Consistent Partial Least Squares estimators for linear and polynomial factor models. A report of a belated, serious and not even unsuccessful attempt. *Comments are invited.*

Theo K. Dijkstra
University of Groningen
Economics & Econometrics
PO BOX 800,
9700 AV Groningen, The Netherlands
t.k.dijkstra@rug.nl

(this version) April 7, 2011.

Abstract

Partial Least Squares algorithms for structural equation models are well-known for their fast convergence, but the ensuing estimators are as a rule inconsistent. The probability limits of the (absolute) loadings on latent variables are generally too large, whereas the limits of (absolute) bivariate and multiple correlations between latent variables are generally too small. Structural coefficients can be highly distorted as a result. In this report we suggest a computationally very simple approach to correct for the inconsistency. We deal with a general class of linear factor models (essentially the 'basic design' in PLS parlance) and invade the field of polynomial factor models. A small simulation study was done using variants of a simple model with interaction that in one form or another has been used for testing many times in the literature. For the models analyzed the approach yielded encouraging results.

Contents

1 Introduction.

This is not a paper. So, meaning no disrespect, an elaborate discussion of the literature in which the approach to be proposed is embedded, contrasted with alternatives and put into perspective, is not (yet) to be found here. It is just a report about a test of some old ideas, and their elaboration, aimed at adjusting certain properties of the Partial Least Squares method. The properties in apparent need of adjustment refer to the bias or inconsistency of PLS, when taken as a tool for estimating factor models[1]. I made embarrasingly simple suggestions ages ago, Dijkstra (1981), but never actually tested them. My rekindled interest in PLS and factor models, driven also by my interest in nonlinear factor models, induced me finally to check them out (and nobody else bothered :-D).

The report is pretty much self-contained, I think. The algebra borders on the trivial, which combined with PLS' relative simplicity and speed could be a bonus. The report has two main sections. Section 2 deals with a class of factor models, with a relatively large scope given the generic character of the class. Section 3 introduces ways to handle certain nonlinear models, polynomial factor models in particular. Here it is much more difficult to specify the scope of the approach. It is certainly not all-encompassing, but it may have practical relevance if the methods stand up to serious scrutiny.

The subtitle of the report is worth repeating: *comments are invited*.

[1] One can refuse to interpret PLS this way, and look at it as a generalization of principal component and canonical variables analysis, which requires different benchmarks, see Dijkstra (2009, 2010) e.g.

2 A linear factor model and Partial Least Squares.

A starting point for PLS-analyses is the so-called 'basic design', in essence a second order factor model. We will use a specification here that is not the most general but will serve our purpose well. It is assumed that we have a number of i.i.d. vectors of indicators, from a population with finite moments of at least order two (the precise order depending on other distributional assumptions or requirements[2]). All indicators have zero mean, and unit variance. The vector of indicators y, say, is composed of at least two subvectors, each measuring a unique latent variable, and each subvector containing at least two components. For the i^{th} subvector y_i we have:

$$y_i = \lambda_i \cdot \eta_i + \epsilon_i \tag{1}$$

where the 'loading vector' λ_i and the 'vector of idiosyncratic errors' ϵ_i have the same dimensions as y_i and the unobservable 'latent variable' η_i is real-valued. All components of all error vectors are assumed to be mutually independent, and independent of all latent variables. The latter have zero mean and unit variance, and they are mutually related via linear equations, be it recursively or with feedback patterns. We will assume here that the coefficients of these structural or 'inner' equations are identifiable from the second order moment matrix of the latent variables[3] via continuous or locally continuously differentiable mappings (continuous when only consistency is required, and continuously differentiable if also asymptotic normality is desired).

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

$$\Sigma_{ii} := Ey_i y_i^\intercal = \lambda_i \lambda_i^\intercal + \Theta_i \tag{2}$$

where Θ_i is diagonal with non-negative diagonal elements, and we have for the covariance:

$$\Sigma_{ij} := Ey_i y_j^\intercal = \rho_{ij} \lambda_i \lambda_j^\intercal \tag{3}$$

where ρ_{ij} stands for the correlation between η_i and η_j. The sample counterparts of Σ_{ii} and Σ_{ij} are denoted by S_{ii} and S_{ij} respectively, and assumptions

[2] If we write $E\left(X^p\right) = Ex_1^{p_1} x_2^{p_2} ... x_q^{p_q}$ for a q-dimensional random vector X and natural numbers $p_1, p_2, ..., p_q$, it will be implicitly understood that the expectation is assumed to exist. A strong law of large numbers will then assure that the sample counterpart based on i.i.d. copies will be a consistent estimator. Asymptotic normality will be the case when $E\left(X^{2 \cdot p}\right)$ exists et cetera.

[3] The original basic design assumed the existence of a 'causal chain', a recursive set of regression equations linking the latent variables, but that restriction is not necessary. Partially identifiable structures can in principle be handled as well, see Bekker and Dijkstra (1990) and Bekker *et al* (1994).

of invertibility are implicit whenever we write S_{ii}^{-1} (or for any other square matrix). Note that the assumptions made sofar entail that (when the sample size tends to infinity) the sample counterparts are consistent estimators of the theoretical variance and covariance matrices.

