Chapter 4
A Perfect Match Between a Model and a Mode

Theo K. Dijkstra

Abstract When the partial least squares estimation methods, the "modes," are applied to the standard latent factor model against which methods are designed and calibrated in PLS, they will not yield consistent estimators without adjustments. We specify a different model in terms of observables only, that satisfies the same rank constraints as the latent variable model, and show that now mode B is perfectly suitable without the need for corrections. The model explicitly uses composites, linear combinations of observables, instead of latent factors. The composites may satisfy identifiable linear structural equations, which need not be regression equations, estimable via 2SLS or 3SLS. Each time practitioners contemplate the use of PLS' basic design model the composites model is a viable alternative. The chapter is conceptual mainly, but a small Monte Carlo study exemplifies the feasibility of the new approach.

4.1 Introduction

Herman (H.O.A.) Wold (1908–1992) developed partial least squares (PLS) in a series of papers, published as well as privately circulated. The *seminal* published papers are Wold (1966, 1975, 1982). A key characteristic of PLS is the determination of composites, linear combinations of observables, by weights that are fixed

This chapter "continues" a sometimes rather spirited discussion with Wold, that started in 1977, at the Wharton School in Philadelphia, via my PhD thesis, Dijkstra (1981), and a paper Dijkstra (1983). There was a long silence, until about 2008, when Peter M. Bentler (UCLA) rekindled my interest in PLS, one of the many things for which I owe him my gratitude. Crucial also is the collaboration with Joerg Henseler (Twente), that led to a number of papers on PLS and on ways to get consistency without the need to increase the number of indicators, PLSc, as well as to a software program ADANCO for composites. I am very much in his debt too. The present chapter expands on Dijkstra (2010) by avoiding unobservables as much as possible while still adhering to Wold's fundamental principle of soft modeling.

T.K. Dijkstra (✉)
Faculty of Economics and Business, University of Groningen, Groningen, Netherlands
e-mail: t.k.dijkstra@rug.nl

H. Latan, R. Noonan (eds.), *Partial Least Squares Path Modeling*,
DOI 10.1007/978-3-319-64069-3_4

points of sequences of alternating least squares programs, called "modes." Wold distinguished three types of modes (not models!): mode A, reminiscent of principal component analysis, mode B, related to canonical variables analysis, and mode C, that mixes the former two. In a sense PLS is an extension of canonical variables and principal components analyses. While Wold designed the algorithms, great strides were made in the estimation, testing, and analysis of structured covariance matrices, as induced by linear structural equations in terms of latent factors and indicators (LISREL first, then EQS et cetera). Latent factor modeling became the dominant backdrop against which Wold designed his tools. One model in particular, the "basic design," became the model of choice in calibrating PLS. Here each latent factor is measured indirectly by a unique set of indicators, with all measurement errors usually assumed to be mutually uncorrelated. The composites combine the indicators for each latent factor separately, and their relationships are estimated by regressions.[1] The basic design embodies Wold's "fundamental principle of soft modeling": all information between the blocks of observables is assumed to be conveyed by latent variables (Wold 1982).[2] However, in this model PLS is not well-calibrated[3]: when applied to the true covariance matrix it yields by *necessity* approximations, see, e.g., Dijkstra (1981, 1983; 2010; 2014). For consistency, meaning that the probability limit of the estimators equals the theoretical value, Wold also requires the number of indicators to increase alongside the number of observations (consistency-at-large).

In this chapter we leave the realm of the unobservables, and build a model in terms of manifest variables that satisfies the fundamental principle of soft modeling, adjusted to read: *all information between the blocks is conveyed solely by the composites.* For this model, mode B is the perfect match, in the sense that estimation via mode B is the natural thing to do: when applied to the true covariance matrix it yields the underlying parameter values, not approximations that require corrections. A latent factor model, in contrast, would need additional structure (like uncorrelated measurement errors) and fitting it would produce approximations.

The chapter is structured as follows. The next section, Sect. 4.2, outlines the new model. We specify for a vector $\mathbf{y}$ of observable variables, "indicators," a structural model that generates via linear composites of separate blocks of indicators all the standard *basic design* rank restrictions on the covariance matrix, without invoking

[1]This includes simultaneous equations systems, which are generally not regressions. They were estimated by a Fix Point method, essentially iterations of 2SLS (two-stage-least-squares) regressions (Boardman et al. 1981). See below for 2SLS and Dijkstra and Henseler (2015a,b).

[2]"Soft modeling" indicates that PLS is meant to perform "substantive analysis of complex problems that are at the same time data-rich and theory-primitive" (Wold 1982).

[3]I am not saying here that methods that are not well-calibrated are intrinsically "bad." This would be ludicrous given the inherent approximate nature of statistical models. Good predictions typically require a fair amount of misspecification, to put it provocatively. But knowing what happens when we apply a statistical method to "the population" helps answering what it is that it is estimating. Besides, consistency, and to a much lesser extent "efficiency," was very important to Wold.

the existence of unobservable latent factors. They, the composites, are linked to each other by means of a "structural," "simultaneous," or "interdependent" equations system, that together with the loadings fully captures the (linear) relationships between the blocks of indicators.

