MIS Quarterly

RESEARCH ESSAY

CONSISTENT PARTIAL LEAST SQUARES PATH MODELING

Theo K. Dijkstra
Faculty of Economics and Business, University of Groningen, Nettelbosje 2,
9747 AE Groningen THE NETHERLANDS {t.k.dijkstra@rug.nl}

Jörg Henseler
Faculty of Engineering Technology, University of Twente, Drienerlolaan 5,
7522 NB Enschede THE NETHERLANDS {j.henseler@utwente.nl}
NOVA IMS, Universidade Nova de Lisboa, 1070-312 Lisbon PORTUGAL {jhenseler@isegi.unl.pt}

Appendix A

In variance-based SEM, the path coefficients can be determined based on the construct score correlation matrix **R**, which for the model depicted in Figure 1 is of the following form ("cor" = correlation; "rel" = reliability):

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \operatorname{cor}(\xi_1,\xi_2)\cdot\sqrt{\operatorname{rel}(\widetilde{\xi}_1)\cdot\operatorname{rel}(\widetilde{\xi}_2)} & \operatorname{cor}(\xi_1,\eta)\cdot\sqrt{\operatorname{rel}(\widetilde{\xi}_1)\cdot\operatorname{rel}(\widetilde{\eta})} \\ \operatorname{cor}(\xi_1,\xi_2)\cdot\sqrt{\operatorname{rel}(\widetilde{\xi}_1)\cdot\operatorname{rel}(\widetilde{\xi}_2)} & 1 & \operatorname{cor}(\xi_2,\eta)\cdot\sqrt{\operatorname{rel}(\widetilde{\xi}_2)\cdot\operatorname{rel}(\widetilde{\eta})} \\ \operatorname{cor}(\xi_1,\eta)\cdot\sqrt{\operatorname{rel}(\widetilde{\xi}_1)\cdot\operatorname{rel}(\widetilde{\eta})} & \operatorname{cor}(\xi_2,\eta)\cdot\sqrt{\operatorname{rel}(\widetilde{\xi}_2)\cdot\operatorname{rel}(\widetilde{\eta})} & 1 \end{bmatrix}$$

From **R**, we can extract the submatrix $\mathbf{R}_{\mathbf{X}} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$

and the subvector $\mathbf{R}_{\mathbf{Xy}} = \begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix}$.

Under the assumption that the construct scores are standardized, the regression equation equals

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \mathbf{R}_{\mathbf{X}}^{-1}\mathbf{r}_{\mathbf{Xy}} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}^{-1}\begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix} = \frac{1}{\det(\mathbf{R}_{\mathbf{X}})}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}\begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix} = \frac{1}{1-r_{12}^2}\begin{bmatrix} r_{13}-r_{12}r_{23} \\ r_{23}-r_{12}r_{13} \end{bmatrix}$$

Since the true β_1 equals zero, we can exploit that $\operatorname{cor}(\xi_1, \eta) = \operatorname{cor}(\xi_1, \xi_2) \cdot \operatorname{cor}(\xi_2, \eta)$ and thus find the following result for the path $\hat{\beta}_1$ coefficient as estimated by variance-based SEM:

$$\hat{\beta}_1 = \frac{1}{1-r_{12}^2}\cdot\left(r_{13}-r_{12}r_{23}\right)$$

$$= \frac{\operatorname{cor}(\xi_1,\eta)\cdot\sqrt{\operatorname{rel}\left(\tilde{\xi}_1\right)\cdot\operatorname{rel}(\tilde{\eta})}-\operatorname{cor}(\xi_1,\xi_2)\cdot\sqrt{\operatorname{rel}\left(\tilde{\xi}_1\right)\cdot\operatorname{rel}\left(\tilde{\xi}_2\right)}\cdot\operatorname{cor}(\xi_2,\eta)\cdot\sqrt{\operatorname{rel}\left(\tilde{\xi}_2\right)\cdot\operatorname{rel}(\tilde{\eta})}}{1-\operatorname{cor}^2\left(\tilde{\xi}_1,\tilde{\xi}_2\right)}$$