PLS features two basic types of algorithms, called mode A and mode B, and a third type which we will not discuss that mixes these two. Each mode generates in case of convergence an estimated weight vector $\widehat{w}$, with typical subvector $\widehat{w}_i$ of the same dimensions as y_i. With these weights *proxies* are defined for the latent variables: $\widehat{\eta}_i := \widehat{w}_i^{\intercal} y_i$ for η_i, with the customary normalization of a unit sampling variance, so $\widehat{w}_i^{\intercal} S_{ii} \widehat{w}_i = 1$. For mode A we have for each i:

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

Here $sign_{ij}$ is the sign of the sample correlation between $\widehat{\eta}_i$ and $\widehat{\eta}_j$, and $Adj\,(i)$ is the set of indices of latent variables *adjacent* to η_i, i.e. the indices of all latent variables who appear on the opposite side of the inner equations in which η_i appears. So $j \in Adj\,(i)$ if η_j 'directly' affects η_i or when it is 'directly' affected by η_i. Clearly, $\widehat{w}_i$ is obtained by a regression of the indicators y_i on the 'sign-weighted sum' of the adjacent proxies: $\sum_{j \in Adj(i)} sign_{ij} \cdot \widehat{\eta}_j$. There are other versions (correlation weights e.g.), this is one of the very simplest, and it is the original one (see Wold(1982)). The particular choice can be shown to be irrelevant for the probability limits of the estimators. The algorithm takes arbitrary starting vectors, 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).

For mode B we have for each i:

$$\widehat{w}_i \propto S_{ii}^{-1} \sum_{j \in Adj(i)} sign_{ij} \cdot S_{ij} \widehat{w}_j. \tag{5}$$

In other words, now we get the weights the other way around, by regressing the sign-weighted sum $\sum_{j \in Adj(i)} sign_{ij} \cdot \widehat{\eta}_j$ on the indicators $y_{i.}$

In PLS the estimated proxies replace the latent variables, and the coefficients of the inner equations are determined by regressions or other appropriate means on the proxies; loadings are estimated by regressions of each block of indicators on their corresponding proxy.

Dijkstra (1981), see also Dijkstra (2010), has shown that the modes converge with a probability tending to one when the sample size tends to infinity, and that the weight vectors are locally continuously differentiable functions of the sample covariance matrix of y. A useful implication is that the probability limit of each $\widehat{w}_i$ can be obtained by replacing S_{ij} and S_{ii} by Σ_{ij} and

Σ_{ii} respectively, in fact this is true for every well-behaved function of S and $\widehat{w}$. We will exploit that in the next sections.

2.1 A consistent version of PLS for the basic design.

2.1.1 Mode *A*.

Substituting Σ for S in

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

and recalling that $\Sigma_{ij} \propto \lambda_i \lambda_j^\intercal$ yields immediately that the probability limit of $\widehat{w}_i$ is proportional to λ_i. If we denote this limit by $\overline{w}_i$ we have in fact

$$\overline{w}_i := p \lim \widehat{w}_i = \lambda_i \div (\lambda_i \Sigma_{ii} \lambda_i)^{\frac{1}{2}} . \tag{6}$$

In PLS loadings are estimated by a regression of the indicators y_i on their direct sample proxy $\widehat{\eta}_i$ but since that takes away the proportionality *we will not follow this tradition for mode A*. Instead we will rescale the estimated weights, see below.

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

$$R^2 \left(\eta_i, \overline{\eta}_i\right) = \left(\overline{w}_i^\intercal \lambda_i\right)^2 \tag{7}$$

which equals

$$\left(\lambda_i^\intercal \lambda_i\right)^2 \div \lambda_i \Sigma_{ii} \lambda_i = \frac{\left(\lambda_i^\intercal \lambda_i\right)^2}{\left(\lambda_i^\intercal \lambda_i\right)^2 + \lambda_i^\intercal \Theta_i \lambda_i}. \tag{8}$$

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:

$$R^2 \left(\overline{\eta}_i, \overline{\eta}_j\right) = \left(\overline{w}_i^\intercal \Sigma_{ij} \overline{w}_j\right)^2 = \rho_{ij}^2 \cdot R^2 \left(\eta_i, \overline{\eta}_i\right) \cdot R^2 \left(\eta_j, \overline{\eta}_j\right) \tag{9}$$

indicating that the PLS-proxies will tend to underestimate the correlations between the latent variables. In fact one can show that this is true for multiple correlations as well, see Dijkstra (2010). So there appears to be a case for a rescaling of the weights.

We propose here to rescale $\widehat{w}_i$ in such a way as to reproduce the off-diagonal elements of S_{ii} as well as possible, i.e. such that

$$trace \left[\left(S_{ii} - diag \left(S_{ii} \right) - \left(c_i^2 \cdot \widehat{w}_i \widehat{w}_i^\intercal - diag \left(c_i^2 \cdot \widehat{w}_i \widehat{w}_i^\intercal \right) \right) \right)^2 \right] \tag{10}$$

is as small as possible for a proper choice of c_i^2. If the latter is positive, which it will be with a probability tending to one, the optimal value, denoted by $\widehat{c}_i$, is taken to be its positive square root. Other target functions present themselves, including all metrics used in covariance matrix fitting, taking corrections for loadings as well as correlations simultaneously into account. But the approach suggested appears to be one of the simplest options, and it does not require additional numerical optimization. In fact we have

$$\widehat{c}_i^2 = \frac{\widehat{w}_i^\intercal \left(S_{ii} - diag\left(S_{ii}\right)\right) \widehat{w}_i}{\widehat{w}_i^\intercal \left(\widehat{w}_i \widehat{w}_i^\intercal - diag\left(\widehat{w}_i \widehat{w}_i^\intercal\right)\right) \widehat{w}_i}. \tag{11}$$

It is straightforward to verify that the correction does its job:

$$\overline{c}_i := p \lim \widehat{c}_i = \left(\lambda_i \Sigma_{ii} \lambda_i\right)^{\frac{1}{2}} \tag{12}$$

and so that in particular (equation (6))

$$\widehat{\lambda}_i := \widehat{c}_i \widehat{w}_i \longrightarrow \lambda_i \text{ in probability.} \tag{13}$$

Also note that

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

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

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

Moreover, see equation (9), we can estimate the correlations between the latent variables consistently:

$$\widehat{\rho}_{ij}^2 := \frac{\widehat{R^2\left(\overline{\eta}_i, \overline{\eta}_j\right)}}{\widehat{R^2\left(\eta_i, \overline{\eta}_i\right)} \cdot \widehat{R^2\left(\eta_j, \overline{\eta}_j\right)}} = \frac{\left(\widehat{w}_i^\intercal S_{ij} \widehat{w}_i\right)^2}{\left(\widehat{w}_i^\intercal \widehat{w}_i\right)^2 \cdot \widehat{c}_i^2 \cdot \left(\widehat{w}_j^\intercal \widehat{w}_j\right)^2 \cdot \widehat{c}_j^2} \tag{16}$$

and therefore all coefficients of the equations linking the latent variables, if they are expressable in a smooth way in terms of the correlations.