Section 4.3 is devoted to estimation issues. We describe a step-wise procedure: first the weights defining the composites via generalized canonical variables,[4] then their correlations and the loadings in the simplest possible way, and finally the parameters of the simultaneous equations system using the econometric methods 2SLS or 3SLS. The estimation proceeds essentially in a non-iterative fashion (even when we use one of the PLS' traditional algorithms, it will be very fast), making it potentially eminently suitable for bootstrap analyses. We give the results of a Monte Carlo simulation for a model for 18 indicators; they are generated by six composites linked to each other via two linear equations, which are *not* regressions. We also show that mode A, when applied to the true covariance matrix of the indicators, can only yield the correct results when the composites are certain principal components. As in PLSc, mode A can be adjusted to produce the right results (in the limit).

Section 4.4 suggests how to test various aspects of the model, via tests of the rank constraints, via prediction/cross-validation, and via global goodness-of-fit tests.

Section 4.5 contains some final observations and comments. We briefly return to "the latent factors versus composites"-issue and point out that in a latent factor model the factors cannot fully be replaced by linear composites, no matter how we choose them: the regression of the indicators on the composites will not yield the loadings on the factors, *or* (inclusive) the composites cannot satisfy the same equations that the factors satisfy.

The Appendix contains a proof for a statement needed in Sect. 4.3.

4.2 The Model: Composites as Factors

Our point of departure is a random vector[5] $\mathbf{y}$ of "indicators" that can be partitioned into N subvectors, "blocks" in PLS parlance, as $\mathbf{y} = (\mathbf{y}_1; \mathbf{y}_2; \mathbf{y}_3; \ldots; \mathbf{y}_N)$. Here the semi-colon stacks the subvectors one underneath the other, as in MATLAB; $\mathbf{y}_i$ is of order $p_i \times 1$ with p_i usually larger than one. So $\mathbf{y}$ is of dimension $p \times 1$ with $p := \sum_{i=1}^{N} p_i$. We will denote $\text{cov}(\mathbf{y})$ by $\mathbf{\Sigma}$, and take it to be positive definite (p.d.), so no indicator is redundant. We will let $\mathbf{\Sigma}_{ii} := \text{cov}(\mathbf{y}_i)$. $\mathbf{\Sigma}_{ii}$ is of order $p_i \times p_i$ and it is of course p.d. as well. It is *not* necessary to have other constraints on $\mathbf{\Sigma}_{ii}$, in particular it need not have a one-factor structure. Each block $\mathbf{y}_i$ is condensed into a composite, a scalar c_i, by means of a conformable weight

[4] It should be pointed out that I see PLS' mode B as one of a family of generalized canonical variables estimation methods (Sect. 4.3.1), to be treated on a par with the others, without necessarily claiming that mode B is the superior or inferior method. None of the methods will be uniformly superior in every sensible aspect.

[5] Vectors and matrices will be distinguished from scalars by printing them in boldface.

vector $\mathbf{w}_i$: $c_i := \mathbf{w}_i^{\mathsf{T}}\mathbf{y}_i$. The composites will be normalized to have variance one: $\text{var}(c_i) = \mathbf{w}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{ii}\mathbf{w}_i = 1$. The vector of composites $\mathbf{c} := (c_1; c_2; c_3; \ldots; c_N)$ has a p.d. covariance/correlation matrix denoted by $\mathbf{R}_c = \left(r_{ij}\right)$ with $r_{ij} = \mathbf{w}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{ij}\mathbf{w}_j$ where $\boldsymbol{\Sigma}_{ij} := \text{E}(\mathbf{y}_i - \text{E}\mathbf{y}_i)\left(\mathbf{y}_j - \text{E}\mathbf{y}_j\right)^{\mathsf{T}}$. A regression of $\mathbf{y}_i$ on c_i and a constant gives a loading vector $\mathbf{L}_i$ of order $p_i \times 1$:

$$\mathbf{L}_i := \text{E}\left(\mathbf{y}_i - \text{E}\mathbf{y}_i\right) \cdot \left(c_i - \text{E}c_i\right) = \text{E}\left(\mathbf{y}_i - \text{E}\mathbf{y}_i\right)\left(\mathbf{y}_i - \text{E}\mathbf{y}_i\right)^{\mathsf{T}}\mathbf{w}_i = \boldsymbol{\Sigma}_{ii}\mathbf{w}_i \tag{4.1}$$

So far all we have is a list of definitions but as yet no real model: there are no constraints on the joint distribution of $\mathbf{y}$ apart from the existence of moments[6] and a p.d. covariance matrix. We will now impose our version of Wold's fundamental principle in soft modeling:

all information between the blocks is conveyed solely by the composites

We deviate from Wold's original formulation in an essential way: whereas Wold postulated that all information is conveyed by unobserved, even unobservable, latent variables, we let the information to be fully transmitted by indices, by composites of observable indicators. So we postulate the existence of weight vectors such that for any two different blocks $\mathbf{y}_i$ and $\mathbf{y}_j$

$$\begin{aligned}\boldsymbol{\Sigma}_{ij} &= r_{ij}\mathbf{L}_i\mathbf{L}_j^{\mathsf{T}} &(4.2)\\ &= \mathbf{w}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{ij}\mathbf{w}_j \cdot \boldsymbol{\Sigma}_{ii}\mathbf{w}_i \cdot \left(\boldsymbol{\Sigma}_{jj}\mathbf{w}_j\right)^{\mathsf{T}}\\ &= \text{corr}\left(\mathbf{w}_i^{\mathsf{T}}\mathbf{y}_i, \mathbf{w}_j^{\mathsf{T}}\mathbf{y}_j\right) \cdot \text{cov}\left(\mathbf{y}_i, \mathbf{w}_i^{\mathsf{T}}\mathbf{y}_i\right) \cdot \left(\text{cov}\left(\mathbf{y}_j, \mathbf{w}_j^{\mathsf{T}}\mathbf{y}_j\right)\right)^{\mathsf{T}} &(4.3)\end{aligned}$$

The cross-covariances between the blocks are determined by the correlation between their corresponding composites and the loadings of the blocks on those composites. Note that line (4.2) is highly reminiscent of the corresponding equation for the basic design, with latent variables. There it would read $\rho_{ij}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_j^{\mathsf{T}}$ with ρ_{ij} representing the correlation between the latent variables, with $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ capturing the loadings. So the rank-one structure of the covariance matrices between the blocks is maintained fully, without requiring the existence of N additional unobservable variables.

