Consistent Partial Least Squares for Nonlinear Structural Equation Models

Theo K. Dijkstra
Department of Economics, Econometrics & Finance
University of Groningen
Groningen, The Netherlands
E-Mail: t.k.dijkstra@rug.nl

Karin Schermelleh-Engel
Department of Psychology
Goethe University
Frankfurt am Main, Germany
E-Mail: schermelleh-engel@psych.uni-frankfurt.de
Abstract

Partial Least Squares as applied to models with latent variables, measured indirectly by indicators, is well-known to be inconsistent. The linear compounds of indicators that PLS substitutes for the latent variables do not obey the equations that the latter satisfy. We propose simple, non-iterative corrections leading to consistent and asymptotically normal (CAN)-estimators for the loadings and for the correlations between the latent variables. Moreover, we show how to obtain CAN-estimators for the parameters of structural recursive systems of equations, containing linear and interaction terms, without the need to specify a particular joint distribution. If quadratic and higher order terms are included, the approach will produce CAN-estimators as well when predictor variables and error terms are jointly normal. We compare the adjusted PLS, denoted by PLSc, with Latent Moderated Structural Equations (LMS), using Monte Carlo studies and an empirical application.

Keywords: consistent partial least squares, latent moderated structural equations, nonlinear structural equation model, interaction effect, quadratic effect
Consistent Partial Least Squares for Nonlinear Structural Equation Models

Introduction

The estimation and assessment of nonlinear structural equation models (SEM) has attracted a great deal of attention in the last two decades. The incorporation of nonlinearities, interaction between latent variables in particular, has elevated fitting and testing issues to new technical heights. There is an extensive literature with a large diversity of approaches in the tradition that builds on covariance-based SEM. Of the many one could refer to, we mention: product indicator approaches using maximum likelihood (e.g., Jöreskog & Yang, 1996; Marsh, Wen, & Hau, 2004; Moosbrugger, Schermelleh-Engel, Kelava, & Klein, 2009; Schermelleh-Engel, Klein, & Moosbrugger, 1998) based on the seminal paper by Kenny and Judd (1984), two-stage least-squares (Bollen, 1996; Bollen & Paxton, 1998; Jöreskog & Yang, 1996), the orthogonalizing approach (Little, Bovaird, & Widaman, 2006), moment-based approaches (Wall & Amemiya, 2003; Mooijaart & Bentler, 2010), Bayesian estimation methods (e.g., Lee, Song, & Poon, 2007; Lee, Song, & Tang, 2007), and distribution-analytic approaches using maximum likelihood or quasi-maximum likelihood methods (Klein & Moosbrugger, 2000; Klein & Muthén, 2007).

There is an alternative tradition, sometimes referred to as components-based SEM, that builds on the seminal works of Herman Wold (1966, 1975, 1982): Partial Least Squares (PLS). PLS has become a vibrant field of both applied and theoretical research (see, e.g., Tenenhaus, Esposito Vinzi, Chatelin, & Lauro, 2005, and the Handbook of Partial Least Squares by Esposito Vinzi, Chin, Henseler, & Wang, 2010, for extensive overviews).

A relatively recent development is the incorporation of nonlinearities, as in
particular for a product indicator approach (Chin, Marcolin, & Newsted, 2003), a 2-stage approach (Chin et al., 2003; Henseler & Fassott, 2010), a hybrid approach (Wold, 1982), and an orthogonalizing approach (Henseler & Chin, 2010; Little et al., 2006).

PLS can be viewed as an extension of principal components or canonical variables analysis, or alternatively, as a technique for estimating latent variable models. It is the latter use that is the more controversial, for the following reason: An essential feature of PLS is the (iterative) construction of linear compounds of indicator variables as proxies or stand-ins for the latent variables, and the use of estimated relationships between the linear compounds as estimates for the relationships between the latent variables. However, the relationships between the former can never replicate those between the latter, apart from sets of measure zero in the parameter space. A practical consequence is that PLS tends to overestimate the loadings in absolute value, and to underestimate multiple and bivariate (absolute) correlations between the latent variables (see below, see also Dijkstra, 1981, 1983, 2010, for an asymptotic analysis; and of the many one could refer to for Monte Carlo simulations, see, e.g., Schermelleh-Engel, Werner, Klein, & Moosbrugger, 2010). It is possible however to correct for this tendency.

Here we take up the suggestions made by Dijkstra (1981, 1985, 2010) and apply them to certain recursive sets of nonlinear structural equations. We go substantially beyond the former references, where the discussions concern corrections to loading vectors and corrections to correlations between latent variables only. Our extension includes the structural parameters that measure interaction and other nonlinear effects that can be captured by polynomial models. We exploit the fact that these structural parameters satisfy certain moment equations whose coefficients can be estimated consistently by PLS provided its estimators are properly corrected. The corrections we propose are simple, and attainable with little effort using standard PLS software. So the approach proposed here, denoted by PLSc, shares with PLS its well-known, high computational stability and
efficiency. The advantage of PLSc is that it is ‘well calibrated’, i.e. it will produce the true parameter values for the models we discuss when applied to the ‘population’. The ensuing parameter estimators are also consistent and asymptotically normal (CAN) under standard assumptions. Some might hope therefore that the estimators are ‘about right on the average in finite samples’. In addition it would be nice when their accuracy as measured by their standard deviations does not leave too much to be desired when compared to a state-of-the-art method like Latent Moderated Structural Equations (LMS), even when applied under ideal conditions. Below we will report some results that are not discouraging. (For a separate analysis of linear structural equations with their own intricacies and possibilities for estimation and testing we refer to Dijkstra & Henseler, 2012).

An important point to appreciate is that PLSc is a ‘partial information method’: the linear compounds and the corrections are determined without using the detailed specification of the entire structural model. It cannot be expected therefore to be as asymptotically efficient as full information methods. Formulated more positively, the implementation of PLSc does not require a complete and correct specification of the structural model. It may therefore be more robust against specification errors in the structural equations. An elaborate investigation of robustness will not be performed here. However, we will first report some results for the standard case. We refer to the conclusion for this point and others on the research agenda.

The paper is structured as follows. In section 2 we elaborate on the basic idea behind PLSc. It will be seen that without the need to specify a joint distribution of the indicators, one can get consistent and asymptotically normal estimators for the parameters of recursive structural equation systems with linear and interaction terms. We show that when the recursive system contains quadratic and higher order terms as well, the approach is also valid when (a sufficient condition) the measurement errors and
equation errors, as well as the explanatory latent variables of the first equation of the recursive system, are jointly normal. (The nonlinearity of the system excludes full joint normality, of course.) Estimators for the loadings and for the correlations between the latent variables are consistent and asymptotically normal always (requiring only that the sample covariance matrix of the indicators has the asymptotic properties referred to). In section 3 we report illustrative Monte Carlo studies for a nonlinear structural model including an interaction term and two quadratic terms, and compare PLSc with LMS. In section 4 an empirical application is presented, using both PLSc and LMS again. Here we offer a computationally simple suggestion to test for the adequacy of the nonlinear model by a ‘chi-squared type of test’, based on the differences between jackknife pseudo-values for PLSc and LMS. We believe this suggestion to be new. In the final section, section 5, we summarize the results, discuss a number of issues that can be raised with regard to PLS’ modeling and estimation, and outline a research agenda. The appendix, that may be of independent interest as well, proves an extension of Stein’s identity that helps interpret estimated coefficients of (linear and) quadratic structural equations, when in fact the true \textit{equation is nonlinear of higher order} and the model employed in estimation is wrong: the estimators are consistent (CAN actually) for the average first and second order partial derivatives of the structural equations (here normality is crucial). So up to a point one could avoid the estimation of nonlinear models that go beyond the linear and quadratic model, and still get meaningful results.

