PLS for path diagrams revisited, and extended.

Prepared for the 6th International Conference on Partial Least Squares, September 4-7, 2009, Beijing, China

Theo K. Dijkstra
University of Groningen
Economics & Econometrics
&
SNS Asset Management
The Netherlands
E-mail: theo.dijkstra@snsam.nl

Abstract: I extend the usual indices as linear compounds, by introducing 'proper indices' and using concepts of fuzzy logic. A proposal is made to always study and report the full range of possible structural parameter values as determined by variations across all admissible indices. An appeal is made to use more structural information in the construction of indices. And finally, for a particular class of empirical applications, it is pointed out that proper indices for non-linear path models can also be obtained by combining Wold's mode B approach with a non-negative least squares routine.

Keywords: linear index, fuzzy logic, quadratic model with interaction, simulation.

1. Introduction.

From the start PLS distinguished itself from mainstream factor modelling by the explicit substitution of (linear) indices for unobserved latent variables in path diagrams. The leading generic cases were principal components and canonical variables, see H.O.A. Wold (1966, 1975, 1982). The choice of weights for the indicators constituting the constructs embedded in a path diagram was dealt with by a large variety of 'alternating least squares'-algorithms. These are cycles of regressions that update natural subsets of weights in turn, and that come to a halt when changes in consecutive updates are no longer numerically significant. In practice the algorithms typically converge, and when they do, they converge rather fast. See T.K. Dijkstra (1981, 2007) for a proof showing that the PLS-algorithms will converge from arbitrary starting points to unique fixed points, with a probability tending to one when the size of the random sample goes to infinity, provided the structure of the true covariance matrix of the indicators is sufficiently close to (a generalization of) Wold's *basic design*. There is a plethora of PLS-algorithms, see Tenenhaus *et al* (2005) for a recent overview, and it is not trivial at all to make an unambiguous, informed choice that does not appear to be opportunistic or arbitrary. Generally, except for mode A and mode B the PLS-algorithms are not known to solve global, as opposed to local, optimization problems[1]. I propose to go back to the point of departure, and partly rethink the process of constructing indices[2]. This will lead to some 'new' concepts, and a research agenda.

---

[1] Mode B maximizes the sum of the absolute correlations between all 'adjacent' indices, see Dijkstra (1981, 2007), and a similar statement, in terms of covariances, is true for mode A, see McDonald (1996). The simplest version of mode C can also be construed as an optimization program.

[2] I witnessed PLS' original development firsthand as a member of Herman Wold's research team at the Wharton School, Philadelphia, in the fall of 1977. My attempts at the time to incorporate structural information in the estimation process complicated it rather substantially. Since Herman Wold saw PLS' main potential in high-dimensional, low-structure environments with soft data, he opted for a radical simplification. He honoured me by serving as the decisive member on my PhD-committee. The present paper is part of a sequence of papers, see

2. Proper indices.

It seems undeniable that in many cases indicators for concepts are selected on the basis of a presumed monotonous relationship with the underlying concept. I contend that the index comprising the indicators should honour this: in the generic case, the weights of the indicators *as well as* the loadings on the index ought to be non-negative. Then an indicator contributes to the index in a non-negative way, and it faithfully reflects its movements 'on the average' as well. I suggest to call such an index *proper*.

In practice one often attempts to construct an index using the first principal component, but when some of its coefficients are negative, it is typically discarded and replaced by a (simple) mean. Checking whether the ensuing loadings are non-negative does not appear to be routine, but a check is certainly not superfluous. I have shown however that non-degenerate indicators always allow of a proper index: there is always a non-empty, convex and compact set of non-negative weight vectors with ensuing non-negative loadings, see Dijkstra (2005, 2007).

It does not appear to be straightforward to instruct the PLS-algorithms to produce proper indices. Mode B, where weight vectors are obtained by regressions of sign-weighted sums of (functions of) indices on a vector of indicators (in Wold's original specification), is perhaps easiest to adjust: just replace the multiple regressions by a non-negative least squares routine. If the sequence of weights converges, and if within each block all indicators are non-negatively correlated, then the ensuing loadings will also be non-negative. Section 6 contains an example.

