What is Predictive about Partial Least Squares?

Galit Shmueli
University of Maryland, USA

Otto Koppius
Erasmus University, The Netherlands

Abstract

Partial least squares (PLS) estimation of path models has become very popular in IS research, as an alternative to covariance-based methods. PLS path modeling is often referred to as being useful for “predictive” applications. In this work, we investigate the predictive aspects of PLS path modeling and its relation to predictive analytics and predictive assessment. In particular, we compare it to neural networks, which share several similarities. We conclude that PLS path modeling (the dominant form of usage in IS) is primarily causal-explanatory in nature.

Introduction

Partial Least Squares (PLS) estimation has become a popular method in information systems research, especially when it comes to analyzing adoption and usage of new electronic commerce applications (e.g., Pavlou and Fygenson (2006)). It is most commonly used to estimate structural equation models in place of covariance-based approaches (e.g., LISREL) in order to overcome the latter’s challenges (“technically demanding, often resulting in analytical errors, and, even if modeled correctly, is not a complete solution”, Chin et al. 2003, p.191). In structural equation models, one or more of the variables is assumed to be unobserved (latent). Theory is used to generate a causal diagram that relates the different variables (observable and latent) to each other. Then, data are used to quantify the hypothesized relationships, and the estimated relationships are used to test causal hypotheses. Hence, at their core, latent variable models are causal-explanatory. There have been claims that PLS path modeling is “causal-predictive” rather than causal-explanatory (Jöreskog and Wold, 1982; Anderson and Garbing, 1988). In this work we investigate the predictive aspects of PLS in light of the distinction between causal explanation and prediction by Shmueli & Koppius (2009). In particular, we look at the three steps in PLS path modeling: model estimation, validation, and use.

PLS and Neural Networks

To highlight the predictive nature of PLS path modeling, we examine it against a classic predictive model that appears somewhat similar: neural networks (NN) (Hackl and Westlund, 2000). NN are popular in data mining for predicting an outcome from a set of predictors. They are considered a “blackbox" in terms of interpretability and have been highly successful in terms of predictive accuracy. Similar to the use of latent variables (LVs) in PLS, in NN there is a diagram that connects the observable inputs and outputs via a network of unobservable (hidden) nodes and the NN algorithm then estimates the weights on the arrows connecting the different nodes. However, unlike path models, there is no underlying causal theoretical model, and the hidden nodes are conceptually unimportant. Hsu et al. (2006) compared PLS and NN numerically on simulated and real data. The simulations yielded similar results, whereas the real data yielded similar LV scores but different path coefficients.

Model estimation: Two sets of models are estimated: *outer models* relate each latent variable to its observed variables (the measurement model), and *inner models* relate the latent variables to each other (the structural model). Each model is estimated using OLS regression (thereby assuming linear relationships). The path coefficients are then obtained via correlation or OLS regression between subsets of the estimated latent variables and/or observed variables. Estimation is done iteratively, until the various parameters are stable, by minimizing the *various sums of squared errors*. In other words, the final model best fits each of the regression models. Neural net estimation is similarly iterative. However, the algorithm tries to directly minimize the error of the *observed output variables*. NN also allow non-linear relationships between nodes. In light of these differences, PLS estimation is more similar to causal-explanatory modeling than to predictive modeling.

Model validation: Validation is achieved by comparing the resulting estimates to the theoretical expectations. In particular, if coefficients do not have the right sign, then the corresponding observed variables are typically removed (Tenenhaus et al., 2005). Popular model validation measures include *communality* and *redundancy.*

When computed by cross-validation (blindfolding), they measure how well the latent variables predict the observed variables (input or output). In contrast, in NN and in general predictive assessment, one examines how well the *observed input variables* predict the *observed output variables*. Hence, blindfolding serves more as a verification of the integrity of the estimated model in fitting the theoretical relationship between the observed and latent variables, rather than evaluates practical predictive power.

Model use: The estimated PLS path model is used for testing causal hypotheses related to the structural diagram. Software output includes estimated weights, loadings, path coefficients, and correlations between latent variables. These are then examined carefully and used for inference. Due to the lack of an explicit overall statistical model and the nature of PLS estimation, statistical inference is achieved by using bootstrapping. Finally, path models are typically used only for causal testing and not for predicting new observations. In contrast, typical NN output does not include any inference-related information beyond the weights, and the focus is on its predictive accuracy on a holdout set (and comparing the training and holdout performance to detect over-fitting). The model is then used to predict new data. In summary, the final use of the estimated path model differs markedly from that of a predictive model.

Conclusions

In line with the differentiation between explanatory-causal statistical models and predictive analytics by Shmueli & Koppius (2009), our preliminary conclusion is that PLS path modeling is a *causal-explanatory* approach. Although PLS has some similarities with the predictive neural networks algorithm, its heavy reliance on a causal theoretical model, the conceptual importance of the latent variables, its method for model validation and its final use deem it very different from predictive models and from practical predictive use.

If predictive power is important to PLS modelers then more emphasis should be given to the relationship between the *observed* input and output variables and the ability of the former to accurately predict the latter. Changes to the estimation procedure can also put more emphasis on prediction: Tenenhaus et al. (2005) suggested using PLS regression in place of OLS regression for outer model estimation. PLS regression is a predictive, shrinkage-based method which is indeed geared towards prediction. Finally, this is work in progress, and we plan to illustrate the various aspects mentioned above via simulation, the results of which will be presented at SCECR.

References

Anderson J C and Gerbing D W (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach", *Psychological Bulletin,* **103**(3), pp. 411-423.

Chin W W, Marcolin B L, and Newsted P R (2003), "A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and an Electronic-Mail Emotion/Adoption Study", *Information Systems Research,* **14**(2), pp. 189-217.

Hackl P and Westlund A H (2000) "On Structural Equation Modelling for Customer Satisfaction Measurement", *Total Quality Management*, **11**, pp. S820-S825.

Hsu S-H, Chen W-H, and Hsieh M-J (2006), "Robustness Testing of PLS, LISREL, EQS and ANN-based SEM for Measuring Customer Satisfaction", *Total Quality Management*, **17**(3), pp. 355–371.

Jöreskog K G and Wold H (1982), "The ML and PLS Techniques For Modeling with Latent Variables: Historical and Comparative Aspects", in *Systems Under Indirect Observation: Causality, Structure, Prediction* (Vol. I), Amsterdam: North-Holland, pp. 263-270.

Pavlou P A and Fygenson M (2006), "Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior", *MIS Quarterly,* **30**(1), pp. 115-143.

Shmueli G and Koppius O (2009), "The Challenge of Prediction in Information Systems Research", *Working Paper RHS 06-058, School of Business, UMD* (http://ssrn.com/abstract=1112893)

Tenenhaus M, Vinzi V E, Chatelin Y-M, and Lauro C (2005), "PLS Path Modeling", *Computational Statistics & Data Analysis*, **48**, pp. 159-205.