\textbf{Consistent Partial Least Squares}

\textit{Weight Vectors}

A starting point for theoretical PLS analyses is what its founder, Herman Wold, called the ‘basic design’ (see, e.g., Wold, 1982 or Dijkstra, 2010). The basic design is in essence a structural equation model that captures the links between latent variables,
Consistent PLS

measurement errors and indicators. We focus first on the measurement model, the relationships between the latent variables can be left unspecified at this stage.

The vector of indicators, denoted by ‘$y$’, is assumed to possess finite moments of appropriate order (as many as are required for consistency and asymptotic normality of the estimators to be developed below). The word ‘basic’ refers to the clear cut character of the measurement model: the vector $y$ can be decomposed into subvectors of indicators, where each subvector measures or reflects a unique latent variable. It is assumed that there are at least two indicators per latent variable. If we denote a typical subvector of $y$ by $y_i$, we have

$$y_i = \lambda_i \cdot \eta_i + \epsilon_i$$

where the vector of factor loadings $\lambda_i$ and the vector of idiosyncratic (‘measurement’) errors $\epsilon_i$ have the same dimensions as $y_i$, and the latent variable $\eta_i$ is one-dimensional. For convenience the sufficient but by no means necessary assumption is made that all components of all error vectors are mutually independent, and independent of all latent variables. Indicators and latent variables will be assumed to have zero means and unit variances. The correlation between $\eta_i$ and $\eta_j$ will be denoted by $\rho_{ij}$.

A particular set of easy implications is that the covariance matrix $\Sigma_{ii}$ of $y_i$ can be written as:

$$\Sigma_{ii} := Ey_iy_i^\top = \lambda_i \lambda_i^\top + \Theta_i$$

where $\Theta_i$ is diagonal with non-negative diagonal elements. The covariance between $y_i$ and $y_j$ is given by:

$$\Sigma_{ij} := Ey_iy_j^\top = \rho_{ij} \lambda_i \lambda_j^\top.$$  

It is assumed as is customary in SEM estimation that we have a random sample, a set of independent vectors, each distributed as $y$. The sample counterparts of $\Sigma_{ii}$ and $\Sigma_{ij}$ are denoted by $S_{ii}$ and $S_{ij}$, respectively. Without repeatedly saying so it is assumed that the
sample data are standardized before being analyzed. So the observed data have zero mean and unit (sample) variance. Note that the assumptions made so far concerning the character of the sample and the existence of moments, and standard results from asymptotic theory, jointly entail that the sample counterparts are CAN estimators for the population covariance and correlation matrices.

PLS features a number of iterative fixed-point algorithms, of which we select one, the so-called ‘mode A algorithm’ (for a discussion of possible rationales for the PLS algorithms see, e.g., Dijkstra, 2010). Mode A is in general the numerically most stable mode. As a rule it converges, and it is usually very fast (four or five iterations, say, to get five significant decimals for the problems analyzed in this paper). The outcome is an estimated weight vector \( \hat{w} \), with typical subvector \( \hat{w}_i \) of the same dimensions as \( y_i \). With these weights sample proxies are defined for the latent variables: \( \hat{\eta}_i := \hat{w}_i^T y_i \) for \( \eta_i \), with the customary normalization of a unit sampling variance, so \( \hat{w}_i^T S_{ii} \hat{w}_i = 1 \). In Wold’s PLS approach the \( \hat{\eta}_i \)'s replace the unobserved latent variables, and loadings and structural parameters are estimated ‘directly’, as opposed to LISREL and other covariance based SEM programs that follow the opposite order. For mode A we have for each \( i \):

\[
\hat{w}_i \propto \sum_{j \in C(i)} sign_{ij} \cdot S_{ij} \hat{w}_j.
\]

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, the set \( C(i) \) contains only the indices of latent variables that are ‘adjacent’ to \( \eta_i \), i.e., only the indices of those latent variables that appear on the other side of the structural or path equations in which \( \eta_i \) appears. This is not always a good idea, particularly when the correlations between the indicators of \( \eta_i \) and the indicators of its adjacent variables are weak. One may consider to use in general all \( j \neq i \) in \( C(i) \). Here we followed Wold’s traditional approach, using ‘adjacency’.

Clearly, \( \hat{w}_i \) is obtained by a regression of the indicators \( y_i \) on the ‘sign-weighted sum’ of
the selected proxies: \( \sum_{j \in C(i)} \text{sign}_{ij} \cdot \hat{\eta}_j \). There are other versions (with correlation weights, e.g.), but this is one of the very simplest, and it is the original one (cf. Wold, 1982). There is little motivation in the PLS literature for the coefficients of \( S_{ij} \hat{w}_j \), but the particular choice can be shown to be irrelevant for the probability limits of the estimators. If we write (4) as the equation \( \hat{w} = F(\hat{w}, S) \), the PLS-algorithm solves the equation iteratively, by repeated substitutions. We take an essentially arbitrary starting vector \( w_0 \) and define the sequence \( \{w_k\} \) of vectors of the same dimension as \( \hat{w} \) for \( k \geq 1 \) by \( w_k = F(w_{k-1}, S) \). So the subscript \( k \) gives the number of iterations, it does not indicate a subvector. If the sequence converges, which in practice is assumed to have happened whenever two consecutive weight vectors differ nowhere more than a prescribed small number, \( \hat{w} \) is by definition the last vector calculated. One can speed things up a bit by inserting the updates as soon as they are available, but that is not really necessary. The function \( F(\ldots) \) varies with the mode (A, B, or C in PLS parlance). Whatever the mode, Dijkstra (1981, 1985, see also Dijkstra, 2010) has shown that \( F(\ldots, S) \) is a local contraction mapping of the weights when \( S \) is sufficiently close to \( \Sigma \). Since the probability of that event is arbitrarily close to one when the size of the random sample is sufficiently large, the modes will converge with a probability tending to one, for in fact essentially arbitrary starting vectors. For very special cases one can prove more: (essentially) global convergence instead of just local convergence. The latter is the rule of course also in SEM. The vector \( \hat{w} \) is a locally continuously differentiable function of the sample covariance (correlation) matrix \( S \). So \( \hat{w} \) and other estimators that depend smoothly on \( \hat{w} \) and \( S \) possess probability limits and are jointly asymptotically normal. There is essentially only one regularity condition to be satisfied: at the true values, for \( S \) equal to \( \Sigma \), the fixed point equations ought to have a solution, see equation (5) below. This excludes trivial, degenerate cases with some loading vectors equal to zero, or zero correlations between \( \eta_i \) and every \( \eta_j \) with \( j \in C(i) \).
Factor Loadings

Let us denote the probability limit of \( \hat{w}_i \), \( \text{plim} \hat{w}_i \), by \( \bar{w}_i \). We can get it from equation (4) by substitution of \( \Sigma \) for \( S \). So:

\[
\bar{w}_i \propto \sum_{j \in C(i)} \text{sign}_{ij} \cdot \Sigma_{ij} \bar{w}_j = \sum_{j \in C(i)} \text{sign}_{ij} \cdot \rho_{ij} \lambda_i \lambda_j^\top \bar{w}_j.
\] (5)

