An R Package to Compare the Wold and Lohmöller PLS Mode B Algorithms

Palabras clave: Partial Least Squares, Wold Algorithm, Lohmöller Algorithm, Simulated Data Sets, R Package

Abstract

Partial least squares path models (PLS-PM) allow the estimation of complex models with multiple dependent and independent latent variables. Each unobserved variable is constructed from a set of manifest variables, or a block of variables measured in the same individuals. The main advantage of PLS-PM is its data analysis approach. The no assumptions about data distribution makes it a very flexible statistical learning technique applicable to many problems. Another advantage is that latent variables are estimated as linear combinations of the manifest variables, so latent variables are constructed as composite variables, supporting the construction, validation, and interpretation of the models. PLS-PM has been successfully applied in marketing, econometrics, psychometrics, education, and in general, all areas where a set of constructs or latent variables can be analyzed and approximated by components. PLS-PM –as a multiblock method– has many applications in other disciplines as chemometrics and engineering and may be very useful to integrate large data sets from many different sources.

The two main algorithms to estimate PLS Mode B path models are the Wold and Lohmöller algorithms. These algorithms aim to find the weights vectors of the linear combinations of the manifest variables in such a way as to maximize a function of the covariance between components. These components are constructed by alternating the estimation of the measurement or outer models –those that relate each latent variable with a set of manifest variables– and the estimation of the structural or inner model –the one that relates the latent variables. The algorithms alternate between the construction of the outer and inner components until they reach a condition for convergence. In a Mode B approach, the algorithms update each weights vector as the vector of regression coefficients in the multiple regression of the component on the corresponding manifest variables. The critical difference between the algorithms is how they update the components when estimating the structural model. While the Lohmöller algorithm updates the components using a Jacobi-type iteration, Wold's algorithm updates the components using a Gauss-Seidel-type iteration. The type of iteration and the properties of the data matrices determine the convergence properties of the algorithms.

In this work, I present an R package to compare the Wold and Lohmöller algorithms with centroid and factorial weighting schemes to estimate PLS Mode B path models. The objective is to compare the performance and main properties of the algorithms by executing

Monte Carlo simulations. The principal features of the implementation are as follows. It allows the generation of simulated data for PLS Mode B models –the literature reports a few procedures to create data for these models because it is difficult to generate data for the dependent or endogenous latent variables. Several PLS Mode B setups can be simulated, and the user can vary conditions such as the number of observations and variables. The values of the weights vectors may also be initialized. The choice of the initial values is especially important to ensure the convergence of the algorithm to a local or global optimum. The estimates of the relationships between the variables are compared with the values assumed as true. Then, it is possible to understand the algorithm's performance in approximating the true values. The comparison is reported in terms of mean bias ($\frac{1}{t}\sum_{i=1}^{t}\theta - E[\theta_i]$), mean relative bias ($MRB = 100 * \frac{1}{t}\sum_{i=1}^{t}\frac{\theta - E[\theta_i]}{\theta}$), and mean square error ($MSE = Bias^2 + Variance$). In addition, the application reports the value of the maximized function depending on the number of iterations for reaching the condition for convergence and the value of the condition versus the number of iterations. Two objective functions are examined: the sum of the square correlations between components (SSQCOR criterion) and the sum of the absolute value of the correlations between components (SABSCOR criterion). All results are reported graphically, then it is possible to understand the growth of the criteria and the rate of convergence of the algorithms.

References

- Chu, M. T., & Watterson, J. L. (1993) On a multivariate eigenvalue problem. I: Algebraic theory and a power method. SIAM Journal of Scientific Computing, 41(5):1089-1106.
- Hanafi, M. (2007) PLS path modeling: Computation of latent variables with the estimation mode B. Computational Statistics, 22:275-292.
- Kettenring, J. R. (1971). Canonical analysis of several sets of variables. Biometrika 58(3):433-451.
- Lohmöller, J. B. (1989) Latent variables path modeling with partial least squares. Heildelberg: Physical-Verlag.
- Martinez-Ruiz, A. (2011) Patent value models: Partial least squares path modelling with mode C and few indicators. PhD Thesis, Universitat Politècnica de Catalunya BarcelonaTech. https://upcommons.upc.edu/handle/2117/94865
- Martinez-Ruiz, A., Montañola-Sales, C. (2019) Big data in multi-block data analysis: An approach to parallelizing Partial Least Squares Mode B algorithm. Heliyon, 5(4), e01451. https://doi.org/10.1016/j.heliyon.2019.e01451
- Mathes, H. (1993) Global optimization criteria of the PLS-algorithm in recursive path models with latent variables. In: Haagen, K., Bartholomew, D.J., & Deistler, M. (Eds.), Statistical modeling and latent variables (pp. 229-248). Elsevier Science Publishers.
- Wold, H. (1982) Soft modeling: The basic design and some extensions. In: Jöreskog, K. G. & Wold, H. (Ed.), Systems under indirect observation, part II (pp. 1-54). Amsterdam: North-Holland.