3. Elastic (proper) indices.

I suggest to construct indices with the help of a method designed by Yager (1988, 1996, 1999) for aggregation purposes in *multi-criteria decision analysis,* MCDA. It is developed in the context of Choquet integrals and 'fuzzy measures'[3]. Specifically, if $x_1, x_2 \ldots, x_n$ are the (standardized) scores on the n indicators for a particular object, item or entity, I propose to use the index:

$$I(x\,|\,w,p) \equiv \sum_{i=1}^{i=n} x_{(i)} * \left[\left(\sum_{j=1}^{j=i} w_{(j)} \right)^{p} - \left(\sum_{j=1}^{j=i-1} w_{(j)} \right)^{p} \right]$$

Here $x_{(i)}$ is the i-th largest of the scores, so (.) is a permutation of the scores. The *same* permutation is used for the weights w, so w is sorted according to the ranking of x. The exponent p is a non-negative real number. For p equal to zero the index reduces to the best score, and when p becomes 'large' the index tends to the worst score. Other special cases are:

- The weighted average: take p=1.
- The ordered weighted average[4]: take equal $w_{(j)}$'s. When p is larger then 1, this will generate an OWA that overweighs the lower scores. In this case good performances or high scores *across the board* are required for a high score on the underlying concept. In contrast, when p is smaller then 1, just a few high scores suffice.

Dijkstra (2005, 2007) and Dijkstra and Henseler (2008), forming a renewed attempt on my part to contribute to 'soft modelling'.

[3] See Grabisch *et al* (2000) for the general theory, and Dijkstra (2008) for an analysis of Yager's approach.

[4] Or rather a subset of the OWA's. We can get all of them if we replace the power functions by non-decreasing real functions Q(.) defined on the unit interval, with Q(0)=0 and Q(1)=1; one could call them 'quantifiers' . Convex quantifiers value good performance across the board, they represent dilations of 'all'. Concavity corresponds to leniency, to 'some'.

In general, the weight a particular score $x_{(i)}$ gets in the index depends on the *importance* of the corresponding indicator, as measured by $w_{(i)}$, and of the importance of those on which a higher score is obtained, as well as on the extent to which one insists on high scores across the board as measured by the power coefficient p. In the parlance of fuzzy logic, if the indicators are criteria and the score measures the extent to which a criterion is satisfied, if p exceeds 1 the index measures to what extent 'most' of the important criteria are satisfied ('most' is replaced by 'some' for p less than 1).

4. Full range analysis.

Both within and outside the PLS family there are many, many ways to construct indices, leading to a variety of structural parameter values or estimates, generically defined by regressions (the latter is particularly true for Wold's PLS). The choice of construction method for the indices, is driven by outcomes, purposes, or ease of interpretation, and sometimes by considerations of an aesthetic nature. I propose to always study and report the possible ranges of structural parameter values as determined by 'all' possible weights (and powers). This can usually be done most easily by a Monte Carlo simulation. I suggest first to generate weights and powers from maximum entropy distributions on natural domains subject to 'natural' constraints on expectations (see section 6 for an example). Then I would calculate for each generated configuration of weights and powers the corresponding loadings and 'inner' structural parameters (regression coefficients, multiple correlation coefficients, et cetera). The ensuing (multidimensional) histogram will indicate for the data at hand the extent to which the choice of method matters for the estimation of parameters of interest; it may establish important invariants or reveal substantial instabilities.
If the available data can be seen as a realization of a random sample from a stable population, the bootstrap can be used to assess the variation of the findings across samples of similar size (again see section 6 for an example).

5. Structure-consistent index construction.

The path diagram is a crucial carrier of information and meaning. Every index derives part of its semantic meaning or informational content by the relationships it holds with the other indices. In a purely statistical or data-analytical approach this entails that the path diagram should be involved in the actual construction of the indices. In the archetypal version of PLS the structural information is reduced to 'adjacency', which means that we only keep track of which variables appear at different sides of structural equations. Some versions take bivariate correlations into account, but in general structural information about aspects of joint distributions are ignored in the construction of the indices. I propose to investigate to what extent it will be helpful when we *do* take this into account, and I developed some suggestions in a linear context[5].

An alternative approach would be to use aggregation procedures as developed in MCDA and fuzzy logic, and choose weights, or {importances and quantifiers} consistently with values, preferences or theoretical considerations. What an index 'means' will depend on the conceptual framework in which it is embedded. In practice one will see a combination of

[5] See Dijkstra (2007), in particular section 6.2, for a suggestion how to proceed. The exclusion constraints of (linear) structural equations are transformed into constraints on the correlation matrix of the indices. Optimization with respect to the parameters of the indices of a suitable function of the correlation matrix of the indices subject to the exclusion constraints (and others such as properness), will lead to indices that honor the path diagram in which they jointly appear. There is absolutely no claim to priority here, doubtless many others will have thought of similar methods way before me. But it could deserve a revival.

approaches, with statistical results partly overruled when they conflict with matters of substance.

