Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares

Theo K. Dijkstra[1][2]

[1] SNS Asset Management, Research & Development, Pettelaarpark 120, P. O. Box 70053, 5201 DZ 's-Hertogenbosch, The Netherlands, theo.dijkstra@snsam.nl
[2] University of Groningen, Econometrics & OR, WSN-gebouw, P. O. Box 800, 9700 AV Groningen, The Netherlands

1 Introduction

Partial Least Squares is a family of regression based methods designed for the analysis of high dimensional data in a low-structure environment. Its origin lies in the sixties, seventies and eighties of the previous century, when Herman O. A. Wold vigorously pursued the creation and construction of models and methods for the social sciences, where 'soft models and soft data' were the rule rather than the exception, and where approaches strongly oriented at prediction would be of great value. The author was fortunate to witness the development firsthand for a few years. Herman Wold suggested (in 1977) to write a PhD-thesis on LISREL versus PLS in the context of latent variable models, more specifically of '*the basic design*'. I was invited to his research team at the Wharton School, Philadelphia, in the fall of 1977. Herman Wold also honoured me by serving on my PhD-committee as a distinguished and decisive member. The thesis was finished in 1981. While I moved into another direction (specification, estimation and statistical inference in the context of model uncertainty) PLS sprouted very fruitfully in many directions, not only as regards theoretical extensions and innovations (multilevel, nonlinear extensions et cetera) but also as regards applications, notably in chemometrics, marketing, and political sciences. The PLS regression oriented methodology became part of main stream statistical analysis, as can be gathered from references and discussions in important books and journals. See e. g. T. Hastie, R. Tibshirani, J. Friedman (2001), or M. Stone & R. J. Brooks (1990), I. E. Frank & J. H. Friedman (1993), M. Tenenhaus et al. (2005), there are many others. This chapter will not cover these later developments, others are much more knowledgeable and are more up-to-date than I am. Instead we will go back in time and return to one of the real starting points of PLS: *the basic design*. We will look at PLS here as a method for structural equation modelling and estimation, as in Tenenhaus et al. (2005). Although I cover ground common to the latter's review I also offer additional insights, in particular into the distributional assumptions behind the basic design, the convergence of the algorithms and the properties of their outcomes. In addition, ways are suggested to modify the outcomes for the tendency to over- or underestimate loadings and correlations. Although I draw from my work from the period 1977-1981, which, as the editor graciously suggested is still of some value and at any rate is not particularly well-known, but I also suggest new developments, by stepping away from the latent variable paradigm and returning to the formative years of PLS, where principal components and canonical variables were the main source of inspiration.

In the next section we will introduce the basic design, somewhat extended beyond its archetype. It is basically a second order factor model where each indicator is directly linked to one latent variable only. Although the model is presented as 'distribution free' the very fact that conditional expectations are always assumed to be linear does suggest that multinormality is lurking somewhere in the background. We will discuss this in section 3, where we will also address the question whether normality is important, and to what extent, for the old 'adversary' LISREL. Please note that as I use the term LISREL it does not stand for a specific well-known statistical software package, but for the maximum likelihood estimation and testing approach for latent variable models, under the *working hypothesis* of multivariate normality. There is no implied value judgement about other approaches or packages that have entered the market in the mean time. In section 3 we also recall some relevant estimation theory for the case where the structural specification is incorrect or the distributional assumptions are invalid.

The next section, number 4, appears to be the least well-known. I sketch a proof there, convincingly as I like to believe, that the PLS algorithms will converge from arbitrary starting points to unique solutions, fixed points, with a probability tending to one when the sample size increases and the sample covariance matrix has a probability limit that is compatible with the basic design, or is sufficiently close to it.

In section 5 we look at the values that PLS attains at the limit, in case of the basic design. We find that correlations between the latent variables will be *under*estimated, that this is also true for the squared mutiple correlation coefficients for regressions among latent variables, and the consequences for the estimation of the structural form parameters are indicated; we note that loadings counterbalance the tendency of correlations to be underestimated, by *over*estimation. I suggest ways to correct for this lack of consistency, in the probabilistic sense.

In the section 6, we return to what I believe is the origin of PLS: the construction of indices by means of linear compounds, in the spirit of principal components and canonical variables. This section is really new, as far as I can tell. It is shown that for any set of indicators there always exist *proper indices*, i. e. linear compounds with non-negative coefficients that have non-negative correlations with their indicators. I hint at the way constraints, implied by the path diagram, can be formulated as side conditions for the construction of indices. The idea is to take the indices as the fundamental objects, as the carriers or conveyers of information, and to treat path diagrams as relationships between the indices in their own right. Basically, this approach calls for the replacement of fullblown unrestricted principal component or generalized canonical variable analyses by the construction of proper indices, satisfying modest, ‘theory poor’ restrictions on their correlation matrix. This section calls for further exploration of these ideas, acknowledging that in the process PLS’s simplicity will be substantially reduced.

The concluding section 7 offers some comments on Roderick P. McDonald(1996)’s thought provoking paper on PLS; the author gratefully acknowledges an unknown referee’s suggestion to discuss some of the issues raised in this paper.

2 A second order factor model, the ‘basic design’

Manifest variables, or indicators, are observable variables who are supposed to convey information about the behavior of *latent* variables, theoretical concepts, who are not directly observable but who are fundamental to the scientific enterprise in almost any field, see A.Kaplan(1946). In the social sciences *factor models* are the vehicle most commonly used for the analysis of the interplay between latent and manifest variables. Model construction and estimation used to be focussed mainly on the specification, validation and interpretation of factor loadings and underlying factors (latent variables), but in the seventies of the previous century the relationships between the factors themselves became a central object of study. The advent of optimization methods for high-dimensional problems, like the Fletcher-Powell algorithm, see J. M. Ortega & W. C. Rheinboldt (1970) e. g., allowed research teams to develop highly flexible and user-friendly software for the analysis, estimation and testing of second order factor models, in which relationships between the factors themselves are explicitly incorporated. First Karl G. Jöreskog from Uppsala, Sweden, and his associates developed LISREL, then later, in the eighties, Peter M. Bentler from UCLA designed EQS, and others followed. However, approaches like LISREL appeared to put high demands on the specification of the theoretical relationships: one was supposed to supply a lot of structural information on the theoretical covariance matrix of the indicators. And also it seemed that, ideally, one needed plenty of independent observations on these indicators from a multinormal distribution! Herman O. A. Wold clearly saw the potential of these methods for the social sciences but objected to their informational and distributional demands, which he regarded as unrealistic for many fields of inquiry, especially in the social sciences. Moreover, he felt that estimation and description had been put into focus, at the expense of prediction. Herman Wold had a lifelong interest in the development of predictive and robust statistical methods. In econometrics he pleaded forcefully for ‘recursive modelling’ where every single equation could be used for prediction and every parameter had a predictive interpretation, against the current of mainstream 'simultaneous equation modelling'. For the latter type of models he developed the Fix-Point estimation method, based on a predictive reinterpretation and rewriting of the models, in which the parameters were estimated iteratively by means of simple regressions. In 1966 this approach was extended to principal components, canonical variables and factor analysis models: using least squares as overall predictive criterion, parameters were divided into subsets in such a way that with any one of the subsets kept fixed at previously determined values, the remaining set of parameters would solve a straightforward regression problem; roles would be reversed and the regressions were to be continued until consecutive values for the parameters differed less then a preassigned value, see Wold (1966) but also Wold (1975). The finalizations of the ideas, culminating into PLS, took place in 1977, when Herman Wold was at the Wharton School, Philadelphia. Incidentally, since the present author was a member of Herman Wold's research team at the Wharton School in Philadelphia in the fall of 1977, one could be tempted to believe that he claims some of the credit for this development. In fact, if anything, my attempts to incorporate structural information into the estimation process, which complicated it substantially, urged Herman Wold to intensify his search for further simplification. I will try to revive my attempts in the penultimate section...

For analytical purposes and for comparisons with LISREL-type of alternatives Herman Wold put up a second order factor model, called the 'basic design'. In the remainder of this section we will present this model, somewhat extended, i. e. with fewer assumptions. The next section then takes up the discussion concerning the 'multivariate normality of the vector of indicators', the hard or 'heroic' assumption of LISREL as Herman Wold liked to call it. Anticipating the drift of the argument: the difference between multinormality and the distributional assumptions in PLS is small or large depending on whether the distance between independence and zero correlation is deemed small or large. Conceptually, the difference *is* large, since two random vectors X and Y are independent if and only if 'every' real function of X is uncorrelated with 'every' real function of Y, not just the linear functions. But any one who has ever given a Stat1 course knows that the psychological distance is close to negligible...