Since \( \lambda_j^\top \bar{w}_j \) is a scalar, and all terms in the sum have \( \lambda_i \) in common, it is clear that \( \bar{w}_i \propto \lambda_i \). Because of the normalization (unit variance) the proportionality constant has to be such that \( \bar{w}_i^\top \Sigma_{ii} \bar{w}_i = 1 \). This entails that

\[
\bar{w}_i = \lambda_i \div (\lambda_i^\top \Sigma_{ii} \lambda_i)^{\frac{1}{2}}. \tag{6}
\]

So \( \text{plim} \hat{w}_i \) is proportional to \( \lambda_i \). We propose here as in Dijkstra (1981, 1985, 2010, 2011) to take \( \hat{\lambda}_i := \hat{c}_i \cdot \hat{w}_i \), where the scalar \( \hat{c}_i \) is such that its probability limit is equal to \( (\lambda_i^\top \Sigma_{ii} \lambda_i)^{\frac{1}{2}} \), so that \( \hat{\lambda}_i \) is a consistent estimator for \( \lambda_i \). A simple choice that will do is

\[
\hat{c}_i := \left[ \frac{\hat{w}_i^\top (S_{ii} - \text{diag}(S_{ii})) \hat{w}_i}{\hat{w}_i^\top (\hat{w}_i \hat{w}_i^\top - \text{diag}(\hat{w}_i \hat{w}_i^\top)) \hat{w}_i} \right]^{\frac{1}{2}}. \tag{7}
\]

This is easy to verify: if one replaces \( S_{ii} \) by \( \Sigma_{ii} \) and \( \hat{w}_i \) by \( \bar{w}_i \) in (7), than the matrix in the denominator will equal the matrix in the numerator, apart from a factor \( 1 \div (\lambda_i^\top \Sigma_{ii} \lambda_i) \), so

\[
\bar{c}_i := \text{plim} \hat{c}_i = (\lambda_i^\top \Sigma_{ii} \lambda_i)^{\frac{1}{2}}. \tag{8}
\]

In small samples \( \hat{c}_i \) may not be well-defined, but for the samples studied in this paper and elsewhere that was never an issue. The particular choice was motivated by the consideration that the off-diagonal elements of \( S_{ii} \) consist of consistent estimators for the products of the loadings, and that therefore we could do well by choosing a scalar that would make a proper ‘distance’ between

\[
[S_{ii} - \text{diag}(S_{ii})] \quad \text{and} \quad [(c \cdot \hat{w}_i) (c \cdot \hat{w}_i)^\top - \text{diag}((c \cdot \hat{w}_i) (c \cdot \hat{w}_i)^\top)] \tag{9}
\]
as a function of the real number $c$ as small as possible. Minimization of the sum of squared differences leads to $\hat{c}_i$. Clearly other distances would be permissible as well, like a sum of weighted squared differences with weights reflecting the relative accuracy with which the elements of $S_{ii}$ estimate the products of the loadings. Note that we explicitly used the uncorrelatedness between all components of $\epsilon_i$. If this assumption is in doubt, one could delete the terms in the difference between the two matrices in equation (9) that correspond with possibly nonzero correlations.

**Correlations between Latent Variables**

It will be useful to define a population proxy $\eta_i$ by $\eta_i := \overline{w}_i^T y_i$. Since $\eta_i = (\overline{w}_i^T \lambda_i) \cdot \eta_i + \overline{w}_i^T \epsilon_i$, the squared correlation between a population proxy and its corresponding latent variable is

$$R^2(\eta_i, \eta_i) = (\overline{w}_i^T \lambda_i)^2.$$  

(10)

Inserting the righthand side of (6) for $\overline{w}_i$ and recalling (2) gives

$$R^2(\eta_i, \eta_i) = \frac{(\lambda_i^T \lambda_i)^2}{\lambda_i^T \Sigma_{ii} \lambda_i} = \frac{(\lambda_i^T \lambda_i)^2}{(\lambda_i^T \lambda_i)^2 + \lambda_i^T \Theta_i \lambda_i}.$$  

(11)

For a ‘large’ number of ‘high quality’ indicators this correlation will be close to one (‘consistency at large’ in PLS parlance). With

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

(12)

and (10) we deduce the trivial but important algebraic relationship:

$$R^2(\eta_i, \eta_j) = \rho_{ij}^2 \cdot R^2(\eta_i, \eta_i) \cdot R^2(\eta_j, \eta_j).$$  

(13)

This indicates 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
Consistent PLS 12

\[ R^2 (\eta_i, \bar{\eta}_i) = (\bar{w}_i^T \lambda_i)^2 = (\bar{w}_i^T \cdot (\bar{w}_i \cdot \bar{\tau}_i))^2 = (\bar{w}_i^T \bar{w}_i)^2 \cdot \bar{e}_i^2 \]  

(14)

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

\[ \hat{R}^2 (\eta_i, \bar{\eta}_i) := (\hat{w}_i^T \hat{w}_i)^2 \cdot \hat{e}_i^2. \]  

(15)

Moreover, with

\[ R^2 (\eta_i, \eta_j) := (\hat{w}_i^T S_{ij} \hat{w}_j)^2 \]  

(16)

we can estimate the correlations between the latent variables consistently by ‘inverting’ (13):

\[ \hat{\rho}_{ij}^2 := \frac{R^2 (\eta_i, \eta_j)}{R^2 (\eta_i, \bar{\eta}_i) \cdot R^2 (\eta_j, \bar{\eta}_j)}. \]  

(17)

So \( E \eta_i \eta_j \) is estimated by the sample covariance between the proxies \( \hat{\eta}_i \) and \( \hat{\eta}_j \), each divided by its estimated quality.

We note in passing that standard PLS software for mode A will produce all the necessary ingredients for consistent estimation, all it takes is a simple rescaling of the weight vectors. It may also be worth noting that a similar approach can be developed for mode B (and C). But first results (not reported here, see Dijkstra, 2011) indicate that the adjusted mode B is numerically much less stable, and convergence is a lot less fast. This is probably as one would expect, since mode B is based on multiple regressions. So it seems to make sense to focus on mode A first in developing PLSc.

**Interactions between Latent Variables**

Now consider a system of recursive equations, say \( q \) equations in \( p + q \) latent variables, where the \( i^{th} \) equation reads:

\[ \eta_{p+i} = L_i (\eta_{1:p+i-1}) + I_i (\eta_{1:p+i-1}) + \zeta_i. \]  

(18)

Here \( L_i (\eta_{1:p+i-1}) \) is a linear function of the preceding \( p + i - 1 \) latent variables, and \( I_i (\eta_{1:p+i-1}) \) is a linear function of their centered cross-products, \( \eta_k \eta_l - \rho_{kl} \). The residual
ζᵢ is uncorrelated with the latent variables and their cross-products on the righthand side of its defining equation (or more simply, independent of η₁₂,...,ηᵢ₋₁), so every equation is a regression equation. The coefficients are therefore determined by the second order moments of the explanatory variables on the righthand side, and their covariances with the dependent variable on the lefthand side. To fix ideas, take the first equation and assume p = 3 (this will in fact cover the general case).

\[ η₄ = γ₁η₁ + γ₂η₂ + γ₃η₃ + γ₁₂(η₁η₂ - ρ₁₂) + γ₁₃(η₁η₃ - ρ₁₃) + γ₂₃(η₂η₃ - ρ₂₃) + ζ₁. \]  (19)