A remark on tests of fit. Since the implied $\widehat{\Sigma}$ based on direct substitution of the corrected PLS-estimators is consistent, one may if so inclined consider the use of overall tests as in the mainstream literature, by defining a proper metric $d\left(S, \widehat{\Sigma}\right)$ like $trace\left(S - \widehat{\Sigma}\right)^2$ e.g. The p-value can be estimated using the bootstrap, preceeded by a pre-multiplication of the observation vectors by $\widehat{\Sigma}^{\frac{1}{2}} S^{-\frac{1}{2}}$ so that the observation vectors can be treated as coming from a population with a covariance matrix with the assumed (H_0) structure. See e.g. Yuan & Hayashi (2003) for a general discussion and elaboration in the contect of covariance analysis. Given the speed of PLS, mode A in particular, this appears to be quite feasible (we will report the results for tests of uniformity of the distribution of p-values later).

A small experiment for mode *A*. Here we will report the outcomes of a small Monte Carlo study. The same setup will be used for mode B and an extended version will be employed in the section about interaction[4]. We have just three latent variables, each measured by six indicators with identical loadings equal to .7. The latent variables η_1 and η_2 are independent, and

$$\eta_3 = .3\eta_1 + .5\eta_2 + \zeta \tag{17}$$

where ζ is the regression residual. So the true R-squared of the inner equation is a mere .34. We take $y := [y_1; y_2; y_3]$ or equivalently $[\eta; \epsilon]$ to be multinormal.

The PLS-proxies have the same quality:

$$R^2\left(\eta_i, \overline{\eta}_i\right) = .8522 \tag{18}$$

for all i. The probability limits of the correlations between the PLS-proxies (cf. equation (9)) are

$$\begin{bmatrix} 1 & 0 & .2557 \\ 0 & 1 & .4261 \\ .2557 & .4261 & 1 \end{bmatrix} \tag{19}$$

to be compared with the true matrix

$$\begin{bmatrix} 1 & 0 & .3000 \\ 0 & 1 & .5000 \\ ..3000 & .5000 & 1 \end{bmatrix}. \tag{20}$$

As promised, PLS when uncorrected underestimates the true correlations. The probability limits of the regression coefficients are too small: $[.2557, .4261]$ versus $[.3000, .5000]$. And its implied multiple R-squared is .2469 which is also too small (it should be .3400).

We tested the proposed corrections by generating 10.000 independent random samples of size 400 from the 18-dimensional normal distribution as specfified above. As usual each dataset was standardized. For all samples mode A was applied with a stopping criterion of .000001 (so the algorithm stops when the largest absolute difference between consecutive weight estimates is less than one millionth for the first time), and starting values were generated randomly. In 98% of the cases 4 iterations were used, the other cases required 5 iterations[5]. All samples yielded admissible values for the correction factors (no negative $\widehat{c}_i^2$).

The average values of the loading estimates were essentially the same, on average .6926 (versus .7). Their average standard deviations were .12, .0682,

[4]We adopt there the setup as used in Schermelleh-Engel *et a*l (2010).

[5]It took a minute to complete the entire exercise (4CPU 2.40 Ghz; RAM 512 MB).

and .0591 for the loadings of y_1, y_2 and y_3 respectively. Heywood cases, in the sense that estimated loadings exceeded 1, did occur for the loadings of y_1 in about 1.8% of the cases, but not for the others. On average the R-squared is .3503 (versus .3400), with a standard deviation of .0482. The average estimates for the correlations between the latent variables, with standard deviations in parentheses are

$$\begin{bmatrix} 1 & .0048\ (.0574) & .3069\ (.0524) \\ & 1 & .5033\ (.0458) \\ & & 1 \end{bmatrix} \tag{21}$$

The average regression coefficents are $[.3047, .5022]$ with standard devations $[.0487, .0446]$ respectively. Finally, the estimated qualities of the proxies are close to unbiased: their average values and their standard deviations are essentially the same, on average .8573 (versus .8522) and .0123 respectively.

So the corrections yield estimators that are close to unbiased. The expected loadings are slightly too low, the expected correlations are slightly too high, but the exercise is encouraging (to the author).

2.1.2 Mode *B*.

Proceeding as before, substituting Σ for S in the fixed point equations for mode B we obtain

$$\overline{w}_i := p\lim \widehat{w}_i = \Sigma_{ii}^{-1}\lambda_i \div \left(\lambda_i^{\intercal}\Sigma_{ii}^{-1}\lambda_i\right)^{\frac{1}{2}}. \tag{22}$$

Now a regression of y_i on the proxy makes sense:

$$\overline{\lambda}_i := p\lim\left(S_{ii}\widehat{w}_i\right) = \Sigma_{ii}\overline{w}_i = \lambda_i \div \left(\lambda_i^{\intercal}\Sigma_{ii}^{-1}\lambda_i\right)^{\frac{1}{2}}. \tag{23}$$

Again as before with $\overline{\eta}_i := \overline{w}_i^{\intercal} y$ we have $R^2\left(\eta_i, \overline{\eta}_i\right) = \left(\overline{w}_i^{\intercal}\lambda_i\right)^2$ which equals here

$$R^2\left(\eta_i, \overline{\eta}_i\right) = \lambda_i^{\intercal}\Sigma_{ii}^{-1}\lambda_i. \tag{24}$$

Using Cauchy-Schwartz one finds that the quality of the mode B proxies is higher than mode *A's* proxies unless λ_i is an eigenvector of Σ_{ii}, (which will happen only when all error variances are equal ($\Theta_i \propto I$)), in which case we have equality. (This happens to apply to the experimental setup above).