More important perhaps is the fact that multinormality and independence of the observational vectors is *not* required for consistency of LISREL-estimators, all that is needed is that the sample covariance matrix S is a consistent estimator for the theoretical covariance matrix Σ. The existence of Σ and independence of the observational vectors is more than sufficient, there is in fact quite some tolerance for dependence as well. Also, asymptotic normality of the estimators is assured without the assumption of multinormality. All that is needed is asymptotic normality of S, and that is quite generally the case. Asymptotic optimality, and a proper interpretation of calculated standard errors as standard errors, as well as the correct use of test-statistics however does indeed impose heavy restrictions on the distribution, which make the distance to multinormality, again psychologically spoken, rather small, and therefore to PLS rather large...

There is however very little disagreement about the difference in structural information, PLS is much more modest and therefore more realistic in this regard than LISREL. See T. K. Dijkstra (1983, 1988, 1992) where further restrictions, relevant for *both* approaches, for valid use of frequentist inference statistics are discussed, like the requirement that the model was *not* specified interactively, using the data at hand.

Now for the 'basic design'.

We will take all variables to be centered at their mean, so the expected values are zero, and we assume the existence of all second order moments. Let η be a vector of latent variables which can be partitioned in a subvector $\eta_{\mathbf{n}}$ of e**n**dogenous latent variables and a subvector $\eta_{\mathbf{x}}$ of e**x**ogenous latent variables. These vectors obey the following set of structural equations with conformable matrices B and $\mathit{\Gamma}$ and a (residual) vector ζ with the property that $E\left(\zeta \mid \eta_{\mathbf{x}}\right) = 0$:

$$\eta_{\mathbf{n}} = B\eta_{\mathbf{n}} + \mathit{\Gamma}\eta_{\mathbf{x}} + \zeta \tag{1}$$

The inverse of $(I - B)$ is assumed to exist, and the (zero-) restrictions on B, $\mathit{\Gamma}$ and the covariance matrices of η_x and ζ are sufficient for identification of the structural parameters. An easy consequence is that

$$E\left(\eta_{\mathbf{n}} \mid \eta_{\mathbf{x}}\right) = (I - B)^{-1}\,\mathit{\Gamma}\eta_{\mathbf{x}} \equiv \mathit{\Pi}\eta_{\mathbf{x}} \tag{2}$$

which expresses the intended use of the reduced form, prediction, since no function of $\eta_{\mathbf{x}}$ will predict $\eta_{\mathbf{n}}$ better than $\mathit{\Pi}\eta_{\mathbf{x}}$ in terms of mean squared error. Note that the original basic design is less general, in the sense that B is sub-diagonal there and that for each i larger than 1 the conditional expectation of ζ_i given $\eta_{\mathbf{x}}$ and the first $i-1$ elements of $\eta_{\mathbf{n}}$ is zero. In other words, originally the model for the latent variables was assumed to be a *causal chain*, where every equation, whether from the reduced or the structural form, has a predictive use and interpretation.

Now assume we have a vector of indicators y which can be divided into subvectors, one subvector for each latent variable, such that for the i-th subvector y_i the following holds:

$$y_i = \lambda_i \eta_i + \epsilon_i \tag{3}$$

where λ_i is a vector of loadings, with as many components as there are indicators for η_i, and the vector ϵ_i is a random vector of measurement errors. It is assumed that $E\left(y_i \mid \eta_i\right) = \lambda_i \eta_i$ so that the errors are uncorrelated with the latent variable of the same equation. Wold assumes that measurement errors relating to different latent variables are uncorrelated as well. In the original basic design he assumes that the elements of each ϵ_i are mutually uncorrelated, so that their covariance matrix is diagonal. We will postulate instead that $V_i \equiv E\epsilon_i\epsilon_i^\top$ has at least one zero element (or equivalently, with more than one indicator, because of the symmetry and the fact that is a covariance matrix, at least two zero elements). To summarize:

$$\Sigma_{ij} \equiv Ey_iy_j^\top = \rho_{ij}\lambda_i\lambda_j \text{ for } i \neq j \tag{4}$$

where ρ_{ij} stands for the correlation between η_i and η_j, adopting the convention that latent variables have unit variance, and

$$\Sigma_{ii} = \lambda_i\lambda_i^\top + V_i. \tag{5}$$

So the ρ_{ij}'s and the loading vectors describe the correlations at the first level, of the indicators, and the structural equations yield the correlations at the second level, of the latent variables. It is easily seen that all parameters are identified: equation (4) determines the direction of λ_i apart from a sign factor and (5) fixes its length, therefore the ρ_{ij}'s are identified (as well as the V_i's), and they on their turn allow determination of the structural form parameters, given Σ of course.

3 Distributional assumptions: multinormality or 'distribution free'?

The (extended) basic design does not appear to impose heavy constraints on the distribution of the indicators: the existence of second order moments, some zero conditional expectations and a linear structure, that's about it. Multinormality seems conceptually way off. But let us take an arbitrary measurement equation

$$y_i = \lambda_i\eta_i + \epsilon_i \tag{6}$$

and instead of assuming that $E\left(\epsilon_i \mid \eta_i\right) = 0$, we let ϵ_i and η_i be stochastically independent, which *implies* a zero conditional expectation. As Wold assumes the elements of ϵ_i to be uncorrelated, let us take them here mutually independent. For $E\left(\eta_i \mid y_i\right)$ we take it to be linear as well, so assuming here and in the sequel invertibility of matrices whenever this is needed

$$E\left(\eta_i \mid y_i\right) = \lambda_i^\top \left(\Sigma_{ii}\right)^{-1} y_i \propto \lambda_i^\top V_i^{-1} y_i \tag{7}$$

If now all loadings, all elements of λ_i, differ from zero, we *must* have multinormality of the vector $(y_i; \eta_i; \epsilon_i)$ as follows from a characterization theorem in Kagan et al (1973), see in particular theorem 10.5.3. Let us modify and extend each measurement equation as just described, and let all measurement errors be mutually independent. Then for one thing each element of η will be normal and ϵ, the vector obtained by stacking the ϵ_i's, will be multinormal.

If we now turn to the structural equations, we will take for simplicity the special case of a *complete causal chain*, where B is square and lower diagonal and the elements of the residual vector ζ are mutually independent. A characterization due to Cramér states that when the sum of independent variables is normal, all constituents of this sum are normal, and Cramér and Wold have shown that a vector is multinormal if and only if every linear function of this vector is normal. Combining these characterizations one is easily led to the conclusion that $(y; \eta; \zeta; \epsilon)$ is multinormal. See T. K. Dijkstra (1981) for a more elaborate discussion and other results.

So, roughly, if one strengthens zero conditional expectations to independence and takes all conditional expectations to be linear, one gets multinormality. It appears that psychologically PLS and multinormality are not far apart. But the appreciation of these conditions is not just a matter of taste, or of mathematical/statistical maturity. Fundamentally it is an empirical matter and the question of their (approximate) validity ought to be settled by a thorough analysis of the data. If one has to reject them, how sad is that? The linear functions we use for prediction are then no longer least squares optimal in the set of *all* functions, but best linear approximations only to these objects of desire (in the population,that is). If we are happy with linear approximations, i. e. we understand them *and* can use them to good effect, then who cares about multinormality, or for that matter about linearity of conditional expectations? In the author's opinion, normality has a pragmatic justification only. Using it as a working hypothesis in combination with well worn 'principles', like least squares or, yes, maximum likelihood, often leads to useful results, which as a bonus usually satisfy appealing consistency conditions.

It has been stated and is often repeated, seemingly thoughtlessly, that LISREL is *based* on normality, in the sense that its use *requires* the data to be normally distributed. This is a prejudice that ought to be cancelled. One can use the maximum entropy principle, the existence of second order moments, and the likelihood principle to motivate the choice of the fitting function that LISREL employs. But at the end of the day this function is just one way of fitting a theoretical covariance matrix $\Sigma(\theta)$ to a sample covariance matrix S, where the fit is determined by the difference between the eigenvalues of $S\Sigma^{-1}$ and the eigenvalues of the identity matrix. To elaborate just a bit:
If we denote the p eigenvalues of $S\Sigma^{-1}$ by $\gamma_1, \gamma_2, \ldots, \gamma_p$ the LISREL fitting-function can be written as $\sum_{i=1}^{i=p} (\gamma_i - \log \gamma_i - 1)$. Recall that for real positive numbers $0 \leq x - \log x - 1$ everywhere with equality only for $x = 1$. Therefore the LISREL criterion is always nonnegative and zero only when *all* eigenvalues are equal to 1. The absolute minimum is reached if and only if a θ can be found such that $S = \Sigma(\theta)$. So if $S = \Sigma(\theta_*)$ for some θ_* and identifiability holds, LISREL will find it. Clearly, other functions of the eigenvalues will do the trick, GLS is one of them. See T. K. Dijkstra (1990) for an analysis of the class of Swain functions. The 'maximum likelihood' estimator $\widehat{\theta}$ is a well-behaved, many times differentiable function of S, which yields θ when evaluated at $S = \Sigma(\theta)$. In other words, if S is close to $\Sigma(\theta)$ the estimator is close to θ and it is locally a linear function of S. It follows that when S tends

in probability to its 'true value', $\Sigma(\theta)$, then $\widehat{\theta}$ will do the same and moreover, if S is asymptotically normal, then $\widehat{\theta}$ is.