The upper triangular part of the covariance matrix of the explanatory variables equals:

\[
\begin{bmatrix}
1 & ρ₁₂ & Eη₁² \eta₂ & Eη₁² \eta₃ & Eη₁₂ \eta₃ \\
1 & ρ₂₃ & Eη₁² \eta₂ & Eη₁₂ \eta₃ & Eη₂² \eta₃ \\
1 & & Eη₁₂ \eta₃ & Eη₂² \eta₃ & Eη₂² \eta₃ \\
& & & Eη₁² \eta₂ - ρ₁₂ & Eη₂² \eta₃ - ρ₁₂ρ₂₃ \\
& & & & Eη₁² \eta₃ - ρ₁₃ \\
& & & & & Eη₂² \eta₃ - ρ₂₃
\end{bmatrix}
\]  (20)

We already know how to estimate the correlations. Using \( \overline{η}_i = Q_i \cdot η_i + δ_i \) with \( Q_i := R(\overline{η}_i, η_i) \) and \( δ_i := \overline{w}_i^T e_i \), where the δ’s are mutually independent and independent of the η’s one easily derives

\[
Eη_i² η_j = Q_i² Q_j \cdot Eη_i² η_j
\]  (21)
\[
Eη_i η_j η_k = Q_i Q_j Q_k \cdot Eη_i η_j η_k
\]  (22)
\[
Eη_i² η_j² = Q_i² Q_j² \cdot (Eη_i² η_j² - 1) + 1
\]  (23)
\[
Eη_i² η_j η_k = Q_i² Q_j Q_k \cdot Eη_i² η_j η_k + ρ_jk Q_j Q_k (1 - Q_i²)
\]  (24)

where different subscripts are different indices. Equations (23) and (24) clearly signal that one cannot generally replace latent variables by estimated proxies divided by their estimated quality. Inserting for the Q’s their consistent estimates, based on (15),
estimating the lefthand side by their obvious sample counterparts, and solving for the
moments of the latent variables, will give a consistent estimate for the covariance matrix.
We also need consistent estimates for the covariances between the dependent variable and
the explanatory variables:

\[ [E\eta_4\eta_1, E\eta_4\eta_2, E\eta_4\eta_3, E\eta_4\eta_1\eta_2, E\eta_4\eta_1\eta_3, E\eta_4\eta_2\eta_3] . \]  

(25)

They are obtained simply by taking sample averages of corresponding products of
estimated proxies, each divided by its estimated quality (\( \hat{Q} \)).

So the \( \gamma \)'s and the equation’s R-square of course can be estimated consistently,
using only estimates for the quality of the proxies and their observed values (the latter
aspect distinguishes this model from the linear model, where one needs to store only the
second order moments, here we have to hold on to the raw data). Since the system is
recursive, every equation can be handled in the same way. Apart from certain
independence assumptions, which can also be weakened somewhat, no distributional
assumptions are needed beyond the existence of moments.

With the same modest proviso, it is important to note that the vector of statistics
\( [\hat{w}; \hat{\lambda}; \hat{\rho}; \hat{\gamma}] \) is also asymptotically jointly normal, around the true (population) values: it
is a (locally) smooth function of sample moments of the indicators plus a term that is in
probability of smaller order than \( 1/\sqrt{n} \) (\( n \) is the number of observations). For \( [\hat{w}; \hat{\lambda}; \hat{\rho}] \)
this is clear because of the joint asymptotic normality of \( [\hat{w}; \text{vec}(S)] \) and the smooth
dependence of \( \hat{\rho} \) on this vector. One can verify that the vector of sample averages of all
terms like \( \hat{\eta}_i^2\hat{\eta}_j, \hat{\eta}_i\hat{\eta}_j\hat{\eta}_k, \hat{\eta}_i^2\hat{\eta}_j^2 \), and \( \hat{\eta}_i^2\hat{\eta}_j\hat{\eta}_k \) is also locally a smooth function of the sample
moments of the indicators plus \( o_p(1/\sqrt{n}) \). And, clearly, \( \hat{\gamma} \) is a smooth function of all
statistics involved. An application of the Cramér-Wold device completes the argument.
The result concerning \( [\hat{w}; \hat{\lambda}; \hat{\rho}; \hat{\gamma}] \) is driven entirely by the asymptotic normality of the
sample moments of the vector of indicators; the indicator’s distribution itself is
(asymptotically) irrelevant.

**Quadratic and Higher Order Models**

With added squares, we will see that we cannot avoid making additional distributional assumptions. So consider a system of recursive equations, say $q$ equations in $p + q$ latent variables, where the $i^{th}$ equation reads:

$$
\eta_{p+i} = L_i (\eta_{1:p+i-1}) + I_i (\eta_{1:p+i-1}) + S q_i (\eta_{1:p+i-1}) + \zeta_i
$$

(26)

and $S q_i (\eta_{1:p+i-1})$ is a linear function of the centered squares, $\eta^2 - 1$, of the preceding $p + i - 1$ latent variables. As before, $\zeta_i$ allows of a regression interpretation of each equation.

Focus on the first equation. It may be verified that the covariances between the dependent variable and the explanatory variables can be estimated as before (by sample averages of corresponding products of powers of estimated proxies, each divided by its estimated quality). The previous section also allows one to estimate most of the entries in the covariance matrix of the explanatory variables consistently, except $E \eta_3^i \eta_j$ (for different $i$ and $j$), $E \eta_3^i$, and $E \eta_4^i$. We have (as before with $\eta_i = Q_i \cdot \eta_i + \delta_i$, $Q_i := R (\eta_i, \eta_i)$ and $\delta_i := w_i \epsilon_i$, where the $\delta$'s are mutually independent and independent of the $\eta$'s):

$$
E \eta_i^3 \eta_j = Q_i^3 Q_j E \eta_i^3 \eta_j + 3 E \eta_i \eta_j (1 - Q_i^2),
$$

(27)

so $E \eta_i^3 \eta_j$ can be estimated consistently. The third and the fourth order moments are spoilsports:

$$
E \eta_i^3 = Q_i^3 E \eta_i^3 + E \delta_i^3
$$

(28)

$$
E \eta_i^4 = Q_i^4 E \eta_i^4 + 6 Q_i^2 (1 - Q_i^2) + E \delta_i^4.
$$

(29)

Clearly we need assumptions about or knowledge of the distribution of the measurement errors that go beyond a zero mean, the existence of moments and independence. The
‘natural’ assumption of ‘symmetry’ of the elements of the $\epsilon_i$’s would only come halfway; even though now $E\delta_i^2 = 0$, the kurtosis is not fixed. We could select the value for the kurtosis of the normal distribution without, however, fully adopting multivariate normality for $\epsilon$. Then, $E\delta_i^4 = 3(E\delta_i^2)^2 = 3(Q_i^2)^2$, and we have consistent estimation of the relevant moments for the first equation, and therefore of its regression parameters, provided of course that the assumed value of the kurtosis is correct. For the estimation of the entire system it will be convenient however to adopt normality here, and so we will. More precisely, we will work with joint normality of $[\eta(1:p); \zeta; \epsilon]$.