We now have:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & r_{12}\mathbf{L}_1\mathbf{L}_2^{\mathsf{T}} & r_{13}\mathbf{L}_1\mathbf{L}_3^{\mathsf{T}} & \cdot & r_{1N}\mathbf{L}_1\mathbf{L}_N^{\mathsf{T}} \\ & \boldsymbol{\Sigma}_{22} & r_{23}\mathbf{L}_2\mathbf{L}_3^{\mathsf{T}} & \cdot & r_{2N}\mathbf{L}_2\mathbf{L}_N^{\mathsf{T}} \\ & & \cdot & \cdot & \cdot \\ & & & \boldsymbol{\Sigma}_{N-1,N-1} & r_{N-1,N}\mathbf{L}_{N-1}\mathbf{L}_N^{\mathsf{T}} \\ & & & & \boldsymbol{\Sigma}_{NN} \end{bmatrix} \tag{4.4}$$

[6] A random sample of indicator-vectors and the existence of second order moments is sufficient for the consistency of the estimators to be developed below; with the existence of fourth-order moments we also have asymptotic normality.

The appendix contains a proof of the fact that $\mathbf{\Sigma}$ is positive definite when and only when the correlation matrix of the composites, $\mathbf{R}_c$, is positive definite. Note that in a Monte Carlo analysis we can choose the weight vectors (or loadings) and the values of $\mathbf{R}_c$ independently.

We can add more structure to the model by imposing constraints on $\mathbf{R}_c$. This is done most conveniently by postulating a set of simultaneous equations to be satisfied by $\mathbf{c}$. We will call one subvector of $\mathbf{c}$ the *exogenous* composites, denoted by $\mathbf{c}_{\text{exo}}$, and the remaining elements will be collected in $\mathbf{c}_{\text{endo}}$, the *endogenous* composites. There will be conformable matrices $\mathbf{B}$ and $\mathbf{C}$ with $\mathbf{B}$ invertible such that

$$\mathbf{B}\mathbf{c}_{\text{endo}} = \mathbf{C}\mathbf{c}_{\text{exo}} + \mathbf{z} \tag{4.5}$$

It is customary to normalize $\mathbf{B}$, i.e., all diagonal elements equal one (perhaps after some re-ordering). The residual vector $\mathbf{z}$ has a zero mean and is uncorrelated with $\mathbf{c}_{\text{exo}}$. In this type of (econometric) model the relationships between the exogenous variables are usually not the main concern. The research focus is on the way they drive the endogenous variables and the interplay or the feedback mechanism between the latter as captured by a matrix $\mathbf{B}$ that has nonzero elements both above and below the diagonal. A special case, with no feedback mechanism at all, is the class of *recursive* models, where $\mathbf{B}$ has only zeros on one side of its diagonal, and the elements of $\mathbf{z}$ are mutually uncorrelated. Here the coefficients in $\mathbf{B}$ and $\mathbf{C}$ can be obtained directly by consecutive regressions, given the composites. For general $\mathbf{B}$ this is not possible, since $\mathbf{c}_{\text{endo}}$ is a linear function of $\mathbf{z}$ so that z_i will typically be correlated with every endogenous variable in the ith equation.[7]

Even when the model is not recursive, the matrices $\mathbf{B}$ and $\mathbf{C}$ will be postulated to satisfy certain zero constraints (and possibly other types of constraints, but we focus here on the simplest situation). So some B_{ij}'s and C_{kl}'s are zero. We will assume that the remaining coefficients are *identifiable* from a knowledge of the so-called reduced form matrix $\mathbf{\Pi}$

$$\mathbf{\Pi} := \boldsymbol{B}^{-1}\mathbf{C} \tag{4.6}$$

Note that

$$\mathbf{c}_{\text{endo}} = \mathbf{\Pi}\boldsymbol{c}_{\text{exo}} + \mathbf{B}^{-1}\mathbf{z} \tag{4.7}$$

so $\mathbf{\Pi}$ is a matrix of regression coefficients. Once we have those, we should be able to retrieve $\mathbf{B}$ and $\mathbf{C}$ from them. Identifiability is equivalent to the existence of certain rank conditions on $\mathbf{\Pi}$, we will have more to say about them later on. We could have additional constraints on the covariance matrices of $\mathbf{c}_{\text{exo}}$ and $\mathbf{z}$ but we will not develop that here, taking the approach that demands the least in terms of knowledge

[7] See Pearl (2009) for an in-depth causal analysis of simultaneous equations systems (based on and extending (Haavelmo 1944), probably the best apologia of econometrics ever).