Things become more involved when the probability limit of S, *plim*(S), does *not* satisfy the structural constraints as implied by the second order factor model at hand, so there is *no* θ for which $\Sigma(\theta)$ equals *plim*(S). We will summarize in a stylized way what can be said about the behavior of estimators in the case of Weighted Least Squares, which with proper weighting matrices include LISREL, i. e. maximum likelihood under normality, and related fitting functions as well. The result will be relevant also for the analysis of reduced form estimators using PLS.

To simplify notation we will let $\sigma(\theta)$ stand for the vector of non-redundant elements of the smooth matrix function $\Sigma(\theta)$ and s does the same for S. We will let $\overline{s}$ stand for *plim*(S). Define a fitting function $F(s, \sigma(\theta) \mid W)$ by

$$F(s, \sigma(\theta) \mid W) \equiv (s - \sigma(\theta))^{\top} W (s - \sigma(\theta)) \tag{8}$$

where W is some symmetric random matrix of appropriate order whose *plim*, $\overline{W}$, exists as a positive definite matrix (non-random matrices can be handled as well). The vector θ varies across a suitable set, non-empty and compact or such that F has a compact level set. We postulate that the minimum of $F\left(\overline{s}, \sigma(\theta) \mid \overline{W}\right)$ is attained in a unique point $\theta\left(\overline{s}, \overline{W}\right)$, depending on the probability limits of S and W. One can show that F tends in probability to $F\left(\overline{s}, \sigma(\theta) \mid \overline{W}\right)$ *uniformly* with respect to θ. This implies that the estimator $\widehat{\theta}(s, W) \equiv \arg\min(F)$ will tend to $\theta\left(\overline{s}, \overline{W}\right)$ in probability. Different fitting functions will produce different probability limits, if the model is incorrect. With sufficient differentiability and asymptotic normality we can say more (see T. K. Dijkstra (1981) e. g.), using the implicit function theorem on the first-order conditions of the minimization problem. In fact, when

$$\sqrt{n}\begin{bmatrix} (s - \overline{s}) \\ vec\left(W - \overline{W}\right) \end{bmatrix} \longrightarrow \mathbf{N}\left(0, \begin{bmatrix} V_{ss} & V_{sw} \\ V_{ws} & V_{ww} \end{bmatrix}\right) \tag{9}$$

where n is the number of observations, *vec* stacks the elements columnwise and the convergence is in distribution to the normal distribution, indicated by $\mathbf{N}$, and we define:

$$\Delta \equiv \partial\sigma / \partial\theta^{\top} \tag{10}$$

evaluated at $\theta\left(\overline{s}, \overline{W}\right)$, and M is a matrix with typical element M_{ij}:

$$M_{ij} \equiv \left[\partial^2 \sigma^{\top} / \partial\theta_i \partial\theta_j\right] W [\sigma - \overline{s}] \tag{11}$$

and $\widetilde{V}$ equals by definition

$$\left[\Delta^{\top}\overline{W}, [\overline{s} - \sigma]^{\top} \otimes \Delta^{\top}\right] \begin{bmatrix} V_{ss} & V_{sw} \\ V_{ws} & V_{ww} \end{bmatrix} \begin{bmatrix} \overline{W}\Delta \\ [\overline{s} - \sigma] \otimes \Delta \end{bmatrix} \tag{12}$$

with σ and its partial derivatives in M and $\widetilde{V}$ also evaluated at the same point $\theta\left(\overline{s},\overline{W}\right)$, then we can say that $\sqrt{n}\left(\widehat{\theta}\left(s,W\right)-\theta\left(\overline{s},\overline{W}\right)\right)$ will tend to the normal distribution with zero mean and covariance matrix $\varOmega$, say, with

$$\varOmega\equiv\left(\varDelta^{\top}\overline{W}\varDelta+M\right)^{-1}\widetilde{V}\left(\varDelta^{\top}\overline{W}\varDelta+M\right)^{-1}. \tag{13}$$

This may appear to be a somewhat daunting expression, but it has a pretty clear structure. In particular, observe that if $\overline{s}=\sigma\left(\theta\left(\overline{s},\overline{W}\right)\right)$, in other words, $\underset{\sim}{if}$ the structural information contained in $\varSigma$ is correct, then M becomes 0 and $\widetilde{V}$ which sums 4 matrices looses 3 of them, and so the asymptotic covariance of the estimator $\widehat{\theta}\left(s,W\right)$ reduces to:

$$\left(\varDelta^{\top}\overline{W}\varDelta\right)^{-1}\varDelta^{\top}\overline{W}V_{ss}\overline{W}\varDelta\left(\varDelta^{\top}\overline{W}\varDelta\right)^{-1} \tag{14}$$

which simplifies even further to

$$\left(\varDelta^{\top}V_{ss}^{-1}\varDelta\right)^{-1} \tag{15}$$

when $\overline{W}=V_{ss}^{-1}$. In the latter case we have asymptotic efficiency: no other fitting function will produce a smaller asymptotic covariance matrix. LISREL belongs to this class, provided the structure it implicitly assumes in V_{ss} is correct. More precisely, it is sufficient when the element in V_{ss} corresponding with the asymptotic covariance between s_{ij} and s_{kl} equals $\sigma_{ik}\sigma_{jl}+\sigma_{il}\sigma_{jk}$. This is the case when the underlying distribution is multinormal. Elliptical distributions in general will yield an asymptotic covariance matrix that is proportional to the normal V_{ss}, so they are efficient as well. The author is unaware of other suitable distributions. So LISREL rests for inference purposes on a *major* assumption, that is in the opinion of the author not easily met. If one wants LISREL to produce reliable standard errors, one would perhaps be well advised to use the bootstrap. By the way, there are many versions of the theorem stated above in the literature, the case of a correct model is particularly well covered. In fact, we expect the results on asymptotic efficiency to be so well known that references are redundant.

To summarize, if the model is correct in the sense that the structural constraints on $\varSigma$ are met, and S is consistent and W has a positive definite probability limit then the classical fitting functions will produce estimators that tend in probability to the true value. If the model is not correct, they will tend to the best fitting value as determined by the particular fitting function chosen. The estimators are normal, asymptotically, when S and W are (jointly), whether the structural constraints are met or not. Asymptotic efficiency is the most demanding property and is not to be taken for granted. A truly major problem that we do not discuss is model uncertainty, where the model itself is random due to the interaction between specification, estimation and validation on the same data set, with hunches taken from the data to improve the model. This wreaks havoc on the standard approach. No statistics school really knows how to deal with this. See for discussions e. g. Leamer (1978), T. K. Dijkstra (1988) or T. Hastie et al. (2001).

In the next sections we will see that under the very conditions that make LISREL consistent, PLS is not consistent, but that the error will tend to zero when the quality of the estimated latent variables, as measured by their correlation with the true values, tends to 1 by increasing the number of indicators per latent variable.

4 On the PLS-algorithms: convergence issues and functional properties of fixed points

The basic approach in PLS is to construct *proxies* for the latent variables, in the form of linear compounds, by means of a sequence of alternating least squares algorithms, each time solving a local, linear problem, with the aim to extract the predictive information in the sample. Once the compounds are constructed, the parameters of the structural and reduced form are estimated with the proxies replacing the latent variables. The particular information embodied in the structural form is not used explicitly in the determination of the proxies. The information actually used takes the presence or absence of variables in the equations into account, but not the implied zero constraints and multiplicative constraints on the reduced form (:the classical rank constraints on submatrices of the reduced form as implied by the structural form).