Under normality, and still focusing on the first equation, the covariance matrix of the explanatory variables can be expressed explicitly in terms of the correlations between the elements of $\eta(1:p)$. The covariances of the latter with the cross-products and squares are zero since

$$
E\eta_i^2 \eta_j = E\eta_i^2 (\rho_{ij} \eta_i + \zeta_{j,i}) = \rho_{ij} E\eta_i^3 + E\eta_i^2 E\zeta_{j,i} = 0
$$

(30)

and

$$
E\eta_i \eta_j \eta_k = E\eta_i \eta_j (\beta_{ki} \eta_i + \beta_{kj} \eta_j + \zeta_{k,ij})
= \beta_{ki} E\eta_i^2 \eta_j + \beta_{kj} E\eta_i \eta_j^2 + E\eta_i \eta_j E\zeta_{k,ij} = 0.
$$

(31)

In other words, the covariance matrix of the explanatory variables for the first equation is block-diagonal. The remaining elements corresponding with the squares and the cross-products can all be calculated using:

$$
E\eta_i^2 \eta_j^2 = 1 + 2\rho_{ij}^2
$$

(32)

$$
E\eta_i^3 \eta_j = 3\rho_{ij}
$$

(33)

$$
E\eta_i^2 \eta_j \eta_k = \rho_{jk} + \rho_{ij}\rho_{ik}.
$$

(34)

For concreteness’ sake we will specify the upper triangular part of the covariance matrix
for $p = 3$, in lexicographic order \{linear, cross-products, squares\}:

\[
\begin{pmatrix}
1 & \rho_{12} & \rho_{13} & 0 & 0 & 0 & 0 & 0 \\
1 & \rho_{23} & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 + \rho_{12}^2 & \rho_{12}\rho_{13} + \rho_{23} & \rho_{12}\rho_{23} + \rho_{13} & 2\rho_{12} & 2\rho_{12} & 2\rho_{13}\rho_{23} & 1 + \rho_{12}^2 & \rho_{12}\rho_{13} + \rho_{23} & \rho_{12} & 2\rho_{12}\rho_{23} & 2\rho_{13} & 2\rho_{13}^2 & 2 & 2\rho_{13}^2 & 2 & 2\rho_{23} & 2 \\
1 + \rho_{13}^2 & \rho_{13}\rho_{23} + \rho_{12} & 2\rho_{13} & 2\rho_{12}\rho_{23} & 2\rho_{23} & 2\rho_{23} & 2 & 2\rho_{13}^2 & 2 & 2\rho_{13}^2 & 2 & 2\rho_{23} & 2 & 2 & 2 & 2 \\
1 + \rho_{23}^2 & 2\rho_{12}\rho_{13} & 2\rho_{13} & 2\rho_{23} & 2\rho_{23} & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\

\end{pmatrix}
\]  
\( (35) \)

All other cases are direct extensions. Consistent estimation of the regression parameters of equations like

\[
\begin{align*}
\eta_4 & = \gamma_1 \eta_1 + \gamma_2 \eta_2 + \gamma_3 \eta_3 + \\
& \quad \gamma_{12} (\eta_1 \eta_2 - \rho_{12}) + \gamma_{13} (\eta_1 \eta_3 - \rho_{13}) + \gamma_{23} (\eta_2 \eta_3 - \rho_{23}) + \\
& \quad \gamma_{11} (\eta_1^2 - 1) + \gamma_{22} (\eta_2^2 - 1) + \gamma_{33} (\eta_3^2 - 1) + \zeta_1
\end{align*}
\]  
\( (36) \)

is straightforward.

Adding equations recursively presents \textit{in principle} no problem either: for each new equation for $\eta_{p+i}$, we can get an estimate for the covariance matrix of the explanatory variables by substitution of the estimated equation for $\eta_{p+i-1}$, expressing the relevant covariances in terms of estimated correlations of the first $p$ latent variables and the estimated regression parameters (block-diagonality will no longer apply). The words \textit{in principle} ought to be stressed: each additional variable and substitution raises the order of the polynomials in $\eta$ by the power 2, and calculation of moments quickly becomes challenging, \textit{even} under normality. A recent article by Kan (2008) provides the most efficient algorithms available, so it will be feasible up to a point, but one would guess that
statistical stability will quickly be a serious issue.

Adding polynomials with higher powers like $\eta_1^3$, $\eta_2^3$, $\eta_1^2\eta_2$, $\eta_1\eta_2^2$, and $\eta_1\eta_2\eta_3$ and so on, will again in principle present no problem with normality, even though one may in practical situations want to reflect upon its wisdom. In this case more work needs to be done for the covariances between the dependent variable and the explanatory variables.

For different numbers $i, j, k$ less than $p+1$ and $m$ larger than $p$ we find

$$E\eta^3_i \eta_m = Q^3_i Q_m E\eta^3_i \eta_m + 3E\eta_i \eta_m (1 - Q^2_i) \quad (37)$$

$$E\eta^2_i \eta_j \eta_m = Q^2_i Q_j Q_m E\eta^2_i \eta_j \eta_m + E\eta_j \eta_m (1 - Q^2_i). \quad (38)$$

We still have: $E\eta_i \eta_j \eta_k \eta_m = Q_i Q_j Q_k Q_m E\eta_i \eta_j \eta_k \eta_m$. And one can proceed as before.

An Observation on Nonlinear Functions and Normality.

What if one knows or suspects that $\eta_{p+1} = f(\eta_{(1:p)}) + \zeta$, where the structural function $f(.)$ cannot be approximated well by a finite polynomial, or if it can, the degree of the approximating polynomial is too high to use nonlinear estimation methods with confidence? Then if $\eta_{(1:p)}$ is normal one could contemplate to use a linear or quadratic specification, regardless of the true nature of $f(.)$, and still get sensible estimates. That is to say, when one uses a linear model, the coefficients are average first order partial derivatives of $f(.)$. A quadratic specification would yield in addition average second order partial derivatives. This observation is based on Stein’s identity for first order partial derivatives; the appendix recalls the identity and extends it to second order partial derivatives. It is a pity that Stein’s identity characterizes the normal distribution: for other distributions it is at best an approximation.

Monte Carlo Study

Here we present some illustrative Monte Carlo simulation results, pitting PLSc against LMS. We would like to check whether both methods are correct on the average. It
will also be interesting to see how accurate PLSc is relative to LMS in terms of standard deviations. Since, in this first study, the setup favors LMS because of normality of all of those variables that can be normal, one will expect PLSc to be less accurate, but the size of the difference is not clear beforehand.

The model we will use is a nonlinear latent structural equation model that includes an interaction effect as well as two quadratic effects. All observed variables as well as the latent variables are standardized. The following structural equation is used:

\[ \eta_3 = \gamma_1 \eta_1 + \gamma_2 \eta_2 + \gamma_{12} (\eta_1 \eta_2 - \rho_{12}) + \gamma_{11} (\eta_1^2 - 1) + \gamma_{22} (\eta_2^2 - 1) + \zeta. \]  

(39)

Values of the structural parameters are given by: \( \gamma_1 = .50, \gamma_2 = - .30, \gamma_{12} = - .20, \gamma_{11} = .10, \gamma_{22} = -.15, \) and the correlation between \( \eta_1 \) and \( \eta_2 \) is \( - .30 \). For the values chosen the variance of \( \zeta \) is equal to \(.4788\). We use the measurement equations \( y_i = \lambda_i \cdot \eta_i + \epsilon_i \) where each \( y_i \) is of order 3 by 1, and all factor loadings are equal to \(.80\). All components of all error vectors are mutually independent, and independent of all latent variables (and therefore also of \( \zeta \)). We take the vector containing \( \eta_1, \eta_2, \zeta, \epsilon_1, \epsilon_2, \) and \( \epsilon_3 \) to be jointly normal. Generating values for those variables is straightforward, and with \( \eta_1, \eta_2, \) and \( \zeta \) we get a value for \( \eta_3 \) (which is of course nonnormally distributed); values for the indicators are now immediate.