about the relationships between the composites. It is perhaps good to note that granted identifiability, the free elements in $\mathbf{B}$ and $\mathbf{C}$ can be interpreted as regression coefficients, provided we replace the "explanatory" endogenous composites by their regression on the exogenous composites. This is easily seen as follows:

$$\begin{aligned}\mathbf{c}_{\text{endo}} &= (\mathbf{I}-\mathbf{B})\,\mathbf{c}_{\text{endo}} + \mathbf{C}\mathbf{c}_{\text{exo}} + \mathbf{z} && (4.8)\\ &= (\mathbf{I}-\mathbf{B})\left(\mathbf{\Pi}\boldsymbol{c}_{\text{exo}} + \mathbf{B}^{-1}\mathbf{z}\right) + \mathbf{C}\mathbf{c}_{\text{exo}} + \mathbf{z} && (4.9)\\ &= (\mathbf{I}-\mathbf{B})\,(\mathbf{\Pi}\boldsymbol{c}_{\text{exo}}) + \mathbf{C}\mathbf{c}_{\text{exo}} + \mathbf{B}^{-1}\mathbf{z} && (4.10)\end{aligned}$$

where $\mathbf{B}^{-1}\mathbf{z}$ is uncorrelated with $\mathbf{\Pi}\boldsymbol{c}_{\text{exo}}$ and $\mathbf{c}_{\text{exo}}$. So the free elements of $(\mathbf{I}-\mathbf{B})$ and $\mathbf{C}$ can be obtained by a regression of $\mathbf{c}_{\text{endo}}$ on $\mathbf{\Pi}\boldsymbol{c}_{\text{exo}}$ and $\mathbf{c}_{\text{exo}}$, equation by equation.[8] Identifiability is here equivalent to invertibility of the covariance matrix of the "explanatory" variables in each equation. A necessary condition for this to work is that we cannot have more coefficients to estimate in each equation than the total number of exogenous composites in the system.

We have for $\mathbf{R}_c$

$$\mathbf{R}_c = \begin{bmatrix}\text{cov}\,(\mathbf{c}_{\text{exo}}) & \text{cov}\,(\mathbf{c}_{\text{exo}})\cdot\mathbf{\Pi}^{\intercal} \\ \mathbf{\Pi}\text{cov}\,(\mathbf{c}_{\text{exo}})\,\mathbf{\Pi}^{\intercal} + \mathbf{B}^{-1}\text{cov}\,(\mathbf{z})\,(\mathbf{B}^{\intercal})^{-1} & \end{bmatrix} \qquad (4.11)$$

Thanks to the structural constraints, the number of parameters in $\mathbf{R}_c$ could be (considerably) less than $\frac{1}{2}N(N-1)$, potentially allowing for an increase in estimation efficiency.

As far as $\mathbf{\Sigma}$ is concerned, the model is now completely specified.

4.2.1 *Fundamental Properties of the Model and Wold's Fundamental Principle*

Now define for each i the measurement error vector $\mathbf{d}_i$ via

$$\mathbf{y}_i - \text{mean}\,(\mathbf{y}_i) = \mathbf{L}_i\,(c_i - \text{mean}\,(c_i)) + \mathbf{d}_i \qquad (4.12)$$

where $\mathbf{L}_i = \mathbf{\Sigma}_{ii}\,\mathbf{w}_i$, the loadings vector obtained by a regression of the indicators on their composite (and a constant).

By construction $\mathbf{d}_i$ has a zero mean and is uncorrelated with c_i. In what follows it will be convenient to have all variables de-meaned, so we have $\mathbf{y}_i = \mathbf{L}_i c_i + \mathbf{d}_i$. It is easy to verify that:

[8] The estimation method based on these observations is called 2SLS, two-stage-least-squares, for obvious reasons, and was developed by econometricians in the 1950s of the previous century.

The measurement error vectors are mutually uncorrelated, and uncorrelated with all composites:

$$\mathrm{E}\mathbf{d}_i\mathbf{d}_j^{\mathsf{T}} = 0 \text{ for all different } i \text{ and } j \tag{4.13}$$

$$\mathrm{E}\mathbf{d}_i c_j = 0 \text{ for all } i \text{ and } j \tag{4.14}$$

It follows that $\mathrm{E}\mathbf{y}_i\mathbf{d}_j^{\mathsf{T}} = 0$ for all different i and j. In addition:

$$\mathrm{cov}\,(\mathbf{d}_i) = \boldsymbol{\Sigma}_{ii} - \mathbf{L}_i\mathbf{L}_i^{\mathsf{T}} \tag{4.15}$$

The latter is also very similar to the corresponding expression in the basic design, but we cannot in general have a diagonal $\mathrm{cov}(\mathbf{d}_i)$, because $\mathrm{cov}(\mathbf{d}_i)\,\mathbf{w}_i$ is identically zero (implying that the variance of $\mathbf{w}_i^{\mathsf{T}}\mathbf{d}_i$ is zero, and therefore $\mathbf{w}_i^{\mathsf{T}}\mathbf{d}_i = 0$ with probability one). The following relationships can be verified algebraically using regression results, or by using conditional expectations formally (so even though we use the formalism of conditional expectations and the notation, we do just mean regression).