There are two basic types of algorithms, called *mode A* and *mode B*, and a third type, *mode C*, that mixes these two. Each mode generates an estimated weight vector $\widehat{w}$, with typical subvector $\widehat{w}_i$ of the same order as y_i. These weight vectors are fixed points of mappings defined algorithmically. If we let S_{ij} stand for the sample equivalent of Σ_{ij}, and $sign_{ij}$ for the sign of the sample correlation between the estimated proxies $\widehat{\eta}_i \equiv \widehat{w}_i^{\top} y_i$ and $\widehat{\eta}_j \equiv \widehat{w}_j^{\top} y_j$, and C_i is the index set that collects the labels of latent variables which appear at least once on different sides of the structural equations in which η_i appears, we have for *mode A:*

$$\widehat{w}_i \propto \sum_{j \epsilon C_i} sign_{ij} \cdot S_{ij} \widehat{w}_j \text{ and } \widehat{w}_i^{\top} S_{ii} \widehat{w}_i = 1. \tag{16}$$

As is easily seen the i-th weight vector is obtainable by a regression of the i-th subvector of indicators y_i on the scalar $\widehat{a}_i \equiv \sum_{j \epsilon C_i} sign_{ij} \cdot \widehat{\eta}_j$, so the weights are determined by the ability of $\widehat{a}_i$ to predict y_i. It is immediate that when the basic design matrix Σ replaces S the corresponding fixed point $\overline{w}_i$, say, is proportional to λ_i. But note that this requires at least two latent variables. In a stand-alone situation mode A produces the first principal component, and there is no simple relationship with the loading vector. See Hans Schneeweiss and Harald Mathes (1995) for a thorough comparison of factor analysis and principal components. Mode A and principal components share a lack of scale-invariance, they are both sensitive to linear scale transformations. McDonald

(1996) has shown essentially that mode A corresponds to maximization of the sum of absolute values of the covariances of the proxies, where the sum excludes the terms corresponding to latent variables which are not directly related. The author gratefully acknowledges reference to McDonald (1996) by an unknown referee.
For *mode B* we have:

$$\widehat{w}_i \propto S_{ii}^{-1} \sum_{j \epsilon C_i} sign_{ij} \cdot S_{ij} \widehat{w}_j \text{ and } \widehat{w}_i^\top S_{ii} \widehat{w}_i = 1. \tag{17}$$

Clearly, $\widehat{w}_i$ is obtained by a regression that reverses the order compared to mode A: here $\widehat{a}_i$, defined similarly, is regressed on y_i. So the indicators are used to predict the sign-weighted sum of proxies. With only two latent variables mode B will produce the first canonical variables of their respective indicators, see Wold (1966, 1982) e. g. Mode B is a genuine generalization of canonical variables: it is equivalent to the maximization of the sum of absolute values of the correlations between the proxies, $\widehat{w}_i^\top S_{ij} \widehat{w}_j$, taking only those i and j into account that correspond to latent variables which appear at least once on different sides of a structural equation. A Lagrangian analysis will quickly reveal this. The author noted this, in 1977, while he was a member of Herman Wold's research team at the Wharton School, Philadelphia. It is spelled out in his thesis (1981). Kettenring (1971) has introduced other generalizations, we will return to this in the penultimate section. Replacing S by $\varSigma$ yields a weight vector $\overline{w}_i$ proportional to $\varSigma_{ii}^{-1}\lambda_i$, so that the 'population proxy' $\overline{\eta}_i \equiv \overline{w}_i^\top y_i$ has unit correlation with the best linear least squares predictor for η_i in terms of y_i. This will be true as well for those generalizations of canonical variables that were analyzed by Kettenring (1971). Mode B is scale-invariant, in the sense that linear scale transformations of the indicators leave $\widehat{\eta}_i$ and $\overline{\eta}_i$ undisturbed.

Mode C mixes the previous approaches: some weight vectors satisfy mode A, others satisfy mode B type of equations. As a consequence the products of mode C mix the properties of the other modes as well. In the sequel we not dwell upon this case. Suffice it to say that with two sets of indicators, two latent variables, mode C produces a variant of the well-known MIMIC-model.

Sofar we have simply assumed that the equations as stated have solutions, that they actually *have* fixed points, and the iterative procedure to obtain them has been merely hinted at. To clarify this, let us discuss a simple case first. Suppose we have three latent variables connected by just one relation $\eta_3 = \beta_{31}\eta_1 + \beta_{32}\eta_2$ plus a least squares residual, and let us use mode B. The fixed point equations specialize to:

$$\widehat{w}_1 = \widehat{c}_1 S_{11}^{-1} \cdot [sign_{13} \cdot S_{13} \widehat{w}_3] \tag{18}$$

$$\widehat{w}_2 = \widehat{c}_2 S_{22}^{-1} \cdot [sign_{23} \cdot S_{23} \widehat{w}_3] \tag{19}$$

$$\widehat{w}_3 = \widehat{c}_3 S_{33}^{-1} \cdot [sign_{13} \cdot S_{31} \widehat{w}_1 + sign_{23} \cdot S_{32} \widehat{w}_2] \,. \tag{20}$$

The scalar $\widehat{c}_i$ forces $\widehat{w}_i$ to have unit length in the metric of S_{ii}.The iterations start with arbitrary nonzero choices for the $\widehat{w}_i$'s, which are normalized as required, the *sign*-factors are determined, and a cycle of updates commences: inserting $\widehat{w}_3$ into (18) and (19) gives updated values for $\widehat{w}_1$ and $\widehat{w}_2$, which on their turn are inserted into (20), yielding an update for $\widehat{w}_3$, then new *sign*-factors are calculated, and we return to (18) et cetera. This is continued until the difference between consecutive updates is insignificant. Obviously, this procedure allows of small variations, but they have no impact on the results. Now define a function G, say by

$$G\left(w_3, S\right) \equiv c_3 S_{33}^{-1} \cdot \left[c_1 S_{31} S_{11}^{-1} S_{13} + c_2 S_{32} S_{22}^{-1} S_{23}\right] \cdot w_3 \tag{21}$$

where c_1 is such that $c_1 S_{11}^{-1} S_{13} w_3$ has unit length in the metric of S_{11}, c_2 is defined similarly, and c_3 gives G unit length in the metric of S_{33}. Clearly G is obtained by consecutive substitutions of (18) and (19) into (20). Observe that:

$$G\left(w_3, \Sigma\right) = \overline{w}_3 \tag{22}$$

for every value of w_3 (recall that $\overline{w}_3 \propto \Sigma_{33}^{-1} \lambda_3$). A very useful consequence is that the derivative of G with respect to w_3, evaluated at $\left(\overline{w}_3, \Sigma\right)$ equals *zero*. Intuitively, this means that for S not too far away from Σ, $G\left(w_3, S\right)$ maps two different vectors w_3 ,which are not too far away from $\overline{w}_3$, on points which are *closer* together than the original vectors. In other words, as a function of w_3, $G\left(w_3, S\right)$ will be a local contraction mapping. With some care and an appropriate mean value theorem one may verify that our function does indeed satisfy the conditions of Copson's *Fixed point theorem with a parameter*, see E. T. Copson (1979), sections 80-82. Consequently, G *has* a unique fixed point $\widehat{w}_3\left(S\right)$ in a neighborhood of $\overline{w}_3$ for every value of S in a neighborhood of Σ, and it *can* be found by successive substitutions: for an arbitrary starting value sufficiently close to $\overline{w}_3$ the ensuing sequence of points converges to $\widehat{w}_3\left(S\right)$ which satisfies $\widehat{w}_3\left(S\right) = G\left(\widehat{w}_3\left(S\right), S\right)$. Also note that if $plim(S) = \Sigma$ then the first iterate from an arbitrary starting point will tend to $\overline{w}_3$ in probability, so if the sample is sufficiently large the conditions for a local contraction mapping will be satisfied with an arbitrarily high probability. Essentially, any choice of starting vector will do. The mapping $\widehat{w}_3\left(S\right)$ is continuous, in fact it is continuously differentiable, as follows quickly along familiar lines of reasoning in proofs of implicit function theorems. So asymptotic normality is shared with S. The other weight vectors are smooth transformations of $\widehat{w}_3\left(S\right)$, so they will be well-behaved as well.

It is appropriate now to point out that what we have done with mode B for three latent variables can also be done for the other modes, and the number of latent variables is irrelevant: reshuffle the equations (16) and (17), if necessary, so that the weights corresponding to the exogenous latent variables are listed first; we can express them in terms of the endogenous weight vectors, $w_{\mathbf{n}}$, say, so that after insertion in the equations for the latter a function $G\left(w_{\mathbf{n}}, S\right)$ can be defined with the property that $G\left(w_{\mathbf{n}}, \Sigma\right) = \overline{w}_{\mathbf{n}}$ and we proceed as before. We obtain again a well-defined fixed point $\widehat{w}\left(S\right)$ by means of successive substitutions. Let us collect this in a theorem (T. K. Dijkstra, 1981; we ignore trivial regularity assumptions that preclude loading vectors like λ_i to consist of zeros only; and similarly, we ignore the case where Σ_{ij} is identically zero for every $j \epsilon C_i$):

Theorem 1. *If* $\operatorname{plim}(S) = \Sigma$ *where* Σ *obeys the restrictions of the basic design, then the PLS algorithms will converge for every choice of starting values to unique fixed points of equations (16) and (17) with a probability tending to one when the number of sample observations tends to* ∞*. These fixed points are continuously differentiable functions of* S*, their probability limits satisfy the fixed point equations with* S *replaced by* Σ*.They are asymptotically normal when* S *is.*

As a final observation in this section: if $plim(S) = \Sigma_*$ which is *not* a basic design matrix but comes sufficiently close to it, then the PLS-algorithms will converge in probability to the fixed point defined by $\widehat{w}\left(\Sigma_*\right)$. We will again have good numerical behavior and local linearity.