To test PLSc, 10,000 samples of size \( n = 400 \) were generated. In order to compare LMS with PLSc, additionally 500 samples of size \( n = 400 \) had to be generated, because LMS is not able to estimate 10,000 samples within a reasonable time (LMS in Mplus uses numerical integration which is very time-consuming because the integration must be done at each iteration). LMS estimates were standardized as with PLSc. Monte Carlo mean estimates of structural parameters and their standard deviations are listed in Table 1.

The results show that both approaches perform about equally well with regard to the estimated means of the structural parameters, while standard deviations of LMS tend
to be smaller than those of PLSc. The loading estimates of PLSc, not shown, are on average very close to the true values, with a very slight tendency to underestimation. The corresponding standard errors are 0.043 for the loadings on $\eta_2$ and $\eta_3$, and 0.059 for the loadings on $\eta_2$. For LMS, the loading estimates are very close to the true values, and the corresponding standard errors are 0.051 for the loadings on $\eta_1$, $\eta_2$, and $\eta_3$.

— Table 1 about here —

We ran a number of additional Monte Carlo studies for PLSc, using $n = 100, 200, 400, 800, 1600$, and $n = 3200$, while keeping the same model and the same parameter values (and 10,000 Monte Carlo samples for each setup). The unbiasedness properties are the same for all sample sizes, with the bias of the estimator for the inner $R^2$ steadily decreasing (it is virtually unbiased for the larger sample sizes). When the sample size is doubled, the standard errors (Monte Carlo standard deviations) decrease: on average they are multiplied by 0.66, 0.69, 0.70, 0.70 and 0.71 when going from $n = 100$ to $n = 200, 400, 800, 1600$, and $n = 3200$. Roughly, doubling the sample size multiplies the standard error by $1/\sqrt{2}$, so the nonlinearity has a mild effect on the accuracy. Also note that the calculation time increased but much less faster than the sample size: the calculations for $n = 800$ took just 1.5 times as long as those for $n = 100$ (and for $n = 3200$ it took 2.6 times as long).

For reasons of comparison, we also ran Monte Carlo studies for LMS, but using only $n = 100, 200, 400$, and $n = 800$ because of time limits, while keeping the same model and the same parameter values (and 500 Monte Carlo samples for each setup). When the sample size is doubled, the size of the standard errors also roughly decreases by the factor $1/\sqrt{2}$. However, in contrast to PLSc, the calculations for $n = 800$ took 15.3 times as long as those for $n = 100$.

One issue that ought to be mentioned here is the ’numerical scope’ of PLSc as
opposed to LMS. The latter uses a highly sophisticated routine for the multiple integrals in its optimization, that may limit its applicability for 'larger' models. We report here a simulation study for the same model as before, except that we now had 6 indicators for each of the exogenous latent variables, and 12 for the endogenous latent variable. PLSc took about 79 seconds on a slow machine (4CPU 2.40 GHz; RAM 512 MB) to generate and estimate 10,000 samples of size 400. The unbiasedness properties are as before, and the accuracy is even higher (see Table 2). For LMS this Monte Carlo study could not be performed, as the model turned out to be too complex: after 45 minutes 5 samples had been analyzed, only one had converged and had produced estimates.

— Table 2 about here —

In empirical applications one will want of course standard errors for the estimators. LMS uses an approximation based on the second order derivatives of the likelihood function. For PLSc the bootstrap will be quite feasible, thanks to its computational efficiency. So we ran also a bootstrap analysis, with 500 samples of size 400 and 500 bootstrap replications. The ensuing standard errors (not reported here) are close to unbiased and relatively stable (the standard errors of the bootstrap standard errors are about 1/5 times or smaller than their mean).

The results of the Monte Carlo studies indicate that PLSc is able to estimate the parameters of the nonlinear structural equation model without bias, as LMS, but PLSc estimates the nonlinear part less accurately. The relatively higher accuracy of LMS is as one would expect: with a correct specification of the structural equation, and a distribution that agrees exactly with the one assumed by LMS, maximum likelihood ‘ought’ to do better. However, as our simulation studies also showed, PLSc has an advantage over LMS when more complex models with a large number of indicators or models with large sample sizes are being analyzed.
An Empirical Application

In the following a small empirical application is presented, comparing PLSc again with LMS. The structural equation includes a latent interaction term. Since the indicators are discrete variables, LMS will not necessarily calculate proper (approximations to) standard errors. We decided to use the (delete-1) jackknife for this purpose, even though it tends to be a bit conservative. But jackknifing is computationally less demanding than the bootstrap, which may be an issue for LMS. An added advantage of using sample re-use methods like jackknife and bootstrap is that the ensuing standard errors are ‘distribution-free’, making the comparison more fair.

The dataset used was put together and analyzed by Chin et al. (2003)\(^6\). It consists of observations on twelve indicators for each of \(n = 250\) individuals. The indicators are coded answers to questions that can be partitioned into three groups. All indicators except one are on a seven point scale (from one to seven), the exception requires eleven points (from zero to ten) (see Chin et al., 2003 for the details). We recoded one of the indicators (‘ENJ2’) to align the scores with the emotional value, consistently with the other indicators (score \(s\) is replaced by \(8 - s\)). Nonnormality is patently obvious. The first group of six indicators is related to the perceived usefulness (\(\eta_1\)) of a particular IT-application, the second group of three to its enjoyment (\(\eta_2\)), and the last group of three to the individuals’ intention-to-use (\(\eta_3\)) the application. The original research question concerned the existence of an interaction between perceived usefulness and enjoyment as related to the intention-to-use.

The structural model that is entertained is just:

\[
\eta_3 = \gamma_1 \eta_1 + \gamma_2 \eta_2 + \gamma_{12} (\eta_1 \eta_2 - \rho_{12}) + \zeta. \tag{40}
\]

PLSc and LMS gave a similar picture of the quality of the indicators (loadings as well as standard errors), therefore only results concerning the structural equation are reported.
(see Table 3). The estimates for PLSc (denoted by ‘Est.’) were obtained in 4 iterations for a stopping criterion of 0.000001. The estimated quality of the proxies was quite high: the squared correlations between the latent variables and corresponding proxies were .9319, .9044, and .9972 for $\eta_1$, $\eta_2$, and $\eta_3$, respectively. The calculations of the jackknife estimates for the standard errors (denoted by ‘J-SE’), that required 250 additional runs of the PLSc-algorithm, took 1.77 seconds on a slow machine (4CPU 2.40 GHz; RAM 512 MB) without optimizing for efficiency. The estimates for LMS were obtained in 44 iterations, the calculation of the jackknife estimates for the standard errors took 1.13 hours on an averagely equipped PC (4CPU 3.00 GHz; RAM 3.00 GB).