$$\mathrm{E}\,(\mathbf{y}_1|c_1) = \mathbf{L}_1 c_1 \tag{4.16}$$

because $\mathrm{E}(\mathbf{y}_1|c_1) = \mathrm{E}(\mathbf{L}_1 c_1 + \mathbf{d}_1|c_1) = \mathbf{L}_1 c_1 + 0$. Also note that

$$\mathrm{E}\,(c_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) \tag{4.17}$$

$$= \mathrm{E}(\mathrm{E}\,(c_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N, \mathbf{d}_2, \mathbf{d}_3, \ldots, \mathbf{d}_N)\,|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) \tag{4.18}$$

$$= \mathrm{E}(\mathrm{E}\,(c_1|c_2, c_3, \ldots, c_N, \mathbf{d}_2, \mathbf{d}_3, \ldots, \mathbf{d}_N)\,|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) \tag{4.19}$$

$$= \mathrm{E}(\mathrm{E}\,(c_1|c_2, c_3, \ldots, c_N)\,|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) \tag{4.20}$$

$$= \mathrm{E}\,(c_1|c_2, c_3, \ldots, c_N) \tag{4.21}$$

We use the "tower property" of conditional expectation on the second line. (In order to project on a target space, we first project on a larger space, and then project the result of this on the target space.) On the third line we use $\mathbf{y}_i = \mathbf{L}_i c_i + \mathbf{d}_i$ so that conditioning on the $\mathbf{y}_i$'s and the $\mathbf{d}_i$'s is the same as conditioning on the c_i's and the $\mathbf{d}_i$'s. The fourth line is due to zero correlation between the c_i's and the $\mathbf{d}_i$'s, and the last line exploits the fact that the composites are determined fully by the indicators. So because $\mathrm{E}(\mathbf{y}_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) = \mathrm{E}(\mathbf{L}_1 c_1 + \mathbf{d}_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) = \mathbf{L}_1\mathrm{E}(c_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N)$ we have

$$\mathrm{E}\,(\mathbf{y}_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) = \mathbf{L}_1\mathrm{E}\,(c_1|c_2, c_3, \ldots, c_N) \tag{4.22}$$

In other words, the best (least squares) predictor of a block of indicators given other blocks is determined by the best predictor of the composite of that block given the composites of the other blocks, together with the loadings on the composite. This

contrasts rather strongly with the model Wold used, with latent factors/variables $\mathbf{f}$. Here instead of $\mathbf{L}_1\mathrm{E}(c_1|c_2, c_3, \ldots, c_N)$ we have

$$\mathrm{E}\,(\mathbf{y}_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) = \boldsymbol{\lambda}_1\mathrm{E}(\mathrm{E}\,(f_1|f_2, f_3, \ldots, f_N)\,|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) \tag{4.23}$$

Basically, we can follow the sequence of steps as above for the composites except the penultimate step, from (4.20) to (4.21). I would maintain that the model as specified answers more truthfully to the fundamental principle of soft modeling than the basic design.

4.3 Estimation Issues

We will assume that we have the outcome of a Consistent and Asymptotically Normal (CAN-)estimator for $\boldsymbol{\Sigma}$. One can think of the sample covariance matrix of a random sample from a population with covariance matrix $\boldsymbol{\Sigma}$ and finite fourth-order moments (the latter is sufficient for asymptotic normality, consistency requires finite second-order moments only). The estimators to be described are all (locally) smooth functions of the CAN-estimator for $\boldsymbol{\Sigma}$, hence they are CAN as well.

We will use a step-wise approach: first the weights, then the loadings and the correlations between the composites, and finally the structural form coefficients. Each step uses a procedure that is essentially non-iterative, or if it iterates, it is very fast. So no explicit overall fit-criterion, although one could interpret the approach as the first iterate in a block relaxation program that aims to optimize a positive combination of target functions appropriate for each step. The view that a lack of an overall criterion to be optimized is a major flaw is ill-founded. Estimators should be compared on the basis of their distribution functions, the extent to which they satisfy computational desiderata, and the induced quality of the predictions. There is no theorem, and their cannot be one, to the effect that estimators that optimize a function are better than those that are not so motivated. For composites a proper comparison between the "step-wise" (partial) and the "global" approaches is still open. Of the issues to be addressed two stand out: *efficiency* in case of a proper, correct specification, and *robustness* with respect to distributional assumptions and specification errors (the optimization of a global fitting function that takes each and every structural constraint seriously may not be as robust to specification errors as a step-wise procedure).

4.3.1 Estimation of Weights, Loadings, and Correlations

The only issue of some substance in this section is the estimation of the weights. Once they are available, estimates for the loadings and correlations present themselves: the latter are estimated via the correlation between the composites, the

former by a regression of each block on its corresponding composite. One could devise more intricate methods but in this stage there seems little point in doing so.

To estimate the weights we will use generalized Canonical Variables (CV's) analysis.[9] This includes of course the approach proposed by Wold, the so-called mode B estimation method. Composites simply *are* canonical variables. Any method that yields CV's matches naturally, "perfectly," with the model. We will describe some of the methods while applying them to $\mathbf{\Sigma}$ and show that they do indeed yield the weights. A continuity argument then gives that when they are applied to the CAN-estimator for $\mathbf{\Sigma}$ the estimators for the weights are consistent as well. Local differentiability leading to asymptotic normality is not difficult to establish either.[10]

For notational ease we will employ a composites model with three blocks, $N = 3$, but that is no real limitation. Now consider the covariance matrix, denoted by $\mathbf{R}(\mathbf{v})$, of $\mathbf{v}_1^\intercal \mathbf{y}_1$, $\mathbf{v}_2^\intercal \mathbf{y}_2$, and $\mathbf{v}_3^\intercal \mathbf{y}_3$ where each $\mathbf{v}_i$ is normalized ($\mathrm{var}(\mathbf{v}_i^\intercal \mathbf{y}_i) = 1$). So

$$\mathbf{R}(\mathbf{v}) := \begin{bmatrix} 1 & \mathbf{v}_1^\intercal \mathbf{\Sigma}_{12} \mathbf{v}_2 & \mathbf{v}_1^\intercal \mathbf{\Sigma}_{13} \mathbf{v}_3 \\ \mathbf{v}_1^\intercal \mathbf{\Sigma}_{12} \mathbf{v}_2 & 1 & \mathbf{v}_2^\intercal \mathbf{\Sigma}_{23} \mathbf{v}_3 \\ \mathbf{v}_1^\intercal \mathbf{\Sigma}_{13} \mathbf{v}_3 & \mathbf{v}_2^\intercal \mathbf{\Sigma}_{23} \mathbf{v}_3 & 1 \end{bmatrix}. \tag{4.24}$$