Now define $\widehat{c}_i$ as the positive square root of

$$\frac{\widehat{\lambda}_i^{\intercal}\left(S_{ii} - diag\left(S_{ii}\right)\right)\widehat{\lambda}_i}{\widehat{\lambda}_i^{\intercal}\left(\widehat{\lambda}_i\widehat{\lambda}_i^{\intercal} - diag\left(\widehat{\lambda}_i\widehat{\lambda}_i^{\intercal}\right)\right)\widehat{\lambda}_i} \tag{25}$$

which will be positive with a probability tending to one. One verifies again that

$$p\lim \widehat{c}_i \widehat{w}_i = \lambda_i, \tag{26}$$

moreover we have

$$p\lim \widehat{c}_i^2 = R^2\left(\eta_i, \overline{\eta}_i\right), \tag{27}$$

and so because (9) is true also[6],

$$\widehat{\rho}_{ij}^2 := \frac{\left(\widehat{w}_i^{\intercal} S_{ij} \widehat{w}_i\right)^2}{\widehat{c}_i^2 \cdot \widehat{c}_j^2} \tag{28}$$

is consistent for ρ_{ij}^2.

We ran the same Monte Carlo experiment for mode B. All samples yielded admissible values for the correction factors (no negative $\widehat{c}_i^2$). As for mode A the corrections lead to estimators that are close to unbiased, except for the qualities of the proxies. In particular $R^2\left(\eta_1, \overline{\eta}_1\right)$ had a large upward bias of 16%. Mode B is numerically less stable and less fast: it took twice as long to complete the exercise, with rather more iterations per sample (varying from 6 to 39 with a median number of 9 iterations). The standard deviations were rather similar to those of mode A for most parameters, the exception being the qualities of the proxies (their standard deviations were roughly twice as large).

In the sequel we will work with mode A.

3 Adding polynomial terms to a simple model.

Here we will show how one can use the approach as just developed, to get consistent estimators for the coefficients of interaction and quadratic terms and higher order polynomials. We will use a simple setup with just three latent variables to illustrate this. First we will add interaction to the linear equation. It will turn out that we do not need to make distributional assumptions beyond those concerning independence and the existence of moments to get consistency (and asymptotic normality), although normality of $[\eta_1; \eta_2]$ will help to simplify the estimation process somewhat. Adding also quadratic or higher order terms necessitates additional distributional assumptions and although normality will not be strictly necessary, it will be quite helpful. The impact of these assumptions when unwarranted are not investigated in

[6] In Dijkstra (1981, 2010) another suggestion is made, aimed at reproducing the covariance matrix between blocks, *given* the loading corrections, as well as possible. For mode A it yields the same number, but not for mode B, it remains to be tested.

this note. Anyway, since the literature seems to be mainly focussed on the Gaussian situation (for those latent variables whose normality is compatible with the nonlinear relationships), it is an acceptable start.

3.1 Interaction.

Now consider

$$\eta_3 = \gamma_1\eta_1 + \gamma_2\eta_2 + \gamma_{12}\left(\eta_1\eta_2 - E\eta_1\eta_2\right) + \zeta \tag{29}$$

where all variables have zero mean, the η's have unit variance, and the (regression) residual ζ is independent of η_1 and η_2 (and therefore also of their product). As before the latent variables are measured indirectly by $y_i = \lambda_i \cdot \eta_i + \epsilon_i$ and all components of all error vectors are mutually independent, and independent of all latent variables (and therefore of ζ).

Adapting Wold's suggestion how to handle quadratic terms in inner equations, see Wold (1982) and Dijkstra & Henseler (2011), we get for mode A the following fixed point equations for the weight vectors:

$$\widehat{w}_1 \propto S_{13}\widehat{w}_3 \tag{30}$$

$$\widehat{w}_2 \propto S_{23}\widehat{w}_3 \tag{31}$$

so $\widehat{w}_1$ and $\widehat{w}_2$ are obtained from a regression of y_1 and y_2 on $\widehat{\eta}_3$ respectively, and $\widehat{w}_3$ is obtained from a regression of y_3 on

$$sign_{13} \cdot \widehat{\eta}_1 + sign_{23} \cdot \widehat{\eta}_2 + sign_{3,1\times 2} \cdot (\widehat{\eta}_1\widehat{\eta}_2 - mean(\widehat{\eta}_1\widehat{\eta}_2)) \tag{32}$$

where $sign_{3,1\times 2}$ is the sign of the correlation between $\widehat{\eta}_3$ and the product of $\widehat{\eta}_1$ and $\widehat{\eta}_2$. There are other options, but again this does not affect the probability limits. As before we will use the weights as estimates for the loadings. We get the same probability limits as for the linear model. In PLS one would treat the proxies as stand-ins for the latent variables and regress $\widehat{\eta}_3$ on a constant and $\widehat{\eta}_1$, $\widehat{\eta}_2$ and $\widehat{\eta}_1\widehat{\eta}_2$ to get estimates $\widehat{\gamma}$ for the γ's[7]. Note that $\overline{\gamma} := p\lim\widehat{\gamma}$ satisfies:

$$\begin{bmatrix} 1 & E\overline{\eta}_1\overline{\eta}_2 & E\overline{\eta}_1^2\overline{\eta}_2 \\ E\overline{\eta}_1\overline{\eta}_2 & 1 & E\overline{\eta}_1\overline{\eta}_2^2 \\ E\overline{\eta}_1^2\overline{\eta}_2 & E\overline{\eta}_1\overline{\eta}_2^2 & E\overline{\eta}_1^2\overline{\eta}_2^2 - \left(E\overline{\eta}_1\overline{\eta}_2\right)^2 \end{bmatrix} \begin{bmatrix} \overline{\gamma}_1 \\ \overline{\gamma}_2 \\ \overline{\gamma}_{12} \end{bmatrix} = \begin{bmatrix} E\overline{\eta}_1\overline{\eta}_3 \\ E\overline{\eta}_2\overline{\eta}_3 \\ E\overline{\eta}_1\overline{\eta}_2\overline{\eta}_3 \end{bmatrix}. \tag{33}$$