5 Correlations, structural parameters, loadings

In this section we will assume without repeatedly saying so that $plim(S) = \Sigma$ for a Σ satisfying the requirements of the extended basic design except for one problem, indicated below in the text. Recall the definition of the *population proxy* $\overline{\eta}_i \equiv \overline{w}_i^\top y_i$ where $\overline{w}_i \equiv plim(\widehat{w}_i)$ depends on the mode chosen; for mode A $\overline{w}_i$ is proportional to λ_i and for mode B it is proportional to $\Sigma_{ii}^{-1}\lambda_i$. Its sample counterpart, the *sample proxy*, is denoted by $\widehat{\eta}_i \equiv \widehat{w}_i^\top y_i$. In PLS the sample proxies replace the latent variables. Within the basic design, however, this replacement can never be exhaustive unless there are no measurement errors. We can measure the quality of the proxies by means of the squared correlation between η_i and $\overline{\eta}_i : R^2\left(\eta_i, \overline{\eta}_i\right) = \left(\overline{w}_i^\top \lambda_i\right)^2$. In particular, for mode A we have

$$R_A^2\left(\eta_i, \overline{\eta}_i\right) = \frac{\left(\lambda_i^\top \lambda_i\right)^2}{\lambda_i^\top \Sigma_{ii} \lambda_i} \tag{23}$$

and for mode B:

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

as is easily checked. It is worth recalling that the mode B population proxy is proportional to the best linear predictor of η_i in terms of y_i, which is not true for mode A. Also note that the Cauchy-Schwarz inequality immediately entails that R_A^2 is always less than R_B^2 unless λ_i is proportional to $\Sigma_{ii}^{-1}\lambda_i$ or equivalently, to $V_i^{-1}\lambda_i$; for diagonal V_i this can only happen when all measurement error variances are equal. For every mode we have that

$$R^2\left(\overline{\eta}_i, \overline{\eta}_j\right) = \left(\overline{w}_i^\top \Sigma_{ij} \overline{w}_j\right)^2 = \rho_{ij}^2 \cdot R^2\left(\eta_i, \overline{\eta}_i\right) \cdot R^2\left(\eta_j, \overline{\eta}_j\right) \tag{25}$$

and we observe that in the limit the PLS-proxies will *under*estimate the squared correlations between the latent variables. This is also true of course for two-block canonical variables: they *under*estimate the correlation between the underlying latent variables eventhough they maximize the correlation between linear compounds. It is not typical for PLS of course. Methods like Kettenring's share this property. The error depends in a simple way on the quality of the proxies, with mode B performing best.

The structural bias does have consequences for the estimation of structural form and reduced form parameters as well. If we let R stand for the correlation matrix of the latent variables, $\overline{R}$ does the same for the population proxies, and K is the diagonal matrix with typical element $R\left(\eta_i, \overline{\eta}_i\right)$, we can write

$$\overline{R} = KRK + I - K^2. \tag{26}$$

So conditions of the Simon-Blalock type, like zero partial correlation coefficients, even if satisfied by R will typically not be satisfied by $\overline{R}$. Another consequence is that *squared multiple correlations* will be *under*estimated as well: the value that PLS obtains in the limit, using proxies, for the regression of η_i on other latent variables never exceeds the fraction $R^2\left(\eta_i, \overline{\eta}_i\right)$ of the 'true' squared multiple correlation coefficient. This is easily deduced from a well-known characterization of the squared multiple correlation: it is the maximum value of $1 - \beta^\top R\beta$ with respect to β where R is the relevant correlation matrix of the variables, and β is a conformable vector whose i-th component is forced to equal 1 (substitution of the expression for $\overline{R}$ quickly yields the upper bound as stated). The upper bound can be attained only when the latent variables other than η_i are measured without flaw.

In general we have that the regression matrix for the population proxies equals $\overline{\Pi}$, say, with

$$\overline{\Pi} = \overline{R}_{\mathbf{nx}} \overline{R}_{\mathbf{xx}}^{-1} = K_{\mathbf{n}} \Pi R_{\mathbf{xx}} K_{\mathbf{x}} \overline{R}_{\mathbf{xx}}^{-1} \tag{27}$$

where subscripts indicate appropriate submatrices, the definitions will be clear. Now we assumed that B and Γ could be identified from Π. It is common knowledge in econometrics that this is equivalent to the existence of rank restrictions on submatrices of Π. But since $\overline{R}$ differs from R these relations will be disturbed and $\overline{\Pi}$ will *not* satisfy them, except on sets of measure zero in the parameterspace. This makes the theory hinted at in section 3 relevant. With p replacing s, and π replacing σ for maximum similarity, if so desired, we can state that classical estimators for the structural form parameters will asymptotically center around (B_*, Γ_*) say, which are such that $\left(I - B_*\right)^{-1} \Gamma_*$ fits $\overline{\Pi}$ 'best'. 'Best' will depend on the estimation procedure chosen and $\overline{\Pi}$ varies with the mode. In principle, the well-known delta method can be used to get standard errors, but we doubt whether that is really feasible (which is something of an understatement). The author, T. K. Dijkstra (1982, 1983), suggested to use the bootstrap as a general tool. Later developments, such as the stationary bootstrap for time series data, has increased the value of the method even more, but care must be used for a proper application; in particular, one should resample the observations on the indicators, not on the sample proxies, for a decent analysis of sampling uncertainty.

Turning now to the loadings, some straightforward algebra easily yields that both modes will tend to *over*estimate them in absolute value, mode B again behaving better than mode A, in the limit that is. The loadings are in fact estimated by

$$\widehat{\lambda}_i \equiv S_{ii}\widehat{w}_i. \tag{28}$$

and the error covariance matrices can be calculated as

$$\widehat{V}_i \equiv S_{ii} - \widehat{\lambda}_i\widehat{\lambda}_i^\top. \tag{29}$$

(Note that $\widehat{V}_i\widehat{w}_i = 0$, so the estimated errors are linearly dependent, which will have some consequences for second level analyses, not covered here). Inserting population values for sampling values we get for mode A that $\overline{\lambda}_i$, the probability limit of $\widehat{\lambda}_i$, is proportional to $\Sigma_{ii}\lambda_i$. For mode B we note that $\overline{\lambda}_i$ is proportional to λ_i with a proportionality factor equal to the square root of 1 over $R^2\left(\eta_i, \overline{\eta}_i\right)$. Mode B, but not mode A, will reproduce Σ_{ij} exactly in the limit. For other results, all based on straightforward algebraic manipulations we refer to Dijkstra (1981).

So in general, not all parameters will be estimated consistently. Wold, in a report that was published as chapter 1 in K. G. Jöreskog and H. O. A. Wold(1982), introduced the auxiliary concept of 'consistency at large' which captures the idea that the inconsistency will tend to zero if more indicators of sufficient quality can be introduced for the latent variables. The condition as formulated originally was

$$\frac{\left[E\left(\overline{w}_i^\top \epsilon_i\right)^2\right]^{\frac{1}{2}}}{\overline{w}_i^\top \lambda_i} \to 0. \tag{30}$$

This is equivalent to $R^2\left(\eta_i, \overline{\eta}_i\right) \to 1$. Clearly, if these correlations are large, PLS will combine numerical expediency with consistency. If the proviso is not met in a sufficient degree the author (T. K. Dijkstra (1981)) has suggested to use some simple 'corrections'. E. g. in the case of mode B one could first determine the scalar $\widehat{f}_i$ say that minimizes, *assuming uncorrelated measurement errors*,

$$trace\left(\left[S_{ii} - diag\left(S_{ii}\right) - \left[f_i^2 \cdot \widehat{\lambda}_i\widehat{\lambda}_i^\top - diag\left(f_i^2 \cdot \widehat{\lambda}_i\widehat{\lambda}_i^\top\right)\right]\right]^2\right) \tag{31}$$

for all real f_i and which serves to rescale $\widehat{\lambda}_i$. We get

$$\widehat{f}_i^2 = \frac{\widehat{\lambda}_i^\top \left[S_{ii} - diag\left(S_{ii}\right)\right]\widehat{\lambda}_i}{\widehat{\lambda}_i^\top \left[\widehat{\lambda}_i\widehat{\lambda}_i^\top - diag\left(\widehat{\lambda}_i\widehat{\lambda}_i^\top\right)\right]\widehat{\lambda}_i}. \tag{32}$$