$$= \frac{\operatorname{cor}(\xi_1,\xi_2)\cdot\operatorname{cor}(\xi_2,\eta)\cdot\sqrt{\operatorname{rel}\left(\tilde{\xi}_1\right)\cdot\operatorname{rel}(\tilde{\eta})}-\operatorname{cor}(\xi_1,\xi_2)\cdot\sqrt{\operatorname{rel}\left(\tilde{\xi}_1\right)\cdot\operatorname{rel}\left(\tilde{\xi}_2\right)}\cdot\operatorname{cor}(\xi_2,\eta)\cdot\sqrt{\operatorname{rel}\left(\tilde{\xi}_2\right)\cdot\operatorname{rel}(\tilde{\eta})}}{1-\operatorname{cor}^2\left(\tilde{\xi}_1,\tilde{\xi}_2\right)}$$

$$= \frac{\operatorname{cor}(\xi_1,\xi_2)\cdot\operatorname{cor}(\xi_2,\eta)\cdot\sqrt{\operatorname{rel}\left(\tilde{\xi}_1\right)\cdot\operatorname{rel}(\tilde{\eta})}\cdot\left(1-\operatorname{rel}\left(\tilde{\xi}_2\right)\right)}{1-\operatorname{cor}^2\left(\tilde{\xi}_1,\tilde{\xi}_2\right)}$$

$$= \frac{\operatorname{cor}\left(\tilde{\xi}_1,\tilde{\xi}_2\right)}{1-\operatorname{cor}^2\left(\tilde{\xi}_1,\tilde{\xi}_2\right)}\cdot\operatorname{cor}\left(\tilde{\xi}_2,\tilde{\eta}\right)\cdot\frac{1-\operatorname{rel}\left(\tilde{\xi}_2\right)}{\operatorname{rel}\left(\tilde{\xi}_2\right)}$$

Appendix B[1]

A starting point for PLS-analyses is the so-called basic design, in essence a factor model. It is assumed that we have N i.i.d.[2] column vectors of observed scores $y_1, y_2, y_3, \ldots, y_N$ (i.e., we have N observations). We assume the usual standardization in PLS of each indicator, which means that each vector y_n for $n = 1, 2, 3, \ldots, N$ has zero mean and its elements have unit variance. The vectors can be partitioned into M subvectors, $M \geq 2$, each with at least two components. For the i^{th} subvector y_{in} of y_n, we have

$$y_{in} = \lambda_i \cdot \eta_{in} + \varepsilon_{in} \tag{1}$$

where the loading vector λ_i and the vector of measurement errors εi_n have the same dimensions as yi_n, and the unobservable latent variable η_{in} is real-valued.[3] For convenience, we make the sufficient but by no means necessary assumption that all components of all error vectors are mutually independent, and independent of all latent variables.

The latent variables have also zero mean and unit variance. The correlation between η_{in} and η_{jn} will be denoted by ρ_{ij}. A particular set of easy implications is that the covariance matrix Σ_{ii} of y_{in} can be written as

$$\sum_{ij} := Ey_{ij}y_{jn}^T = \rho_{ij}\lambda_i\lambda_j^T \tag{2}$$

where the covariance matrix of the measurement errors Θ_i is diagonal with non-negative diagonal elements, and we have for the covariance between y_{in} and y_{jn}

---

[1]To keep the paper self-contained, we collect here the main algebraic results underlying PLSc (for a more elaborate explication, see Dijkstra 2010). Note that there are many ways to correct for inconsistency and the one we use here is quite possibly the simplest one. See Dijkstra (2013) for alternatives and extensions. Any combination of PLS with one of the alternatives could legitimately be called PLSc.

[2]The use of i.i.d. (independent and identically distributed) signifies that the data are seen as a random sample from a population. This is for conceptual convenience only. Together with the classical laws of large numbers and the central limit theorem (Cramér 1946), this entails the consistency and asymptotic normality properties we need. We can make much more general assumptions, but the statististical subtleties would be a distraction and would not lead to different results (CAN-estimators).