— Table 3 about here —

The results show (see Table 3) that interaction is present and it is clearly negative: The lower the enjoyment, the higher the impact of an increase in usefulness on the intention to use the application (for alternative analyses with an identical conclusion based on extensions of canonical variables and principal components, cf. Chin et al., 2003, Dijkstra & Henseler, 2011, and Dijkstra, 2009). All alternatives considered yield the same picture: perceived usefulness interacts negatively with enjoyment. The results allow of the same interpretation. Nevertheless, there are notable differences in the size of some of the parameters.

Here we suggest to use the jackknife pseudo-values to test the model’s validity. The idea is simply that when the model is correct, different but consistent estimators like LMS and PLSc are likely to be close in sufficiently large samples, but when the model is incorrect, we expect LMS and PLSc to tend to different points in the parameter space. The two approaches are so different in nature, one ‘solving moment equations’ and the other ‘maximizing a highly complex likelihood function with multiple integrals’, that
similar results with a *wrong* model would be surprising indeed (it will depend of course on the ‘size’ and the ‘direction’ of the misspecification, something that is in need of further investigation). So a ‘large difference’ between estimates consistent under the hypothesized model is ‘strong evidence’ against it. If we let $D$ stand for the 250 by 5 matrix of the differences in pseudo-values, and $d$ for the row-vector of its mean, we treat $250 \cdot d \cdot (\text{cov} (D))^{-1} \cdot d^\top$ as an observation on a $\chi^2 (5)$-variable. Its value is 20.33 with an associated probability value of a mere 0.11%. One may be inclined to reject the hypothesis that LMS and PLSc have the same probability limits and therefore reject the model (note that *equality* of probability limits does not prove, strictly speaking, that the model is correct). For the individual parameters one can use the separate $t$-tests which yield significant results at conventional levels for the difference between the interaction coefficients and the R-squares (absolute $t$-values of 2.01 and 2.53 resp., trimmed values not reported are much more outspoken). We could have chosen other parameters to look at instead of the structural parameters, but they seemed to be of obvious and central interest. It is well-known that the $\chi^2$ test can be much more powerful than the individual $t$-tests.

The heuristic nature of the ‘Jackknife-model-test’ will be evident, but theoretical underpinnings of the treatment of pseudo-values as i.i.d. variables used here can be found, e.g., in Gray and Schucany (1972) or Shao and Tu (1995). We intend to do more work on this test.

**Conclusion**

There are basically two ways of looking at PLS. One view is that PLS is a generalization of principal components and canonical variables, as maintained at various occasions by Herman Wold (see any of his references). Its focus was essentially prediction, using least squares criteria for fitting and evaluation (for some discussions and extensions in this vein see, e.g., Dijkstra & Henseler, 2011, and Dijkstra, 2009). This view, that
dominated the field from the eighties of the previous century till now, has led to many fruitful developments, as testified by the Handbook of Partial Least Squares by Esposito Vinzi et al. (2010). The alternative view, which was also maintained at times by Herman Wold, is that it is a distribution-free way to estimate structural models. The ‘basic design’, inspired by LISREL, was set up to facilitate discussion and analyses. This second view is in our opinion less tenable or productive. Its case rests mainly upon ‘consistency at large’, the assumption that the quality of the proxies is sufficiently high\(^7\).

Generally, as an estimation method for structural models, linear or nonlinear, we contend that PLS is not well-calibrated: it cannot reproduce the true parameters when supplied with the population or true distribution. A complicating factor is that the central value or probability limit of its estimators varies with the mode chosen, which may burden the interpretation of the estimation results with issues unrelated to content. So our proposition is that if one wants to use PLS as a computationally efficient and potentially robust alternative to any of the well-established estimation methods for structural models, one would need to consider corrections of the type we propose in this paper. The development of PLSc is not complete of course. For the basic design we made a good start, but it is essential that more general measurement models can be accommodated. In addition, more work should be done on mode B, and perhaps other generalizations of canonical variables. From a practical point of view one would want to see many more empirical tests and Monte Carlo simulations. From a theoretical point of view there seem to be a number of challenges and issues that will be of relevance to the mainstream approaches as well. We close this section with a list of the more ‘pressing’ issues:

- The extension to non-recursive nonlinear models. There is no problem for non-recursive linear models (see Dijkstra & Henseler, 2012), but nonlinearity and simultaneity may prove to be a challenge.

- Models, where the nonlinearity is due to interaction terms only, do not require
specific distributional assumptions. But squared and higher order terms appear to be rather more demanding. Here we chose to work with the normal distribution. The moment equations that this choice produces are probably the simplest possible, but that by itself is no justification in empirical work. How critical is normality? In linear models the effect of nonnormality is frequently tested using the Fleishman, Vale, and Maurelli (FVM) approach (see Fleishman, 1978, and Vale & Maurelli, 1983), where latent variables are replaced by well-chosen polynomials of the third degree in normal variables. But this will not work for nonlinear models without substantial adjustments: The traditional FVM approach leads to a distortion of the relationships between the latent variables so one does not estimate the same parameters. Polynomials of a much higher degree are required to neutralize this. Headrick (2010, in particular chapter 4), has all the necessary tools, but it still has to be implemented.

- If nonnormality is an issue, and one cannot help but feel that it is, is there a way around it that combines empirical relevance with computational efficiency? Ideally, it would integrate smoothly with the first point as well.

- We suggested to use the jackknife pseudo-values of alternative estimators, consistent under the hypothesized (nonlinear) model, for model testing. (A bootstrap version is a possibility also, but depending on the sample size it may be computationally too demanding.) Do we get proper $p$-values for empirically relevant sample sizes? To have some power, the estimators would have to be sensitive in different ways to misspecifications. Are there (possibly model-dependent) rules for choosing them?

- Due to its ‘partial information’ nature, PLSc may be more robust against misspecifications in the structural equations than ‘full information’ methods. It would be of interest to find out to what extent this is borne out in experiments, and one would also like to have some insight into the trade-off between robustness and efficiency.
Acknowledgement

The authors wish to acknowledge the constructive comments and suggestions by the editor, an associate editor, and three reviewers, that led to substantial improvements in contents and readability.
References


structural equation modeling: is partial least squares an alternative?


Appendix

Stein’s Identity

The normal distribution has many interesting (and characterizing) properties. One of them is the property known as Stein’s identity. This says that when $X$ is a $p$-dimensional normal vector with mean $\mu$ and covariance matrix $\Omega$, then $f(X)$, where $f(.)$ is a real valued smooth function that satisfies (cf. Liu, 1994):

$$\text{cov}(X, f(X)) = \Omega E[\nabla f(X)].$$

(41)

Here $\nabla f(.)$ is $f(.)$’s gradient. So when $\Omega$ is invertible, the regression of $f(X)$ on $X$ equals:

$$f(X) = Ef(X) + E[\nabla f(X)]^\top \cdot (X - \mu) + \text{regression residual}.$$  

(42)

Consequently, the regression coefficients are the average first partial derivatives. We add here an extension of Stein’s identity that covers the quadratic case:

$$f(X) = Ef(X) + E[\nabla f(X)]^\top \cdot (X - \mu) + \frac{1}{2} \cdot \text{trace} [E(H(X)) \cdot ((X - \mu)(X - \mu)^\top - \Omega)] + \nu$$

(43)

where $H(.)$ is the Hessian of $f(.)$ and $\nu$ is the regression residual. So the regression coefficients for the interaction and quadratic terms (apart from the multiple $\frac{1}{2}$) are average second order partial derivatives. They are estimated consistently when a quadratic specification is used instead of the true nonlinear function. (There does not appear to be a comparably neat expression for higher order partial derivatives).