One can check that $\widehat{f}_i\widehat{\lambda}_i$ tends in probability to λ_i. In addition we have that $p\lim\left(\widehat{f}_i^2\right)$ equals $R_B^2\left(\eta_i, \overline{\eta}_i\right)$. So one could in principle get consistent estimators for R, the correlation matrix of the latent variables by reversing equation

(25) so to speak. But a more direct approach can also be taken by minimization of

$$trace\left\{\left[S_{ij}-r_{ij}\widehat{f}_i\widehat{f}_j\cdot\widehat{\lambda}_i\widehat{\lambda}_j^\top\right]^\top\cdot\left[S_{ij}-r_{ij}\widehat{f}_i\widehat{f}_j\cdot\widehat{\lambda}_i\widehat{\lambda}_j^\top\right]\right\} \tag{33}$$

for r_{ij}. This produces the consistent estimator

$$\widehat{r}_{ij}\equiv\frac{\widehat{\lambda}_i^\top S_{ij}\widehat{\lambda}_j}{\widehat{f}_i\widehat{f}_j\cdot\widehat{\lambda}_i^\top\widehat{\lambda}_i\cdot\widehat{\lambda}_j^\top\widehat{\lambda}_j}. \tag{34}$$

With a consistent estimator for R we can also estimate B and Γ consistently. We leave it to the reader to develop alternatives. The author is not aware of attempts in the PLS-literature to implement this idea or related approaches. Perhaps the development of second and higher order levels has taken precedence over refinements to the basic design because that just comes naturally to an approach which mimics principal components and canonical variables so strongly. But clearly, the bias can be substantial if not dramatic, whether it relates to regression coefficients, correlations, structural form parameters or loadings as the reader easily convinces himself by choosing arbitrary values for the $R^2(\eta_i,\overline{\eta}_i)$'s; even for high quality proxies the disruption can be significant, and it is parameter dependent. So if one adheres to the latent variable paradigm, bias correction as suggested here or more sophisticated approaches seems certainly to be called for.

6 Two suggestions for further research

In this section we depart from the basic design with its adherence to classical factor analysis modelling, and return so to speak to the original idea of constructing indices by means of linear compounds. We take the linear indices as the fundamental objects and we read path diagrams as representing relationships between the indices in their own right. What we try to do here is to delineate a research program that should lead to the construction of *proper indices*, more about them below, that satisfy the restrictions implied by a path diagram. In the process PLS will loose a lot of its simplicity: proper indices impose inequality restrictions on the indices, and we will no longer do regressions with sums of sign weighted indices, if we do regressions at all, but with sums that somehow reflect the pattern of relationships. The approach is highly provisional and rather unfinished.

As a general principle indicators are selected on the basis of a presumed monotonous relationship with the underlying concept: they are supposed to reflect increases or decreases in the latent variable on an empirically relevant range (without loss of generality we assume that indicators and latent variable are supposed to vary in the same direction). The ensuing *index* should mirror this: not only the *weights* (the coefficients of the indicators in the index) but also the *correlations* between the indicators and the index ought to be positive, or at least non-negative. In practice, a popular first choice for the index is the first principal component of the indicators, the linear compound that best explains total variation in the data. If the correlations between the indicators happen to be positive, Perron-Frobenius' theorem tells us that the first principal component will have positive weights, and of course it has positive correlations with the indicators as well. If the proviso is not met we cannot be certain of these appealing properties. In fact, it often happens that the first principal component is not acceptable as an index, and people resort to other weighting schemes, usually rather simple ones, like sums or equally weighted averages of the indicators. It is not always checked whether this simple construct is positively correlated with its indicators.

Here we will establish that with every non-degenerate vector of indicators is associated a set of *admissible indices*: linear compounds of the indicators with non-negative coefficients whose correlations with the indicators are non-negative. The set of admissible or *proper* weighting vectors is a convex polytope, generated by a finite set of extreme points. In a stand-alone situation, where the vector of indicators is not linked to other indicator-vectors one could project the first principal component on this convex polytope in the appropriate metric, or choose another point in the set,e.g. the point whose average squared correlation with the indicators is maximal. In the regular situation, with more than one block of manifest variables, we propose to choose weighting vectors from each of the admissible sets, such that the ensuing correlation matrix of the indices optimizes one of the distance functions suggested by Kettenring (1971), like: GENVAR (the generalized variance or the determinant of the correlation matrix), MINVAR, its minimal eigenvalue or MAXVAR, its maximal eigenvalue. GENVAR and MINVAR have to be minimized, MAXVAR maximized. The latter approach yields weights such that the total variation of the corresponding indices is explained as well as possible by one factor. The MINVAR-indices will move more tightly together than any other set of indices, in the sense that the variance of the minimum variance combination of the indices will be smaller, at any rate not larger, than the corresponding variance of any other set of indices. GENVAR is the author's favorite, it can be motivated in terms of total variation, or in terms of the volume of (confidence) ellipsoids; see T. W. Anderson (1984, in particular chapter 7.5), or F. R. Gantmacher (1977, reprint of 1959, in particular chapter IX section 5). Alternatively, GENVAR can be linked to *entropy.* The latent variables which the indices represent are supposed to be mutually informative, in fact they are analyzed together for this very reason. If we want indices that are mutually as informative as possible, we should minimize the entropy of their distribution. This is equivalent to the minimization of the determinant of their covariance or correlation matrix, if we adopt the 'most neutral' distribution for the indicators that is consistent with the existence of the second order moments: the normal distribution. (The expression 'most neutral' is a non-neutral translation of 'maximum entropy'. . .). Also, as pointed out by Kettenring (1971), the GENVAR indices satisfy an appealing *consistency* property: the index of every block, given the indices of the other blocks, is the first canonical variable of the block in question relative to the other indices; so every index has maximum multiple correlation with the vector of the other indices.

For the situation where the latent variables are arranged in a path diagram, that embodies a number of zero constraints on the structural form matrices (the matrix linking the exogenous latent variables to the endogenous latent variables, and the matrix linking the latter to each other), we suggest to optimize one of Kettenring's distance functions subject to these constraints. Using Bekker & Dijkstra (1990) and Bekker et al. (1994) the zero constraints can be transformed by symbolic calculations into zero constraints and multiplicative constraints on the regression equations linking the endogenous variables to the exogenous latent variables. In this way we can construct admissible, mutually informative indices, embedded in a theory-based web of relationships.
Now for some detail.

6.1 Proper indices

Let Σ be an arbitrary positive definite covariance or correlation matrix of a random vector X of order p by 1, where p is any natural number. We will prove that there is always a p by 1 vector w with non-negative elements, adding up to 1, such that the vector Σw that contains the covariances between X and the 'index' $w^\top X$,has no negative elements as well (note that at least one element must be positive, since the positive definiteness of Σ and the fact that the weights add up to one preclude the solution consisting of zeros only). Intuitively, one might perhaps expect such a property since the angle between any w and its image Σw is acute due to Σ's positive definiteness.
Consider the set:

$$\left\{x\epsilon \ ^{p} : x \geq 0, \iota^\top x = 1, \Sigma x \geq 0\right\} \tag{35}$$

where ι is a column vector containing p ones. The defining conditions can also be written in the form $Ax \leq b$ with

$$A \equiv \begin{bmatrix} +\imath^\top \\ -\imath^\top \\ -I \\ \Sigma \end{bmatrix} \; and \; b \equiv \begin{bmatrix} +1 \\ -1 \\ 0 \\ 0 \end{bmatrix} \tag{36}$$

where I is the p by p identity matrix, and the zero vectors in b each have p components. Farkas' lemma (see e. g. Alexander Schrijver, 2004, in particular corollary 2.5a in section 2.3.) implies that the set

$$\{x\epsilon \ ^{p} : Ax \leq b\} \tag{37}$$

is not empty if and only if the set

$$\left\{y\epsilon \quad ^{2p+2} : y \geq 0, y^{\top}A = 0, y^{\top}b < 0\right\} \tag{38}$$

is empty. If we write $y^{\top}$ as $\left(y_1, y_2, u^{\top}, v^{\top}\right)$ where u and v are both of order p by 1, we can express $y^{\top}A = 0$ as