6. A numerical, empirical example.

I will use for demonstrational purposes a dataset that has been put together and analyzed by Chin *et* al (2003)[6]. It consists of observations on 12 indicators for 250 people. The indicators can be partitioned into three groups, each group measuring one construct: the *perceived-usefulness*, with six indicators represented by *x,* the *mediator* index *enjoyment,* with three indicators, *m,* and the *intention-to-use* a particular IT-application*,* with also three indicators, represented by *y.* One research goal was to ascertain whether there is some evidence for interaction between *perceived- usefulness* and *enjoyment*. Here I will impose the following 'inner structural equation', a *quadratic (regression) model with interaction*:

$$I_y = \beta_0 + \beta_x * I_x + \beta_m * I_m + \beta_{x^2} * I_x^{\,2} + \beta_{m^2} * I_m^{\,2} + \beta_{xm} * I_x * I_m + residual$$

where the β's are scalars. I_x is a *standardized* (elastic) index for x, and similarly for I_m en I_y, so all have zero mean and unit standard deviation. Note that the indices may have different power coefficients. Incidentally, it can be argued that a latent variable model with the same path diagram is 'never' consistent with the path diagram in terms of indices, see Dijkstra & Henseler (2008). This is similar to principal components versus factor models: a principal component can 'never' fully replace a latent variable. Perhaps one could say that PLS does not model the data in the classical sense that (structural) restrictions are imposed, but that it constructs 'prescriptions' for (predictive) use of the data[7].

For a *full range analysis* a large number of weight vectors are generated independently from the maximum entropy distribution on the relevant simplexes, independently for each set of indicators; for each index powers are generated independently from the maximum entropy distribution on $[0, \infty)$ with mean equal to one[8] (recall that p=1 produces the simple weighted average). For each set of weights and powers the inner regression parameters were calculated, including the R-squared, as well as the loadings, defined as the correlations between indicators and corresponding indices. The ensuing histogram[9] and some optimizations yielded an interesting story. We note a few salient points:

- The R-squared ranged from .38 to .56. The best weighted averages (so with p's equal to one) got very close to the maximum value: .55, as did Wold's adjusted mode B (see below). So the additional flexibility due to the powers led here to just a small improvement in the fit.
- Optimizing the fit over both weights and powers revealed that the graph of R-squared for the data at hand has a rather flat maximum. A global optimum appeared to be ill-

[6] I am grateful for permission to use this dataset. It was also employed by Dijkstra & Henseler (2008). I recoded one of the indicators ('ENJ2') for *enjoyment*, thereby aligning the scores with the emotional value, consistent with the other indicators. As a result, *all* correlations between the indicators are positive.

[7] See the in many ways seminal paper by Stone (1974) for the introduction of 'prescriptions' and a discussion.

[8] If U is uniformly distributed on [0, 1] then $p = -\log(U)$ has the maximum entropy distribution with the required mean. To obtain a realization of a uniform (: maximum entropy) distribution on the unit simplex in R^N, whose coordinates are non-negative and sum to one, proceed as follows: take N-1 independent random drawings from a uniform distribution on [0, 1], append the numbers 0 and 1, sort the N+1 numbers so obtained in ascending order, and finally, calculate the latter's first differences.

[9] I generated 100.000 parameter estimates. The bootstrap reported below was rather time consuming, so 400 bootstrap samples were generated and for each sample 400 parameter estimates. Repeating this once yielded very similar results.

defined. For a number of given, fixed weights the optimized powers were well-defined and always below one. In other words, statistically optimal elastic indices tended to assign a high value when just a few important indicators scored high.

- The interaction coefficient β_{xm} was always negative: it ranged from -.31 to -.08. So if the goal was to find indices that support negative interaction, any choice would do! A shortest 95%-confidence interval for the upper bound of the shortest interval covering 95% of the full range distribution, based on a bootstrap analysis, was [-.27, -.04]. In other words, for 'almost all' samples, 'almost all' indices yield a negative interaction coefficient.
- Loadings of *x*- and *y*-indicators are relatively high for almost all methods. The mediator indicators however are relatively weak with unstable loadings.
- Equally weighted averages as indices, which are proper as well, produced quite acceptable values for all parameters. In particular, the R-squared of .50 is the 73rd percentile of the full range distribution, and its interaction coefficient of -.20 is the 30th percentile. In addition, the loadings are quite high. A related study by Dijkstra & Henseler (2008) found that this 'no-brainer' was very hard to beat in terms of prediction, as tested by cross-validation for the *y*-indicators. Perhaps one should always include this simple contender and test more sophisticated alternatives against it.

Table 1 collects some information about the variation, both across the parameter space and across the sample space, assuming the data at hand are a random sample from a stable population. The smaller numbers in parentheses give the bootstrap based shortest 95% (basic) *confidence* intervals for the upper and lower bounds, as an indication of the sampling uncertainty:

Table 1: Lower and upper bounds of the shortest 95%-intervals of the full range distribution.