[7]This would be Wold's original version. There are alternative approaches that use products of indicators as indicators of the product, as is popular in mainstream structural equation modelling, see Schumacker & Marcoulides (1998) e.g.

The coefficient vector γ satisfies the same equation, without the bars. But since the moments of the proxies differ as a rule from the moments of the latent variables one expects that $\overline{\gamma} \neq \gamma$. Using $\overline{\eta}_i = R\left(\overline{\eta}_i, \eta_i\right)\eta_i + \delta_i$ with $\delta_i := \overline{w}_i^{\intercal}\epsilon_i$ where the δ's are mutually independent and independent of the η's we have in fact:

$$E\overline{\eta}_1^k\overline{\eta}_2^l\overline{\eta}_3^m = E\eta_1^k\eta_2^l\eta_3^m \cdot R^k\left(\eta_1, \overline{\eta}_1\right) \cdot R^l\left(\eta_2, \overline{\eta}_2\right) \cdot R^m\left(\eta_3, \overline{\eta}_3\right)$$

where $k, l, m = 0, 1$ or 2 for any of the moments in (33) except for:

$$E\overline{\eta}_1^2\overline{\eta}_2^2 = \left(E\eta_1^2\eta_2^2 - 1\right) \cdot R^2\left(\eta_1, \overline{\eta}_1\right) \cdot R^2\left(\eta_2, \overline{\eta}_2\right) + 1. \tag{34}$$

As with the linear model we can solve the equations for the moments of the latent variables, in terms of the qualities of the proxies and their moments. Inserting sample counterparts in the appropriate normal equations yields consistent estimators for γ.

So far distributional assumptions like normality are not needed, the approach is a general one. But we note that if one is willing to work with normality of $[\eta_1; \eta_2]$, some expressions simplify. In fact[8] we then have $E\eta_1^2\eta_2 = E\eta_1\eta_2^2 = 0$ and $E\eta_1^2\eta_2^2 = 1 + 2\left(E\eta_1\eta_2\right)^2$.

We ran the same experiment as before, with the difference that now

$$\eta_3 = .3\eta_1 + .5\eta_2 + .3\eta_1\eta_2 + \zeta \tag{35}$$

so the true R-squared is .43. The other parameters were kept the same, so the quality of the proxies is the same ($R^2\left(\eta_i, \overline{\eta}_i\right) = .8522$ for all i). PLS uncorrected will yield asymptotically [.2557, .4261, .2360] for γ instead of [.3, .5, .3] and the R-squared wil be .3026. As for the linear model the (absolute) coefficients and R-squared are too small.

We tested the proposed corrections by generating 10.000 independent random samples of size 400 from the 18-dimensional *non*-normal distribution as specfified above. In particular, this means that η_1, η_2, ζ and the components of ϵ_1, ϵ_2, and ϵ_3 are mutually independent and normal, with zero means. The distribution of $[\eta_3; y_3]$ is decidedly non-normal; all indicators and latent variables (η) have unit variance. As usual each dataset was standardized. For all samples mode A was applied *with the general correction formulae,* normality was not used, with a stopping criterion of .000001 and starting values were generated randomly. In 98.5% of the cases 4 iterations were used, the other

[8] $E\eta_1^2\eta_2 = E\eta_1^2\left(\rho_{12}\eta_1 + \zeta_{2.1}\right) = \rho_{12}E\eta_1^3 + E\eta_1^2\zeta_{2.1} = 0$ because of the symmetry and the regression residual's ($: \zeta_{2.1}$) independence of η_1. Similarly $E\eta_1\eta_2^2 = 0$, and $E\eta_1^2\eta_2^2 = \rho_{12}^2E\eta_1^4 + E\eta_1^2E\zeta_{2.1}^2 = 3\rho_{12}^2 + 1 - \rho_{12}^2 = 1 + 2\left(E\eta_1\eta_2\right)^2$

cases required 3 or 5 iterations[9]. All samples yielded admissible values for the correction factors (no negative $\widehat{c}_i^2$).

The average value of the loading estimates and the standard deviations were essentially as before, and similarly for the estimated qualities of the proxies. Heywood cases, in the sense that estimated loadings exceeded 1, did occur for the loadings of y_1 in about 2.2% of the cases, but not for the others. On average the R-squared is .4402 (versus .4300), with a standard deviation of .0527. The average regression coefficents are $[.3033, .5028, .2996]$ with standard devations $[.0498, .0462, .0561]$ respectively.

So again the corrections yield estimators that are close to unbiased.

3.2 Interaction and quadratic terms.

As expected, we will now consider

$$\eta_3 = \gamma_1\eta_1 + \gamma_2\eta_2 + \gamma_{11}\left(\eta_1^2 - 1\right) + \gamma_{22}\left(\eta_2^2 - 1\right) + \gamma_{12}\left(\eta_1\eta_2 - E\eta_1\eta_2\right) + \zeta \quad (36)$$

where as before all variables have zero mean, the η's have unit variance, and the (regression) residual ζ is independent of η_1 and η_2 (and therefore also of their product and squares). Also, the latent variables are measured indirectly by $y_i = \lambda_i \cdot \eta_i + \epsilon_i$, with all components of all error vectors mutually independent, and independent of all latent variables (and ζ).