We will prove (43), assuming the existence of partial derivatives and boundedness of their expected absolute values.

Write $X = \mu + \Omega^{\frac{1}{2}}Z$, where $Z$ is $p$-dimensional standard normal. Consider first, for a real function $g(.)$ from the same class as $f(.)$, the regression of $g(Z) - Eg(Z)$ on $Z$ and the squares and cross-products of its components. The covariance matrix of the regressors
is diagonal, with ones everywhere, except at the entries corresponding with the squares where we have 2. So the regression coefficient of $Z_i$ equals $EZ_i g(Z) = E \nabla g(Z)_i$ by Stein’s identity. For twice the coefficient of $Z_i^2 - 1$ we get:

$$E (Z_i^2 - 1) g(Z) = EZ_i^2 g(Z) - Eg(Z) = (44)$$

$$EZ_i (Z_i g(Z)) - Eg(Z) = E (g(Z) + Z_i \nabla g(Z)_i)) - Eg(Z) = (45)$$

$$E (Z_i \nabla g(Z)_i) = EH_{ii}(Z).$$

On the second line we applied Stein’s identity to $Z_i g(Z)$ and on the third line to $\nabla g(Z)_i$.

One obtains similarly for the coefficient of a cross-product ($i \neq j$):

$$EZ_i (Z_j g(Z)) = E (Z_j \nabla g(Z)_i) = EH_{ij}(Z).$$

Collecting terms yields (with $H$ subscripted by $g$ for identification):

$$g(Z) = Eg(Z) + E \nabla g(Z)^T \cdot Z + 1/2 \cdot \text{trace} |EH_g(Z) \cdot (ZZ^T - I)| + \nu.$$  

Finally take a smooth function $f(.)$ of $X$, and define $g(Z) := f \left( \mu + \Omega^{\frac{1}{2}} Z \right)$. A substitution of $Z = \Omega^{-\frac{1}{2}} (X - \mu)$, $\nabla g(Z) = \Omega^{\frac{1}{2}} \nabla f(X)$, and $H_g(Z) = \Omega^{\frac{1}{2}} H_f(X) \Omega^{\frac{1}{2}}$ into (48) yields the desired expression for general $f(X)$. 

Footnotes

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

2 Things like the sample average of the product \( \hat{\eta}_i \hat{\eta}_j \hat{\eta}_k \) estimate \( E \eta_i \eta_j \eta_k \) consistently as a little algebra and a routine application of asymptotic theory will easily show. The established consistency of \( \hat{\eta} \) and of the sample moments of the indicators is sufficient.

3 The colon indicates stacking, as in Matlab.

4 \( \zeta_{j,i} \) is the residual that is left after regressing \( \eta_j \) on \( \eta_i \). It is independent of \( \eta_i \) and has mean zero. Also \( \eta_k = \beta_{ki} \eta_i + \beta_{kj} \eta_j + \zeta_{k,ij} \) is the regression of \( \eta_k \) on \( \eta_i \) and \( \eta_j \); the residual \( \zeta_{k,ij} \) is independent of the regressors and has mean zero.

5 Only (34) requires some work:

\[
E \eta_i^2 \eta_j \eta_k = E \eta_i^2 \eta_j (\beta_{ki} \eta_i + \beta_{kj} \eta_j + \zeta_{k,ij}) = \beta_{ki} E \eta_i^3 \eta_j + \beta_{kj} E \eta_i^2 \eta_j^2 = 3 \beta_{ki} \rho_{ij} + \beta_{kj} \left(1 + 2 \rho_{ij}^2\right).
\]

Inserting \( (\rho_{ik} - \rho_{ij} \rho_{jk}) \div (1 - \rho_{ij}^2) \) for \( \beta_{ki} \) and an analogous expression for \( \beta_{kj} \) yields (34).

6 We gratefully thank Wynne W. Chin from the Department of Decision and Information Sciences, Bauer College of Business, University of Houston, Texas, USA, for providing the empirical data.

7 Matlab code, that exemplifies how sensitive the PLS structural coefficients can be for less than perfect proxies (in linear models) is available from the first author.

8 This happens, e.g., in Klein and Muthén (2007), see in particular pp. 660 and 661.
Table 1

Comparison of PLSc with LMS: Mean Parameter Estimates (Mean) and Standard Deviations (SD) of the Monte Carlo Study (Model with 3 Indicator Variables measuring each Latent Variable)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>PLSc Mean</th>
<th>PLSc SD</th>
<th>LMS Mean</th>
<th>LMS SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$</td>
<td>0.500</td>
<td>0.500</td>
<td>0.050</td>
<td>0.500</td>
<td>0.047</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>-0.300</td>
<td>-0.300</td>
<td>0.055</td>
<td>-0.297</td>
<td>0.052</td>
</tr>
<tr>
<td>$\gamma_{12}$</td>
<td>-0.200</td>
<td>-0.198</td>
<td>0.086</td>
<td>-0.203</td>
<td>0.063</td>
</tr>
<tr>
<td>$\gamma_{11}$</td>
<td>0.100</td>
<td>0.099</td>
<td>0.061</td>
<td>0.100</td>
<td>0.040</td>
</tr>
<tr>
<td>$\gamma_{22}$</td>
<td>-0.150</td>
<td>-0.148</td>
<td>0.058</td>
<td>-0.152</td>
<td>0.043</td>
</tr>
<tr>
<td>$\rho_{12}$</td>
<td>-0.300</td>
<td>-0.300</td>
<td>0.054</td>
<td>-0.303</td>
<td>0.057</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.521</td>
<td>0.541</td>
<td>0.073</td>
<td>0.533</td>
<td>0.051</td>
</tr>
</tbody>
</table>
Table 2

*Monte Carlo Study of a Complex Model with 6 Indicator Variables measuring each Latent Exogenous Variable and 12 Indicators for the Latent Endogenous Variable: PLSc Mean Parameter Estimates (Mean) and Standard Deviations (SD)*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$</td>
<td>0.500</td>
<td>0.500</td>
<td>0.044</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>-0.300</td>
<td>-0.300</td>
<td>0.048</td>
</tr>
<tr>
<td>$\gamma_{12}$</td>
<td>-0.200</td>
<td>-0.200</td>
<td>0.076</td>
</tr>
<tr>
<td>$\gamma_{11}$</td>
<td>0.100</td>
<td>0.100</td>
<td>0.054</td>
</tr>
<tr>
<td>$\gamma_{22}$</td>
<td>-0.150</td>
<td>-0.150</td>
<td>0.051</td>
</tr>
<tr>
<td>$\rho_{12}$</td>
<td>-0.300</td>
<td>-0.301</td>
<td>0.050</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.521</td>
<td>0.539</td>
<td>0.064</td>
</tr>
</tbody>
</table>
Table 3

*Results of the Empirical Interaction Model: Parameter Estimates (Est.) and Jackknife Standard Errors (J-SE) of PLSc and LMS*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>PLSc</th>
<th>LMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$</td>
<td>0.435</td>
<td>0.073</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.262</td>
<td>0.053</td>
</tr>
<tr>
<td>$\gamma_{12}$</td>
<td>-0.134</td>
<td>0.034</td>
</tr>
<tr>
<td>$\rho_{12}$</td>
<td>0.508</td>
<td>0.076</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.551</td>
<td>0.074</td>
</tr>
</tbody>
</table>