$$v^{\mathsf{T}} \ \varSigma + u^{\mathsf{T}} + (y_2 - y_1) \cdot \imath^{\mathsf{T}} \ = 0 \tag{39}$$

and the inequalities in (38) require that u and v must be non-negative and that $y_2 - y_1$ is positive. If we postmultiply (39) by v we get:

$$v^{\top}\varSigma v + u^{\top}v + (y_2 - y_1) \cdot \imath^{\top}v = 0 \tag{40}$$

which entails that v is zero and therefore from (39) that u as well as $y_2 - y_1$ are zero. (Note that this is true even when $\varSigma$ is just positive semi-definite). We conclude that the second set is empty, so the first set is nonempty indeed! Therefore there are always admissible indices for any set of indicators. We can describe this set in some more detail if we write the conditions in 'standard form' as in a linear programming setting. Define the matrix $\boldsymbol{A}$ as:

$$\boldsymbol{A} \ \equiv \begin{bmatrix} \imath^{\mathsf{T}} & 0^{\mathsf{T}} \\ \varSigma & -I \end{bmatrix} \tag{41}$$

where $\imath$ is again of order p by 1, and the dimensions of the other entries follow from this. Note that $\boldsymbol{A}$ has $2p$ columns. It is easily verified that the matrix $\boldsymbol{A}$ has full rowrank $p+1$ if $\varSigma$ is positive definite. Also define a $p+1$ by 1 vector $\boldsymbol{b}$ as $[1; 0]$, a 1 stacked on top of p zeros, and let s be a p by 1 vector of 'slack variables'. The original set can now be reframed as:

$$\left\{x\epsilon \quad ^{p}, s\epsilon \quad ^{p} : \boldsymbol{A} \cdot \begin{bmatrix} x \\ s \end{bmatrix} = \boldsymbol{b}, \ x \geq 0, s \geq 0\right\} \tag{42}$$

Clearly this is a convex set, a convex polytope in fact, that can be generated by its extreme points. The latter can be found by selecting $p+1$ independent columns from $\boldsymbol{A}$, resulting in a matrix $\boldsymbol{A}_B$, say, with B for 'basis', and checking whether the product of the inverse of $\boldsymbol{A}_B$ times $\boldsymbol{b}$ has nonnegative elements only (note that $\boldsymbol{A}_B^{-1}\boldsymbol{b}$ is the first column of the inverse of $\boldsymbol{A}_B$). If so, the vector $[x; s]$ containing zeros corresponding to the columns of $\boldsymbol{A}$ which were not selected, is an extreme point of the enlarged space (x, s). Since the set is bounded, the corresponding subvector x is an extreme point of the original (x)-space. In principle we have to evaluate $\binom{2p}{p+1}$ possible candidates. A special and trivial case is where the elements of $\varSigma$ are all non-negative: all weighting vectors are acceptable, and, as pointed out before, the first principal component (suitably normalized) is one of them.

6.2 Potentially useful constraints

As indicated before we propose to determine for every block of indicators its set of admissible proper indices, and then choose from each of these sets an index such that some suitable function of the correlation matrix of the selected indices is optimized; we suggested the determinant (minimize) or the first eigenvalue (maximize), and others. A useful refinement may be the incorporation of a priori constraints on the relationships between the indices. Typically one employs a pathdiagram that embodies zero or multiplicative constraints on regression coefficients. It may happen e.g. that two indices are believed to be correlated only because of their linear dependence on a third index, so that the conditional correlation between the two given the third is zero: $\rho_{23.1}$,say, equals 0. This is equivalent to postulating that the entry in the second row and third column of the inverse of the correlation matrix of the three indices is zero (see D. R. Cox and Nanny Wermuth (1998) in particular the sections 3.1-3.4). More complicated constraints are generated by zero constraints on structural form matrices. E. g. the matrix that links three endogenous latent variables to each other might have the following structure:

$$B = \begin{bmatrix} \beta_{11} & 0 & 0 \\ \beta_{21} & \beta_{22} & \beta_{23} \\ 0 & \beta_{32} & \beta_{33} \end{bmatrix} \tag{43}$$

and the effect of the remaining exogenous latent variables on the first set is captured by

$$\Gamma = \begin{bmatrix} 0 & \gamma_{12} \\ \gamma_{21} & 0 \\ 0 & 0 \end{bmatrix} \tag{44}$$

Observe that not all parameters are identifiable, not even after normalization (β_{23} will be unidentifiable). But the matrix of regression coefficients, of the regressions of the three endogenous latent variables on the two endogenous latent variables, taking the given structure into account, satisfies both zero constraints as well as multiplicative constraints. In fact, this matrix, Π, say, with $\Pi \equiv B^{-1}\Gamma$, can be parameterized in a minimal way as follows (see Bekker et al. (1994) section 5.6):

$$\Pi = \begin{bmatrix} 0 & \theta_3 \\ \theta_1 & \theta_1\theta_4 \\ \theta_2 & \theta_2\theta_4 \end{bmatrix} \tag{45}$$

So $\Pi_{11} = 0$ and $\Pi_{21}\Pi_{32} - \Pi_{22}\Pi_{31} = 0$. These restrictions should perhaps not be wasted when constructing indices. They can be translated into restrictions on the inverses of appropriate submatrices of the correlation matrix of the latent variables. Bekker et al. (1994) have developed software for the automatic generation of minimal parameterizations.

Some small scale experiments by the author, using the constraints of properness and those implied by a path diagram, were encouraging (to the author), and only a few lines of MATLAB-code were required. But clearly a lot of development work and testing remains to be done. For constructing and testing indices a strong case can be made for *cross-validation*, which naturally honoures one of *the* purposes of the entire exercise: prediction of observables. It fits rather naturally with the low-structure environment for which PLS was invented, with its soft or fuzzy relationships between (composite) variables. See e. g. S. Geisser (1993) and T. Hastie et al. (2002) for cross-validation techniques and analyses. Cross-validation was embraced early by Herman Wold. He also saw clearly the potential of the related *Jackknife*-method, see Wold (1975).

7 Conclusion

I have described and analyzed some of PLS' properties in the context of a latent variable model. It was established that one may expect the algorithms to converge, from essentially arbitary starting values, to unique fixed-points. As a function of the sample size these points do not necessarily converge to the parameters of the latent variable model, in fact their limits or theoretical values may differ substantially from the 'true' value if the quality of the proxies is not (very) high. But in principle it is possible to adjust the PLS-estimators in a simple way to cancel the induced distortions, within the context of the (extended) basic design. I also outlined an approach where the indices are treated as the fundamental objects, and where the path diagrams serve to construct meaningful, proper indices, satisfying constraints that are relatively modest.

There are other approaches construed as alternatives to PLS. One such approach, as pointed out by a referee, is due to McDonald (1996) who designed six methods for the estimation of latent variable models as the basic design. These methods all share a least squares type of fitting function and a deliberate distortion of the underlying latent variable model. His method I e. g. minimizes the sum of squares of the difference between S and $\Sigma(\theta)$ as a function of θ, where θ contains the loadings as well as the structural parameters of the relationships between the latent variables, and where all measurement error variances are *a priori* taken to be zero. Once the optimal value for θ is obtained, weighting vectors for the composites are chosen proportional to the estimated loading vectors. McDonald applies his methods as well as PLS to a particular, simple population correlation matrix, with known parameters. Method I is the favorite of the referee who referred me to McDonald (1996), but McDonald himself carefully avoids to state his overall preferences. Clearly, one set of parameters is no basis for a well-established preference, as McDonald takes care to point out on page 254, and again on page 262: the results will typically be rather parameter dependent. I think it is relevant to note the fact, which is not difficult to show, that Method I's loading vectors based on true parameters, their probability limits, are typically *not* proportional to the true loadings, as opposed to PLS mode B. Table 2 of McDonald (1996) confirms this. Moreover, the ensuing proxies are *not* proportional to the best linear predictors of the latent variables (in terms of their direct indicators), again unlike PLS mode B. A necessary and sufficient condition for proportionality in the context of the basic design with unrestricted correlations between the latent variables, is that the loading vectors are eigenvectors of the corresponding error covariance matrices; if the latter are diagonal the unique factors of each block should have identical variances.