Canonical variables are composites whose correlation matrix has "maximum distance" to the identity matrix of the same size. They are "collectively maximally correlated." The term is clearly ambiguous for more than two blocks. One program that would seem to be natural is to maximize with respect to $\mathbf{v}$

$$z(\mathbf{v}) := \mathrm{abs}(R_{12}) + \mathrm{abs}(R_{13}) + \mathrm{abs}(R_{23}) \tag{4.25}$$

subject to the usual normalizations. Since

$$\mathrm{abs}\left(R_{ij}\right) = \mathrm{abs}\left(r_{ij}\right) \cdot \mathrm{abs}\left(\mathbf{v}_i^\intercal \mathbf{\Sigma}_{ii} \mathbf{w}_i\right) \cdot \mathrm{abs}\left(\mathbf{v}_j^\intercal \mathbf{\Sigma}_{jj} \mathbf{w}_j\right) \tag{4.26}$$

we know, thanks to Cauchy–Schwarz, that

$$\mathrm{abs}\left(\mathbf{v}_i^\intercal \mathbf{\Sigma}_{ii} \mathbf{w}_i\right) = \mathrm{abs}\left(\mathbf{v}_i^\intercal \mathbf{\Sigma}_{ii}^{\frac{1}{2}} \mathbf{\Sigma}_{ii}^{\frac{1}{2}} \mathbf{w}_i\right) \leq \sqrt{\mathbf{v}_i^\intercal \mathbf{\Sigma}_{ii}^{\frac{1}{2}} \mathbf{\Sigma}_{ii}^{\frac{1}{2}} \mathbf{v}_i \cdot \mathbf{w}_i^\intercal \mathbf{\Sigma}_{ii}^{\frac{1}{2}} \mathbf{\Sigma}_{ii}^{\frac{1}{2}} \mathbf{w}_i} \tag{4.27}$$

$$= \sqrt{\mathbf{v}_i^\intercal \mathbf{\Sigma}_{ii} \mathbf{v}_i \cdot \mathbf{w}_i^\intercal \mathbf{\Sigma}_{ii} \mathbf{w}_i} = 1 \tag{4.28}$$

with equality if and only if $\mathbf{v}_i = \mathbf{w}_i$ (ignoring irrelevant sign differences). Observe that the upper bound can be reached for $\mathbf{v}_i = \mathbf{w}_i$ for all terms in which $\mathbf{v}_i$ appears,

[9] Kettenring (1971) is *the* reference for generalized canonical variables.

[10] These statements are admittedly a bit nonchalant if not cavalier, but there seems little to gain by elaborating on them.

so maximization of the sum of the absolute correlations gives $\mathbf{w}$. A numerical, iterative routine[11] suggests itself by noting that the optimal $\mathbf{v}_1$ satisfies the first order condition

$$0 = \operatorname{sgn}(R_{12}) \cdot \boldsymbol{\Sigma}_{12}\mathbf{v}_2 + \operatorname{sgn}(R_{13}) \cdot \boldsymbol{\Sigma}_{13}\mathbf{v}_3 - l_1 \boldsymbol{\Sigma}_{11}\mathbf{v}_1 \tag{4.29}$$

where l_1 is a Lagrange multiplier (for the normalization), and two other quite similar equations for $\mathbf{v}_2$ and $\mathbf{v}_3$. So with arbitrary starting vectors one could solve the equations recursively for $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, respectively, updating them after each complete round or at the first opportunity, until they settle down at the optimal value. Note that each update of $\mathbf{v}_1$ is obtainable by a regression of a "sign-weighted sum"

$$\operatorname{sgn}(R_{12}) \cdot \mathbf{v}_2^{\intercal}\mathbf{y}_2 + \operatorname{sgn}(R_{13}) \cdot \mathbf{v}_3^{\intercal}\mathbf{y}_3 \tag{4.30}$$

on $\mathbf{y}_1$, and analogously for the other weights. This happens to be the classical form of PLS' mode B.[12] For $\boldsymbol{\Sigma}$ we do not need many iterations, to put it mildly: the update of $\mathbf{v}_1$ is already $\mathbf{w}_1$, as straightforward algebra will easily show. And similarly for the other weight vectors. In other words, we have in essentially just one iteration a fixed point for the mode B equations that is precisely $\mathbf{w}$.

If we use the correlations themselves in the recursions instead of just their signs, we regress the "correlation weighted sum"

$$R_{12} \cdot \mathbf{v}_2^{\intercal}\mathbf{y}_2 + R_{13} \cdot \mathbf{v}_3^{\intercal}\mathbf{y}_3 \tag{4.31}$$

on $\mathbf{y}_1$ (and analogously for the other weights), and end up with weights that maximize

$$z(\mathbf{v}) := R_{12}^2 + R_{13}^2 + R_{23}^2 \tag{4.32}$$

the simple sum of the squared correlations. Again, with the same argument, the optimal value is $\mathbf{w}$.