[3]We follow exactly the specification of Joreskög (1969). Please note that variables are arranged per rows, and observations per columns. Moreover, note that, as opposed to our custom in the main body of the paper, it pays here to use a subscript for the blocks.

$$\sum_{ij} := Ey_{in}y_{jn}^T = \rho_{ij}\lambda_i\lambda_j^T \tag{3}$$

The covariance matrix of each y_n will be denoted by Σ, and the sample covariance matrix by S. The sample counterparts of Σ_{ii} and Σ_{ij} are denoted by S_{ii} and S_{ij}, respectively. Since we assume that the sample data are standardized before being analyzed, we have $S_{ij} = \frac{1}{N}\sum_{n=1}^{N} y_{in}y_{jn}^T$. Note that the assumptions made so far entail that the sample counterparts are consistent and jointly asymptotically normal estimators of the theoretical variance and covariance matrices (Cramér 1946, Chapter 28).

PLS features a number of iterative *fixed-point* algorithms, of which we select one, the so-called Mode A algorithm. This is, in general, numerically the most stable algorithm.[4] As a rule, it converges from arbitrary starting vectors, and it is usually very fast. The outcome is an estimated weight vector $\hat{w}$, with typical subvector $\hat{w}_i$ of the same dimensions as y_{in}. With these weights, *sample proxies* are defined for the latent variables: $\hat{\eta}_{in} := \hat{w}_i^T y_{in}$ for η_{in}, with the customary normalization of a unit sampling variance, so $\frac{1}{N}\sum_{n=1}^{N}\left(\hat{w}_i^T y_{in}\right)^2 = \hat{w}_i^T S_{ii}\hat{w}_i = 1$. In Wold's PLS approach the $\hat{\eta}_{in}$'s replace the unobserved latent variables, and loadings and structural parameters are estimated using ordinary least squares.[5] For Mode A, we have for each i ($\propto$ stands for "proportional to")

$$\hat{w}_i \propto \sum_{j\in C(i)} sign_{ij}\cdot S_{ij}\hat{w}_j \tag{4}$$

Here, $sign_{ij}$ is the sign of the sample correlation between $\hat{\eta}_i$ and $\hat{\eta}_j$, and $C(i)$ is a set of indices of latent variables. Traditionally, $C(i)$ contains the indices of latent variables adjacent to $\hat{\eta}_i$ (i.e., the indices of latent variables that appear on the other side of the structural or path equations in which $\hat{\eta}_i$ appears). Clearly, $\hat{w}_i$ is obtained by a regression of the indicators yi_n on the sign-weighted sum: $\sum_{y\in C(i)} sign_{ij}\cdot\hat{\eta}_{jn}$. There are other versions (with correlation weights, for example); this is one of the very simplest, and it is the original one (see Wold, 1982). There is little motivation in the PLS literature for the coefficients of $S_{ij}\hat{w}_j$,[6] but the particular choice can be shown to be irrelevant for the probability limits of the estimators. The algorithm takes an arbitrary starting vector, and then basically follows the sequence of regressions for each i, each time inserting updates when available (or after each full round; the precise implementation is not important).

Dijkstra (1981, 2010) has shown that the PLS modes converge with a probability tending to one when the sample size tends to infinity, for essentially arbitrary starting vectors. Moreover, the weight vectors that satisfy the fixed-point equations, are locally continuously differentiable functions of the sample covariance matrix of y.They, as well as other estimators that depend smoothly on the weight vectors and S, are therefore jointly asymptotically normal.

Let us denote the probability limit of $\hat{w}_i$, plim $\hat{w}_i$, by $\overline{w}_i$. We can get it from the equation for $\hat{w}_i$ by substitution of Σ for S.

$$\overline{w}_i \propto \sum_{j\in C(i)} sign_{ij}\cdot\sum_{ij}\overline{w}_j = \sum_{j\in C(i)} sign_{ij}\cdot\rho_{ij}\lambda_i\lambda_i^T\overline{w}_j = \lambda_i\sum_{j\in C(i)} sign_{ij}\cdot\rho_{ij}\lambda_j^T\overline{w}_j \tag{5}$$

Since $\lambda_i^T\overline{w}_j$ is a scalar, and all terms in the sum have the vector λ_i in common, it is clear that $\overline{w}_i \propto \lambda_i$. We must have $\overline{w}_i^T\sum_{ii}\overline{w}_i = 1$, so

$$\overline{w}_i = \frac{\lambda_i}{\sqrt{\lambda_i^T\sum_{ii}\lambda_i}} \tag{6}$$

---

[4]This is because Mode A ignores collinearity between the observed variable predictors of the proxy, while Mode B takes account of that colllinearity and thus must find the inverse of the covariance matrix of the predictions, which itself is sometimes unstable.

