) are used as the single proxies (with unit weight) of every lagged variable: by forcing the algorithm to operate only on contemporary data, we avoid that future values may be used in the estimation of lagged latent variables.



We may observe that the dynamic relationships between the latent variables Y_5 , Y_6 and the corresponding lagged Y_{11} , Y_{12} are very weak, though their coefficient estimates are significant. Moreover the average CS scores, expressed in the original 1-10 scale, have increased from $M(Y_{11})=7.34$, in 2004, to $M(Y_5)=7.43$ in 2005 (Y_6 and Y_{12} scores may not be compared since they are referred to different proxy variables in the 2 questionnaires).

3 Simulation of pseudo-panel data

Whenever panel data were not available, one may refer to the pseudo-panel data (cfr. Bernini, Lubisco, 2005), which make it possible to investigate the dynamic relationships among homogeneous groups, aggregated with respect to socio-demographic factors. In the present context the following variables have been considered: branch office, age class (18-30, 31-40, 41-55, 56-75) and customer type (Mass market/Affluent). Consequently, 85 pseudo-cohorts have been obtained, clustering homogeneous customers (observe that the demographic classification of customers has been defined with respect to 2004, since units belonging to the same group should be of the same cohort).

Robustness of the procedure was then evaluated by using the average values of these artificial cohorts. The differences in the β_{jk} coefficient estimates, observed when pseudo-panel data are considered, may be determined by the following factors: a) pseudo-cohorts do not include the same respondents over-time, b) the aggregation loses information about the individuals, c) the groups are not homogeneous with respect to their dimension (one possible solution to this last issue may be weighting data in the PLS algorithm).

In order to outline the effects of the preceding sources of estimating errors, results referred to the whole sample (Table 1, Panel) have been compared with: 1) a PLS analysis referred to the average of data clustered in homogeneous groups (Table 1, Average), and 2) two PLS analyses on data obtained by simulating two different pseudo-cohorts of customers, randomly extracting from the 2005 data set two distinct samples of dimensions 368 and 277 individuals, clustered in 85 pseudo-panels (Pseudo 1 and 2).

Table 1. *Pseudo-panel effects on the β_{jk} coefficient estimates.*

	1→3	1→4	1→5	3→4	3→5	4→5	4→6	5→6	7→1	9→3	10→4	11→5	12→6
Panel	0.36	0.40	0.24	0.30	0.24	0.38	0.22	0.85	0.60	0.56	0.32	0.13	0.18
Average	0.30	0.34	0.20	0.37	0.33	0.47	0.21	1.01	0.78	0.69	0.42	ns	ns
Pseudo1	0.76	0.48	0.18	0.57	0.34	0.47	ns	1.33	0.37	0.39	ns	ns	ns
Pseudo2	0.53	0.39	0.24	0.49	0.41	0.31	ns	1.27	0.84	0.57	0.39	ns	ns

The results reported in Table 1 give evidence of the effects of the previous causes of variability in the β_{jk} coefficient estimates.

We may observe that major differences may be observed by comparing the first 2 rows of the table with the remaining ones: in our opinion, the reason is due to the non-homogeneity of the sub-group dimensions, whose dissimilarities are more evident in the sub-samples selected for the simulations.

The effects of the lagged variables on latent variables Y_5 and Y_6 , which are quasi-negligible with panel data, became non significant in pseudo-panel.

References

- Bayol M.P., de la Foye A., Tellier C., Tehenhaus M. (2000) Use of the PLS Path Modelling to estimate the European Customer Satisfaction Index (ECSI) model, *Statistica Applicata*, vol. 12(3), 361-375.
- Bernini C., Lubisco A. (2005) Customer Satisfaction dinamica e indagini sezionali ripetute nel tempo: un'applicazione ai servizi di trasporto urbano, *Atti della Riunione Scientifica: Valutazione e Customer Satisfaction per la Qualità dei Servizi*, Roma 8-9 settembre 2005, Facoltà di Scienze Statistiche, 35-38.
- Boari G., Cantaluppi G. (2004) PLS-VB programme for path modeling with latent variables, *Atti della XLII Riunione Scientifica SIS*, Bari 9-11 giugno 2004, Cleup, 747-750.
- Boari G., Sala E. (2004) Versione dinamica dei modelli ad equazioni strutturali, Serie E.P., n. 121, Istituto di Statistica, Università Cattolica del S. Cuore, Milano.
- Ronsisvalle F. (2005) *L'impiego di tecniche statistiche multivariate nella valutazione della Customer Satisfaction. L'esperienza del Gruppo SANPAOLO IMI*, Tesi di Laurea, Facoltà di Economia, Università degli Studi di Torino.
- Sala E. (2005) *Modelli dinamici strutturali con variabili latenti*, Tesi di Dottorato di Ricerca in Statistica Metodologica ed Applicata, XVII ciclo, Università degli Studi di Milano Bicocca.
- Zanella A., Boari G., Cantaluppi G. (2002) Indicatori statistici complessivi per la valutazione di un sistema per la gestione della qualità: esame del problema ed un esempio di applicazione, in *Atti della Riunione Satellite della XLI Riunione Scientifica SIS*, Cleup, Padova, 1-26.