One reviewer of McDonald's paper, apparently a member of the 'PLS-camp', suggested that among users of PLS there is an emerging consensus that PLS represents a philosophy rather different from the standard philosophy of what quantitative behavioral science is doing: PLS is mainly prediction-oriented whereas the traditional approach is mainly inference-oriented. I tend to agree with this reviewer, if only for the fact that in each and every one of Wold's contributions to statistics 'prediction' and 'predictive specifications' are central, key terms. And there is also the embryonic PLS-model of principal components, which served as one of the starting points of PLS (or NIPALS as it was called then in 1966): loadings *as well as* 'latent' variables are viewed and treated as parameters to be estimated with a least squares 'prediction' criterion leading to linear compounds as estimates for the latent variables. So in this context at least, the approach appears to be entirely natural. But I would maintain that it is still in need of serious development and explication. Somehow the latent variable model, the basic design, seems to have interfered in a pernicious way by posturing as *the* unique and proper way to analyze and model high-dimensional data; this may have (as far as I can see) impeded further developments. Without wanting to sound presumptuous, my contribution contained in section 6 can be seen as an attempt to revive what I believe to be the original program. Perhaps PLS could re-orient itself by focussing on (proper) index building through prediction-based cross-validation. McDonald clearly disagrees with the reviewer of his paper about the prediction versus inference issue, and counters by claiming that, if it were true, since 'we cannot do better than to use multivariate regressions or canonical variate analysis', one would expect to see a preference among PLS users for multivariate regressions, or if they *must* use a path model they should prefer mode B to mode A. Since this does not seem to happen in practice he infers the invalidity of the reviewer's statement. McDonald has a point when the true parameters are known, but not when they are subject to estimation. If the goal is prediction, this goal is as a rule served best by simplifying the maintained model even more than we would do if description were just what we were after. In fact, predictors based on a moderately incorrect version of the 'true model' usually outperform those constructed on the basis of a more elaborate, more correct version, see T. K. Dijkstra (1989) or T. Hastie et al. (2002). In other words, one can certainly not dismiss path models and indices if prediction is called for.

The final issue raised by McDonald at the very end of his paper concerns the use and appropriateness of latent variable models (in what follows the emphasis is mine). He contends that because of factor score indeterminacy, a small number of indicators makes a latent variable model quite inappropriate;

indeed, we need lots of them if we want to do any *serious* work using the model (this is an 'inescapable fact'). But if we have a large number of indicators per latent variable, a simple average of the former will do an adequate job in replacing the latter, so we then no longer need the model (in other words, the model is either inappropriate or redundant). In my opinion this point of view is completely at odds with the notion of an acceptable model being a useful approximation to part of reality, latent variable modelling is no exception. If a model is to be any good for empirical explanation, prediction or otherwise, it should *not* be a complete and correct specification. See among many e. g. Kaplan (1946, 1964), or T. Hastie *et al.* (2002). A suitable methaphor is a map, that by its very nature *must* yield a more or less distorted picture of 'angles and distances'; maps that are one-to-one can't get us anywhere. The technical merits of McDonald's paper are not disputed here, but the philosophical and methodological content I find hard to understand and accept.

The reviewer of the present chapter concludes from McDonalds results that 'PLS was a mistake, and Method I should have been invented instead. PLS should simply be abandoned'. I disagree. I contend that PLS' philosophy potentially has a lot to offer. In my view there is considerable scope in the social sciences, especially in high-dimensional, low-structure, fuzzy environments, for statistical approaches that specify and construct rather simple 'index-models' through serious predictive testing. PLS in one version or the other still appears to have untapped sources, waiting to be exploited.

References

Anderson, T. W. (1984). *An introduction to Multivariate Statistical Analysis*, Wiley, New York.

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.

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

Copson, E. T. (1968). *Metric Spaces*, Cambridge University Press, Cambridge.

Cox, D. R. and Wermuth, N. (1998). *Multivariate Dependencies- models, analysis and interpretation,* Chapman & Hall, Boca Raton.

Dijkstra, T. K. (1981). *Latent Variables in Linear Stochastic Models,* PhD-thesis, its second edition was published in 1985 by Sociometric Research Foundation, Amsterdam.

Dijkstra, T. K. (1982). Some comments on Maximum Likelihood and Partial Least Squares Methods, *Research Report UCLA, Dept. Psychology,* a shortened version was published in 1983.

Dijkstra, T. K. (1983). Some comments on Maximum Likelihood and Partial Least Squares Methods, *Journal of Econometrics* 22: 67-90.

Dijkstra, T. K. (1988). *On Model Uncertainty and its Statistical Implications, (ed.),* Springer Verlag, Heidelberg.

Dijkstra, T. K. (1989). Reduced Form Estimation, Hedging against possible Misspecification, *International Economic Review* 30(2): 373-390.

Dijkstra, T. K. (1990). Some properties of estimated scale invariant covariance structures, *Psychometrika* 55: 327-336.

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

Frank, I. E., Friedman, J. H. (1993). A Statistical View of Some Chemometric Regression Tools, *Technometrics* 35: 109-135.

Gantmacher, F. R. (1977). *The Theory of Matrices,* Volume 1, Chelsea Publishing Company.

Geisser, S. (1993). *Predictive Inference: An Introduction,* Chapman&Hall, New York.

Hastie, T., Tibshirani, R., Friedman, J. (2002). *The elements of statistical learning,* Springer Verlag, New York.

Jöreskog, K. G., Wold, H. O. A. (1982). *Systems under indirect observation, Part II, (eds.),* North-Holland, Amsterdam.

Kagan, A. M., Linnik, Y. V., Rao, C. R. (1973). *Characterization problems in mathematical statistics,* Wiley, New York.

Kaplan, A. (1946). Definition and specification of meaning, *The Journal of Philosophy* 43: 281-288.

Kaplan, A. (1964). *The Conduct of Inquiry,* Chandler, New York.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables, *Biometrika* 58: 433-451.

Leamer, Edward E. (1978). *Specification Searches, Wiley,New York.*

McDonald, R.P. (1996). Path Analysis with Composite Variables, *Multivariate Behavioral Research* 31(2): 239-270.

Ortega, J. M., Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables,* Academic Press, New York.

Schrijver, A. (2004). *A course in combinatorial optimization,* Springer Verlag, Berlin.

Schneeweiss, H., Mathes, H. (1995). Factor Analysis and Principal Components, *Journal of Multivariate Analysis* 55: 105-124.

Stone, M., Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression, *Journal of the Royal Statistical Society, Series B (Methodological)* 52: 237-269.

Tenenhaus, M.,Vinzi, V. E., Chatelin, Y-M., Lauro, C. (2005). PLS path modelling, *Computational Statistics & Data Analysis* 48: 159-205.

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

Wold, H. O. A. (1975). Path Models with Latent Variables: The NIPALS Approach, *in* Blalock, H. M., Aganbegian, A., Borodkin, F. M., Boudon, R., Capecchi, V., *Quantitative Sociology,* Academic Press, New York, 307-359.

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, 1-55.

Index

asymptotic efficiency, 9
asymptotic normality, 4

basic design, 1, 4, 5, 22
  extended ∼, 6
best linear approximations, 7
best linear predictor, 13, 21
bias, 16
  structural ∼, 14
bootstrap, 9, 14

canonical variables, 4, 11, 14
causal chain, 5
  complete ∼, 6
characterization theorem, 6
consistency
  ∼ at large, 15
  ∼ of LISREL-estimators, 4
  ∼ property, 18
cross-validation, 20

distribution free, 2, 6
distributional assumptions, 4, 6

elliptical distributions, 9
endogenous latent variables, 5
entropy, 17
EQS, 3
exogenous latent variables, 5

Fixed point theorem with a parameter, 12

GENVAR, 17

GLS, 7

implicit function theorem, 8
indices
  admissible ∼, 17
  GENVAR ∼, 18
  proper ∼, 2, 16, 18, 19

Jackknife, 21

likelihood principle, 7
linear compounds, 16
LISREL, 2, 4, 5, 7–10
local contraction mapping, 12

maximum entropy principle, 7
MAXVAR, 17
MIMIC-model, 11
MINVAR, 17
mode
  ∼ A, 10, 13, 15, 22
  ∼ B, 10, 11, 13, 15, 21, 22
  ∼ C, 10, 11
model
  ∼ uncertainty, 9
  acceptable ∼, 23
multinormality, 6
  distance to ∼, 5
multiplicative constraints, 18

path diagram, 16
PLS, 2, 5, 10
  ∼ algorithms, 2
  alternatives to ∼, 21

prejudice, 7
principal component, 4, 10, 22
  first ∼, 17, 19
proxy, 10
  high quality ∼, 16
  population ∼, 13
  sample ∼, 13

rank restrictions, 14
reduced form parameter, 14
regression matrix for the population proxies, 14

scale-invariance, 10, 11
second order factor model, 4
squared correlation, 13
squared multiple correlations, 14
structural equation modeling, 1
structural equations, 5
structural form parameter, 14
symbolic calculations, 18

vector of loadings, 5

Weighted Least Squares, 8