[5]Other estimators like 2SLS are also possible.

[6]In favor of the sign-weights, one could argue that in sufficiently large samples $sign_{ij}\cdot S_{ij}\hat{w}_j$ is approximately equal to $\lambda_i\cdot\left|\rho_{ij}\right|\cdot\left(\lambda_j^T\overline{w}_j\right)$, where the term in brackets measures the (positive) correlation between η_j and its proxy; see below for results that help justify this claim. So the tighter the connection between η_j, and the better η_j can be measured, the more important $\hat{\eta}_j$ is in determining $\hat{w}_i$

We conclude that PLS, Mode A, produces estimated weight vectors that tend to vectors proportional to the true loadings. One would like to have a simple estimate for the proportionality factor. We propose here as in Dijkstra (1981, 2010, 2011) to define $\hat{\lambda}_i := \hat{c}_i \cdot \hat{w}_i$, where the scalar $\hat{c}_i$ is such that the *off-diagonal* elements of S_{ii} are reproduced as well as possible in a least squares sense. So we minimize the Euclidean distance between[7]

$$\left[S_{ii} - diag(S_{ii})\right] \text{ and } \left[(c_i \cdot \hat{w}_i)(c_i \cdot \hat{w}_i)^T - diag\left((c_i \cdot \hat{w}_i)(c_i \cdot \hat{w}_i)^T\right)\right] \tag{7}$$

as a function of c_i and obtain

$$\hat{c}_i := \left[\frac{\hat{w}_i^T\left(S_{ii} - diag(S_{ii})\right)\hat{w}_i}{\hat{w}_i^T\left(\hat{w}_i\hat{w}_i^T - diag\left(\hat{w}_i\hat{w}_i^T\right)\right)\hat{w}_i}\right]^{\frac{1}{2}} \tag{8}$$

The use of "diag" means that only the off-diagonal elements of the matrices are taken into account. So the numerator within brackets is just $\sum_{a \neq b} \hat{w}_{ia}\hat{w}_{ib}S_{ii,ab}$ and similarly for the denominator. In sufficiently large samples, $\hat{c}_i$ will be well-defined, real, and positive. (In all samples in this paper and those in other studies, $\hat{c}_i$ attained proper values.) Its calculation does not require additional numerical optimization. It is straightforward to verify, by replacing S_{ii} by Σ_{ii} and $\hat{w}_i$ by $\overline{w}_i$, that the correction does its job: the matrix in the denominator equals the matrix in the numerator, apart from a factor $\frac{1}{\lambda_i^T \sum_{ii} \lambda_i}$, so

$$\overline{c}_i := plim\ \hat{c}_i = \sqrt{\lambda_i^T \sum_{ii} \lambda_i} \tag{9}$$

Now, in particular

$$plim\ \hat{\lambda}_i = plim\left(\hat{c}_i \cdot \hat{w}_i\right) = \overline{c}_i \cdot \overline{w}_i = \lambda_i \tag{10}$$

It will be useful to define a population proxy $\overline{\eta}_{in}$ by $\overline{\eta}_{in} := \overline{w}_i^T y_{in}$. Clearly, the squared correlation between a population proxy and its corresponding latent variable is

$$R^2\left(\eta_i, \overline{\eta}_i\right) = \left(\overline{w}_i^T \lambda_i\right)^2 \tag{11}$$

which equals

$$\left(\lambda_i^T \lambda_i\right)^2 \div \lambda_i \Sigma_{ii} \lambda_i = \frac{\left(\lambda_i^T \lambda_i\right)^2}{\left(\lambda_i^T \lambda_i\right)^2 + \lambda_i^T \Theta_i \lambda_i} \tag{12}$$