	Lower bound	Upper bound
β_x	.42 (.34, .53)	.53 (.42, .61)
β_m	.11 (.02, .19)	.28 (.19, .35)
β_{x^2}	-.05 (-.13, .07)	.05 (-.10, .14)
β_{m^2}	.01 (-.05, .09)	.15 (.06, .19)
β_{xm}	-.25 (-.33, -.13)	-.13 (-.27, -.04)
R^2	.42 (.31, .57)	.53 (.37, .61)

So the upper bound of the shortest 95%-interval for the interaction coefficients based on the given data is minus.13, and its 95% confidence interval is (-.27, -.04). As one can see the sampling variation is still substantial. But the signs of the coefficients are robust with the exception of the coefficient of the squared *perceived- usefulness* index I_x , which might as well be replaced by zero. In fact, as far is prediction of the y-indicators is concerned, the squared *mediator* index I_m could be deleted also. But the interaction term turned out to be of real value for prediction.

Finally we note that Wold's suggestion in Wold (1982) for the incorporation of the squared value of a latent variable in addition to a linear term, can be easily adjusted to include the full quadratic specification with interaction. In case of mode B, the weights for I_x and for I_m are obtained by a regression of I_y on *x* and *m* respectively, and the weights for I_y by a regression of the sign-weighted sum of all the terms in the quadratic equation on *y*. For the data at hand

the cycles of regressions quickly converge (about 6 iterations for an accuracy of .000001 with very little variation in the iterations for random starting points), but not all of the weights are non-negative. However, replacing the regressions by MATLAB's *lsqnonneg.m* routine yielded with similar ease proper indices, with the same exclusions of indicators as a full-blown maximization via MATLAB's *fmincon.m* of the R-squared for proper indices. The numerical results for the latter were rather similar to Wold's adjusted approach with an R-squared equal to .55. Since all correlations between the indicators were positive all loadings were positive as well, so the constructed indices were 'proper'. If not all correlations are non-negative, proper indices can be obtained by a full-blown optimization of the R-squared subject to appropriate constraints. How to best ensure properness for PLS in general is still an open question.

References

[1] Chin W.W., Marcolin B.L., Newsted P.R. (2003) A partial least squares latent variable modeling approach for measuring interaction effects, *Information Systems Research,* 14(2), 189-217.

[2] Dijkstra T.K. (1981) *Latent variables in linear stochastic models: reflections on 'Maximum Likelihood' and 'Partial Least Squares' methods,* PhD thesis, Groningen.

[3] Dijkstra T.K. (2005) A note on the construction of indices from indicators, working paper 3-1-2005, University of Groningen.

[4] Dijkstra T.K. (2007) Latent variables and indices: Herman Wold's basic design and partial least squares, in: Vinzi V.E., Chin W.W., Henseler J., Wang H. (Eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications. Vol II of Computational Statistics.* Springer Verlag, Heidelberg (in print 2009).

[5] Dijkstra T.K. (2008) 'Equally important' criteria and demanding balanced aggregation, working paper University of Groningen, 1-5-2008. Submitted.

[6] Dijkstra T.K., Henseler J. (2008) Prescriptions for dimension reduction, with interacting factors, working paper University of Groningen & Radboud University Nijmegen, 4-6-2008. A shortened version is accepted for publication in *Quality & Quantity.*

[7] Grabisch M., Murofushi T., Sugeno M. (2000) *Fuzzy Measures and Integrals,* Physica-Verlag, Heidelberg.

[8] Stone M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions,*Journal of the Royal Statistical Society, Series B,* 36, 111-147.

[9] Tenenhaus M., Vinzi V.E., Chatelin Y.M., Lauro C. (2005) PLS Path Modelling, *Computational Statistics and Data Analysis,* 48(1), 159-205.

[10] Wold H.O.A. (1966) Nonlinear estimation by iterative least squares procedures, in: *Research Papers in Statistics, Festschrift for J. Neyman*, David F.N. (ed.), Wiley, New York, 411-444.

[11] Wold H.O.A. (1975) Path models with Latent Variables: The NIPALS Approach, in: *Quantitative Sociology,* Blalock H.M. *et al* (Eds.) Academic Press, New York, 307-359.

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

[13] Yager R.R. (1988) On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision making, *IEEE Transactions on Systems, Man and Cybernetics,*18, 183-190.

[14] Yager R.R. (1996) Quantifier Guided Aggregation Using OWA Operators, *International Journal of Intelligent Systems,* 11, 49-73.

[15]Yager R.R. (1999) Criteria Aggregations Functions Using Fuzzy Measures and the Choquet Integral, *International Journal of Fuzzy Systems,* 1, 96-112.