Mode A calculates $\widehat{w}_1$ and $\widehat{w}_2$ as before. $\widehat{w}_3$ is now obtained from a regression of y_3 on the sign-weighted sum

$$\begin{aligned} &sign_{13} \cdot \widehat{\eta}_1 + sign_{23} \cdot \widehat{\eta}_2 + sign_{3,1\times1} \cdot \left(\widehat{\eta}_1^2 - 1\right) + sign_{3,2\times2} \cdot \left(\widehat{\eta}_2^2 - 1\right) \\ &+ sign_{3,1\times2} \cdot \left(\widehat{\eta}_1\widehat{\eta}_2 - mean(\widehat{\eta}_1\widehat{\eta}_2)\right) \end{aligned} \quad (37)$$

where $sign_{3,1\times1}$ is the sign of the correlation between $\widehat{\eta}_3$ and $\widehat{\eta}_1^2$ et cetera.

When we denote the second order moment matrix for the regressors of (36) by V, so $V :=$

$$\begin{bmatrix} 1 & E\eta_1\eta_2 & E\eta_1^3 & E\eta_1\eta_2^2 & E\eta_1^2\eta_2 \\ E\eta_1\eta_2 & 1 & E\eta_1^2\eta_2 & E\eta_2^3 & E\eta_1\eta_2^2 \\ E\eta_1^3 & E\eta_1^2\eta_2 & E\eta_1^4 - 1 & E\eta_1^2\eta_2^2 - 1 & E\eta_1^3\eta_2 - E\eta_1\eta_2 \\ E\eta_1\eta_2^2 & E\eta_2^3 & E\eta_1^2\eta_2^2 - 1 & E\eta_2^4 - 1 & E\eta_1\eta_2^3 - E\eta_1\eta_2 \\ E\eta_1^2\eta & E\eta_1\eta_2^2 & E\eta_1^3\eta_2 - E\eta_1\eta_2 & E\eta_1\eta_2^3 - E\eta_1\eta_2 & E\eta_1^2\eta_2^2 - (E\eta_1\eta_2)^2 \end{bmatrix} \quad (38)$$

[9]It took 71 seconds to complete the entire exercise (4CPU 2.40 Ghz; RAM 512 MB).

we have for $\gamma := [\gamma_1; \gamma_2; \gamma_{11}; \gamma_{22}; \gamma_{12}]$:

$$V\gamma = \begin{bmatrix} E\eta_1\eta_3 \\ E\eta_2\eta_3 \\ E\eta_1^2\eta_3 \\ E\eta_2^2\eta_3 \\ E\eta_1\eta_2\eta_3 \end{bmatrix}. \tag{39}$$

The previous section allows one to estimate most of the entries in the relevant matrices consistently, except for $E\eta_1^3\eta_2$ (and $E\eta_1\eta_2^3$), $E\eta_1^3$ (and $E\eta_2^3$) , $E\eta_1^4$ (and $E\eta_2^4$), and $E\eta_1^3\eta_2$ (and $E\eta_1\eta_2^3$). With $\overline{\eta}_i = R\left(\overline{\eta}_i, \eta_i\right)\eta_i + \delta_i$ we get with some algebra

$$E\overline{\eta}_1^3\overline{\eta}_2 = R^3\left(\overline{\eta}_1, \eta_1\right) R\left(\overline{\eta}_2, \eta_2\right) E\eta_1^3\eta_2 + 3E\overline{\eta}_1\overline{\eta}_2\left(1 - R^2\left(\overline{\eta}_1, \eta_1\right)\right) \tag{40}$$

so $E\eta_1^3\eta_2$ (and similarly $E\eta_1\eta_2^3$) can be estimated consistently. The third and the fourth order moments spoil things a bit:

$$\begin{aligned} E\overline{\eta}_1^3 &= R^3\left(\overline{\eta}_1, \eta_1\right) E\eta_1^3 + E\delta_1^3 & (41) \\ E\overline{\eta}_1^4 &= R^4\left(\overline{\eta}_1, \eta_1\right) E\eta_1^4 + 6R^2\left(\overline{\eta}_1, \eta_1\right)\left(1 - R^2\left(\overline{\eta}_1, \eta_1\right)\right) + E\delta_1^4 & (42) \end{aligned}$$

with analogous expressions for $E\overline{\eta}_2^3$ and $E\overline{\eta}_2^4$. So 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 ϵ_i's would only come halfway; eventhough now $E\delta_i^3 = 0$ the *kurtosis* is not fixed. For want of reasonable alternatives we will adopt normality here. More precisely, we will work with joint normality of $[\eta_1; \eta_2; \zeta; \epsilon]$. So $E\delta_i^4 = 3\left(E\delta_i^2\right)^2 = 3\left(1 - R^2\left(\overline{\eta}_i, \eta_i\right)\right)^2$ and we have consistent estimation of the relevant moments, and therefore of γ.

The other regressor moments simplify as well. For completeness sake we will specify them:

$$V = \begin{bmatrix} 1 & \rho_{12} & 0 & 0 & 0 \\ \rho_{12} & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2\rho_{12}^2 & 2\rho_{12} \\ 0 & 0 & 2\rho_{12}^2 & 2 & 2\rho_{12} \\ 0 & 0 & 2\rho_{12} & 2\rho_{12} & 1 + \rho_{12}^2 \end{bmatrix}. \tag{43}$$

For the special case $\rho_{12} = 0$ the matrix V is diagonal and

$$\gamma = \left[E\eta_1\eta_3, E\eta_2\eta_3, \frac{1}{2}E\eta_1^2\eta_2, \frac{1}{2}E\eta_1\eta_2^2, E\eta_1\eta_2\eta_3\right] \tag{44}$$

so that the R-squared of the inner equation equals $\gamma_1^2 + \gamma_2^2 + 2\gamma_{11}^2 + 2\gamma_{22}^2 + \gamma_{12}^2$. If we use PLS uncorrected the elements of γ will be underestimated in the limit, as well as the R-squared of course.