Observe that for this $z(\mathbf{v})$ we have

$$\operatorname{tr}\left(\mathbf{R}^2\right) = 2 \cdot z(\mathbf{v}) + 3 = \sum_{i=1}^{3} \gamma_i^2 \tag{4.33}$$

where $\gamma_i := \gamma_i(\mathbf{R}(\mathbf{v}))$ is the ith eigenvalue of $\mathbf{R}(\mathbf{v})$. We can take other functions of the eigenvalues, in order to maximize the difference between $\mathbf{R}(\mathbf{v})$ and the identity matrix of the same order. Kettenring (1971) discusses a number of alternatives. One

[11] With $\boldsymbol{\Sigma}$ one does not really need an iterative routine of course: $\boldsymbol{\Sigma}_{ij} = r_{ij}\boldsymbol{\Sigma}_{ii}\mathbf{w}_i\mathbf{w}_j^{\intercal}\boldsymbol{\Sigma}_{jj}$ can be solved directly for the weights (and the correlation). But in case we just have an estimate, an algorithm comes in handy.

[12] See chapter two of Dijkstra (1981).

of them minimizes the product of the γ_i's, the determinant of $\mathbf{R}(\mathbf{v})$, also known as the generalized variance. The program is called GENVAR. Since $\sum_{i=1}^{N} \gamma_i$ is always N (three in this case) for every choice of $\mathbf{v}$, GENVAR tends to make the eigenvalues as diverse as possible (as opposed to the identity matrix where they are all equal to one). The determinant of $\mathbf{R}(\mathbf{v})$ equals $\left(1 - R_{23}^2\right)$, which is independent of $\mathbf{v}_1$, times

$$1 - \begin{bmatrix} R_{12} & R_{13} \end{bmatrix} \begin{bmatrix} 1 & R_{23} \\ R_{23} & 1 \end{bmatrix}^{-1} \begin{bmatrix} R_{12} \\ R_{13} \end{bmatrix} \tag{4.34}$$

$$= 1 - \left(\mathbf{v}_1^{\intercal} \boldsymbol{\Sigma}_{11} \mathbf{w}_1\right)^2 \begin{bmatrix} r_{12} \mathbf{v}_2^{\intercal} \boldsymbol{\Sigma}_{22} \mathbf{w}_2 & r_{13} \mathbf{v}_3^{\intercal} \boldsymbol{\Sigma}_{33} \mathbf{w}_3 \end{bmatrix} \begin{bmatrix} 1 & R_{23} \\ R_{23} & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{12} \mathbf{v}_2^{\intercal} \boldsymbol{\Sigma}_{22} \mathbf{w}_2 \\ r_{13} \mathbf{v}_3^{\intercal} \boldsymbol{\Sigma}_{33} \mathbf{w}_3 \end{bmatrix}$$

where the last quadratic form does not involve $\mathbf{v}_1$ either and we have with the usual argument that GENVAR produces $\mathbf{w}$ also. See Kettenring (1971) for an appropriate iterative routine (this involves the calculation of ordinary canonical variables of $\mathbf{y}_i$ and the $(N-1)$-vector consisting of the other composites).

Another program is MAXVAR, which maximizes the largest eigenvalue. For every $\mathbf{v}$ one can calculate the linear combination of the corresponding composites that best predicts or explains them: the first principal component of $\mathbf{R}(\mathbf{v})$. No other set is as well explained by the first principal component as the MAXVAR composites. *There is an explicit solution here, no iterative routine is needed for the estimate of* $\boldsymbol{\Sigma}$, if one views the calculation of eigenvectors as non-iterative, see Kettenring (1971) for details.[13] One can show again that the optimal $\mathbf{v}$ equals $\mathbf{w}$ when MAXVAR is applied to Σ, although this requires a bit more work than for GENVAR (due to the additional detail needed to describe the solution).

As one may have expected, there is also MINVAR, the program aimed at minimizing the smallest eigenvalue (Kettenring 1971). The result is a set of composites with the property that no other set is "as close to linear dependency" as the MINVAR set. We also have an explicit solution, and $\mathbf{w}$ is optimal again.

4.3.2 Mode A and Mode B

In the previous subsection we recalled that mode B generates weight vectors by iterating regressions of certain weighted sums of composites on blocks. There is also mode A (and a mode C which we will not discuss), where weights are found iteratively by reversing the regressions: now blocks are regressed on weighted sums of composites. The algorithm generally converges, and the probability limits of

[13] This is true when applied to the estimate for $\boldsymbol{\Sigma}$ as well. With an estimate the other methods will usually require more than just one iteration (and all programs will produce different results, although the differences will tend to zero in probability).