With a large number of high quality indicators, this correlation can be close to one (:"consistency at large" in PLS parlance). A trivially deduced but important algebraic relationship is

$$R^2\left(\overline{\eta}_i, \overline{\eta}_j\right) = \left(\overline{w}_i^T \Sigma_{ij} \overline{w}_j\right)^2 = \rho_{ij}^2 \cdot R^2\left(\eta_i, \overline{\eta}_i\right) \cdot R^2\left(\eta_j, \overline{\eta}_j\right) \tag{13}$$

indicating that the PLS proxies will tend to underestimate the squared correlations between the latent variables. In fact, one can show that this is true for multiple correlations as well (see Dijkstra 2010). Also note that

$$R^2\left(\eta_i, \overline{\eta}_j\right) = \left(\overline{w}_i^T \lambda_i\right)^2 = \left(\overline{w}_i^T \cdot \left(\overline{w}_i \cdot \overline{c}_i\right)\right)^2 = \left(\overline{w}_i^T \overline{w}_i\right)^2 \cdot \overline{c}_i^2 \tag{14}$$

so that we can estimate the (squared) quality of the proxies consistently by

[7]This assumes that the measurement errors within a block are uncorrelated, the basic design. If we know that some errors are correlated or we have doubts about them, we can delete the items from the difference to be minimized.

$$\hat{R}^2\left(\eta_i, \overline{\eta}_i\right) := \left(\hat{w}_i^T \hat{w}_i\right)^2 \cdot \hat{c}_i^2 \tag{15}$$

Moreover, with

$$\hat{R}^2\left(\overline{\eta}_i, \overline{\eta}_j\right) := \left(\hat{w}_i^T S_{ij} \hat{w}_j\right)^2 \cdot \hat{c}_i^2 \tag{16}$$

we can estimate the correlations between the latent variables consistently (see Equation 13)

$$\hat{\rho}_{ij} := \sqrt{\frac{\hat{R}^2\left(\overline{\eta}_i, \overline{\eta}_j\right)}{\hat{R}^2\left(\eta_i, \overline{\eta}_i\right) \cdot \hat{R}^2\left(\eta_j, \overline{\eta}_j\right)}} \tag{17}$$

We close this appendix with four observations:

1. Standard PLS software for Mode A will produce all the necessary, simple ingredients for consistent estimation.

2. The approach can be and has been extended to Mode B (Dijkstra 1981, 2010, 2011), but since Mode A is numerically more stable and faster than Mode B, it has first priority.

3. In the main body of this paper, we used $\rho(\eta_i)$ for $R(\eta_i, \overline{\eta}_i)$.

4. In the main body of this paper, we used (without subscript i)

$$\rho_{A,i} := \left(\hat{w}_i^T \hat{w}_i\right)^2 \cdot \frac{\hat{w}_i^T\left(S_{ii} - diag\left(S_{ii}\right)\right)\hat{w}_i}{\hat{w}_i^T\left(\hat{w}_i, \hat{w}_i^T - diag\left(\hat{w}_i, \hat{w}_i^T\right)\right)\hat{w}_i} \tag{18}$$

   This is just $\hat{R}^2\left(\eta_i, \overline{\eta}_i\right)$.

References

Cramér, H. 1946. *Mathematical Models of Statistics* (Volume 9), Princeton, NJ: Princeton University Press.

Dijkstra, T. K. 1981. *Latent Variables in Linear Stochastic Models: Reflections on "Maximum Likelihood" and "Partial Least Squares" Methods*, Ph.D. Thesis, Groningen University. (A second edition was published in 1985 by Sociometric Research Foundation, Amsterdam.)

Dijkstra, T. K. 2010. "Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares," in *Handbook of Partial Least Squares: Concepts, Methods, and Applications*,V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang (eds.), New York: Springer, pp. 23-46.

Dijkstra, T. K. 2011. "Consistent Partial Least Squares Estimators for Linear and Polynomial Factor Models," working paper, University of Groningen, Groningen, The Netherlands (http://www.rug.nl/staff/t.k.dijkstra/research).

Dijkstra, T. K. 2013. "A Note on How to Make PLS Consistent," working paper, University of Groningen, Groningen, The Netherlands (http://www.rug.nl/staff/t.k.dijkstra/how-to-make-pls-consistent.pdf).

Jöreskog, K. G. 1969. "A General Approach to Confirmatory Maximum Likelihood Factor Analysis," *Psychometrika* (34:2), pp. 183-202.

Wold, H. O. A. 1982. "Soft Modeling: The Basic Design and Some Extensions," in *Systems under indirect observation,* K. G. Jöreskog and H. O. A. Wold (eds.), Amsterdam: North-Holland, pp. 1-54.