The numerical experiment continued. We will keep exactly the same setup as on previous occasions, with η_1 and η_2 independent[10], except that the inner equation now reads:

$$\eta_3 = .3\eta_1 + .5\eta_2 + .2\left(\eta_1^2 - 1\right) + .2\left(\eta_2^2 - 1\right) + .3\left(\eta_1\eta_2 - E\eta_1\eta_2\right) + \zeta. \quad (45)$$

The R-squared is .59. As a reminder, each latent variable is measured by 6 indicators with identical loadings (:.7), so the quality of the proxies is still .8522. The PLS probability limits for γ are

$$\overline{\gamma} = [.2557, .4261, .1573, .1573, .2557] \quad (46)$$

and for its R-squared we note .3459.

We tested the proposed corrections by generating 10.000 independent random samples of size 400 from the 18-dimensional *non*-normal distribution as specified above. In particular, this means that η_1, η_2, ζ and the components of ϵ_1, ϵ_2, and ϵ_3 are mutually independent and normal, with zero means. The distribution of $[\eta_3; y_3]$ is decidedly non-normal; all indicators and latent variables (η) have unit variance. Again each dataset was standardized. For all samples mode A was applied *with the correction formulae based on normality.* This means that V, see (43), is estimated by inserting the corrected PLS-value for ρ_{12} (the entries of the righthand-side of (39) are estimated by the PLS values corrected using the estimated qualities). As usual a stopping criterion of .000001 was employed and starting values were generated randomly. In a bit over 99% of the cases 4 iterations would do, the other cases required 3 or 5 iterations[11]. All samples yielded admissible values for the correction factors (no negative $\widehat{c}_i^2$).

The average value of the loading estimates and the standard deviations were essentially as before, and similarly for the estimated qualities of the proxies. Heywood cases, in the sense that estimated loadings exceeded 1, did occur for the loadings of y_1 in about 2.3% of the cases, but not for the others. On average the R-squared is .6119 (versus .5900), with a standard deviation of .1104 (the median was .6002). The standard deviation is rather larger than for the other models analyzed. In fact, in 3 out of 1000 samples the estimated value was larger than 1.

The average regression coefficents are

$$[.3029, .5017, .1962, .1974, .2957] \quad (47)$$

with standard devations

$$[.0550, .0497, .0541, .0576, .0777] \quad (48)$$

[10] This fact is not used in the estimation procedure.

[11] It took 81 seconds to complete the entire exercise (4CPU 2.40 Ghz; RAM 512 MB).

respectively.

3.3 Higher order terms.

It will be clear that adding yet more terms like η_1^3, η_2^3, $\eta_1^2\eta_2$, $\eta_1\eta_2^2$, and η_1^4 and so on, while maintaining the joint normality of η_1 and η_2, will in principle present no problem, eventhough one may in practical situations want to reflect upon its wisdom. With only two exogenous latent variables all one needs to fill out the moment matrix of the regressors, where entries contain expressions of the form $E\eta_1^k\eta_2^l = E\eta_1^k\left(\rho_{12}\eta_1 + \zeta_{2.1}\right)^l$, is the well-known formula for a standard normal variable z and natural number r:

$$Ez^{2r} = \frac{(2r)!}{r!}\frac{1}{2^r} \tag{49}$$

and all odd moments are zero. The righthand side of the normal equations present no new problems either. With a larger number of exogenous latent variables, there is more work to be done (for a third order polynomial one may wish to use the expressions as given by Mooijaart and Bentler (2010) e.g.). We refer to Meijer (2005) for compact matrix expressions for the moments of the multinormal distribution.

An observation that may help interpret the estimated inner coefficients. 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 μ and covariance matrix Ω, then $f(X)$, where $f(.)$ is a real valued smooth function satisfies (see Liu (1994)):

$$cov\left(X, f(X)\right) = \Omega E\left[\nabla f(X)\right]. \tag{50}$$

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

$$f(X) = Ef(X) + E\left[\nabla f(X)\right]^{\intercal}\cdot(X-\mu) + \text{ regression residual.} \tag{51}$$

Consequently, the regression coefficients are the *average first partial derivatives*. We observe here that when the true specification of the relationship between the endogenous latent variable and the exogenous latent variables requires a general, nonlinear function of the latter, corrected PLS estimation based on a linear relationship yields consistent estimators for the average first partial derivatives. This applies of course to all methods that are consistent in case of linearity.

We would like to add an extension of Stein's identity that covers the quadratic case:

$$\begin{aligned} f(X) &= Ef(X) + E[\nabla f(X)]^{\intercal} \cdot (X-\mu) + \\ &\quad 1/2 \cdot trace\left[E(H(X)) \cdot ((X-\mu)(X-\mu)^{\intercal} - \Omega)\right] + \nu \end{aligned} \quad (52)$$

where $H(.)$ is the Hessian of $f(.)$ and ν is the regression residual. So the regresssion 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 close this section by a proof of the extension, assuming the existence of partial derivatives and boundedness of their expected absolute values[12].

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:

$$E\left(Z_i^2 - 1\right) g(Z) = EZ_i^2 g(Z) - Eg(Z) = \quad (53)$$

$$EZ_i\left(Z_i g(Z)\right) - Eg(Z) = E\left(g(Z) + Z_i \nabla g(Z)_i\right)\right) - Eg(Z) = \quad (54)$$

$$E\left(Z_i \nabla g(Z)_i\right) = EH_{ii}(Z). \quad (55)$$

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 $(i \neq j)$:

$$EZ_i\left(Z_j g(Z)\right) = E\left(Z_j \nabla g(Z)_i\right) = EH_{ij}(Z). \quad (56)$$

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

$$g(Z) = Eg(Z) + E\nabla g(Z)^{\intercal} \cdot Z + 1/2 \cdot trace\left[EH_g(Z) \cdot (ZZ^{\intercal} - I)\right] + \nu. \quad (57)$$

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 (57) yields the desired expression for general $f(X)$.

[12] I have not been able to find a reference. Comments are welcome!

4 Conclusion.

As a way of concluding this preliminary report we will list a number of open questions, not necessarily in the order of significance. Questions 6 and 1 have the highest priority for me.

1. How can one characterize the set of nonlinear models that allow of the treatment as exemplified for the simple model? It encompasses at least systems of equations in terms of latent variables where each equation is a regression, either directly or in 'reduced form'. But one expects there are other systems as well, though certainly not all that can be treated by the standard approaches. Perhaps a detailed study of nonlinear models used in practice can help to delineate the *practical* scope of our approach.

2. What use do negative values of the correction factors $\widehat{c}_i^2$ have in detecting model misspecification? (They could also simply indicate that the sample is too small). Are there other ways of correcting PLS that are more robust and/or lead to better estimators?

3. How should Heywood cases (any instance of a parameter estimate outside its logical region) be handled? They are notoriously difficult to handle in structural equation modelling (see Dijkstra (1992), there is probably more recent work...), we cannot expect anything better here.

4. Using sign-weighted sums of 'adjacent' proxies seems to work well. But are there 'better' alternatives (e.g. correlation-weighted sums)? And would dropping adjacency, using *all* proxies irrespective of the theoretical constraints on the structural equations, improve the quality of the estimators (the loadings in particular)?

5. How does one best specify the metric that measures the distance between S and $\widehat{\Sigma}$ so that both the p-value is robustly estimated and the power is 'maximized' for meaningful prespecified alternatives? (I may be slightly out of touch here...).

6. And, quite possibly the most interesting question, how *good* is the approach developed here for nonlinear models when set against the proper benchmark, i.e. maximum likelihood estimation that takes the nonlinear relationships fully into account? Reference is made to the LMS method, see e.g. Schermelleh-Engel *et al* (1998). (The other somewhat *ad hoc* pragmatic approaches, who use products of indicators, instrumental variables or third order moments cannot in general claim to be (asymptotically) efficient under normality).

References

[1] Bekker, P. A. and Dijkstra, T. K. (1990). On the nature and number of the constraints on the reduced form as implied by the structural form, *Econometrica* 58: 507-514.

[2] Bekker, P. A., A. Merckens, T. J. Wansbeek (1994). *Identification, Equivalent Models, and Computer Algebra.* Academic Press, Boston.

[3] Dijkstra, T. K. (1981). *Latent variables in linear stochastic models.* PhD thesis. (second edition (1985), Amsterdam: Sociometric Research Foundation).

[4] Dijkstra, T. K. (1992). On statistical inference with parameter estimates on the boundary of the parameter space. *British Journal of Mathematical and Statistical Psychology,* vol. 45, pp. 289-309.

[5] Dijkstra, T. K. (2009). PLS for path diagrams revisited, and extended. *Proceedings of the 6^{th} International Conference on Partial Least Squares*, September 4-7, 2009, Beijing.

[6] Dijkstra, T. K. (2010). Latent variables and indices. In: Vinzi, V. Esposito, Chin, W. W., Henseler, J., Wang, H. (eds.), *Handbook of Partial Least Squares,* chapter 1, pp. 23-46. Springer-Verlag, Heidelberg.

[7] Dijkstra, T. K. and J. Henseler (2011). Presciptions for Dimension Reduction, with Interacting Factors. *Quality & Quantity* (forthcoming).

[8] Liu, J. S. (1994). Siegel's formula via Stein's identities. *Statistics & Probability Letters* 21, no. 3, pp. 247-251.

[9] Meijer, E. (2005). Matrix algebra for higher order moments. *Linear Algebra and its Applications,* vol. 410, pp.112-134.

[10] A. Mooijaart and P. M. Bentler (2010). An Alternative Approach for Nonlinear Latent Variable Models. *Structural Equation Modeling: A Multidisciplinary Journal* 29, no. 3, pp. 317-331.

[11] Schermelleh-Engel, K., A. Klein, H. Moosbrugger (1998). Estimating Nonlinear Effects Using a Latent Moderated Structural Equations Approach. In: Schumacker, R. E. and G. A. Marcoulides, eds. (1998).*Interaction and Nonlinear Effects in Structural Equation Modeling.* Lawrence Erlbaum, Mahwah, chapter 10, pp. 239-251.

[12] Schermelleh-Engel, K., C. S. Werner, A. G. Klein, H. Moosbrugger (2010). Nonlinear structural equation modelling: is partial least squares an alternative?. *Adv. Stat. Anal.* 94: 167-184.

[13] Schumacker, R. E. and G. A. Marcoulides, eds. (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling.* Lawrence Erlbaum, Mahwah.

[14] Serfling, R. J. (2004). Multivariate Symmetry and Asymmetry. *Encyclopedia of Statistical Sciences.* Wiley.

[15] Yuan, K-H and K. Hayashi (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology* 56, pp. 93-110.

[16] Wold, H. O. A. (1982). Soft modelling: the basic design and some extensions, *in* Jöreskog, K. G., Wold, H. O. A. (eds), *Systems under indirect observation, Part II,* North-Holland, Amsterdam, chapter 1, 1-55.