the weights can be found as before by applying mode A to $\boldsymbol{\Sigma}$. If we denote the probability limits (plims) of the (normalized) mode A weights by $\widetilde{\mathbf{w}}_i$, we have in the generic case that $\mathbf{y}_i$ is regressed on $\sum_{j\neq i}\text{sgn}(\text{cov}(\widetilde{\mathbf{w}}_i^\intercal\mathbf{y}_i, \widetilde{\mathbf{w}}_j^\intercal\mathbf{y}_j))\cdot\widetilde{\mathbf{w}}_j^\intercal\mathbf{y}_j$ so that

$$\widetilde{\mathbf{w}}_i \propto \sum_{j\neq i}\text{sgn}(\text{cov}(\widetilde{\mathbf{w}}_i^\intercal\mathbf{y}_i, \widetilde{\mathbf{w}}_j^\intercal\mathbf{y}_j))\cdot\boldsymbol{\Sigma}_{ij}\widetilde{\mathbf{w}}_j \tag{4.35}$$

$$= \sum_{j\neq i}\text{sgn}(\text{cov}(\widetilde{\mathbf{w}}_i^\intercal\mathbf{y}_i, \widetilde{\mathbf{w}}_j^\intercal\mathbf{y}_j))\cdot r_{ij}\mathbf{L}_i\mathbf{L}_j^\intercal\widetilde{\mathbf{w}}_j \tag{4.36}$$

$$= \mathbf{L}_i\cdot\left(\sum_{j\neq i}\text{sgn}(\text{cov}(\widetilde{\mathbf{w}}_i^\intercal\mathbf{y}_i, \widetilde{\mathbf{w}}_j^\intercal\mathbf{y}_j))\cdot r_{ij}\mathbf{L}_j^\intercal\widetilde{\mathbf{w}}_j\right) \tag{4.37}$$

and so

$$\widetilde{\mathbf{w}}_i \propto \mathbf{L}_i, \text{infact}\widetilde{\mathbf{w}}_i = \frac{1}{\sqrt{\mathbf{L}_i^\intercal\boldsymbol{\Sigma}_{ii}\mathbf{L}_i}}\mathbf{L}_i \tag{4.38}$$

An immediate consequence is that the plim of mode A's correlation, $\widetilde{r}_{ij}$, equals

$$\widetilde{r}_{ij} = \widetilde{\mathbf{w}}_i^\intercal\left(r_{ij}\mathbf{L}_i\mathbf{L}_j^\intercal\right)\widetilde{\mathbf{w}}_j = r_{ij}\cdot\frac{\mathbf{L}_i^\intercal\mathbf{L}_i}{\sqrt{\mathbf{L}_i^\intercal\boldsymbol{\Sigma}_{ii}\mathbf{L}_i}}\frac{\mathbf{L}_j^\intercal\mathbf{L}_j}{\sqrt{\mathbf{L}_j^\intercal\boldsymbol{\Sigma}_{jj}\mathbf{L}_j}} \tag{4.39}$$

One would expect this to be smaller in absolute value than r_{ij}, and so it is, since

$$\frac{\mathbf{L}_i^\intercal\mathbf{L}_i}{\sqrt{\mathbf{L}_i^\intercal\boldsymbol{\Sigma}_{ii}\mathbf{L}_i}} = \frac{\mathbf{w}_i^\intercal\boldsymbol{\Sigma}_{ii}^2\mathbf{w}_i}{\sqrt{\mathbf{w}_i^\intercal\boldsymbol{\Sigma}_{ii}^3\mathbf{w}_i}} \tag{4.40}$$

$$= \frac{\mathbf{w}_i^\intercal\boldsymbol{\Sigma}_{ii}^{1/2}\Sigma_{ii}^{3/2}\mathbf{w}_i}{\sqrt{\mathbf{w}_i^\intercal\Sigma_{ii}^3\mathbf{w}_i}} \leq \frac{\sqrt{\mathbf{w}_i^\intercal\boldsymbol{\Sigma}_{ii}\mathbf{w}_i}\sqrt{\mathbf{w}_i^\intercal\Sigma_{ii}^3\mathbf{w}_i}}{\sqrt{\mathbf{w}_i^\intercal\Sigma_{ii}^3\mathbf{w}_i}} = 1 \tag{4.41}$$

because of Cauchy–Schwarz. In general, mode A's composites, $\widetilde{\mathbf{c}}$, will not satisfy $\mathbf{B}\widetilde{\mathbf{c}}_{\text{endo}} = \mathbf{C}\widetilde{\mathbf{c}}_{\text{exo}} + \widetilde{\mathbf{z}}$ with $\widetilde{\mathbf{z}}$ uncorrelated with $\widetilde{\mathbf{c}}_{\text{exo}}$. Observe that we have $\widetilde{r}_{ij} = r_{ij}$ when and only when $\Sigma_{ii}\mathbf{w}_i \propto \mathbf{w}_i$ & $\boldsymbol{\Sigma}_{jj}\mathbf{w}_j \propto \mathbf{w}_j$, in which case each composite is *a principal component* of its corresponding block.

For the plim of the loadings, $\widetilde{\mathbf{L}}_i$, we note

$$\widetilde{\mathbf{L}}_i = \frac{1}{\sqrt{\mathbf{L}_i^\intercal\boldsymbol{\Sigma}_{ii}\mathbf{L}_i}}\boldsymbol{\Sigma}_{ii}\mathbf{L}_i \tag{4.42}$$

So mode A's loading vector is in the limit proportional to the true vector when and only when $\boldsymbol{\Sigma}_{ii}\mathbf{w}_i \propto \mathbf{w}_i$.

To summarize:

1. *Mode A will tend to underestimate the correlations in absolute value.*[14]
2. *The plims of the correlations between the composites for Mode A and Mode B will be equal when and only when each composite is a principal component of its corresponding block, in which case we have a perfect match between a model and two modes as far as the relationships between the composites are concerned.*
3. *The plims of the loading vectors for Mode A and Mode B will be proportional when and only when each composite is a principal component of its corresponding block.*

A final observation: we can "correct" mode A to yield the right results in the general situation via

$$\frac{\boldsymbol{\Sigma}_{ii}^{-1}\widetilde{\mathbf{w}}_i}{\sqrt{\widetilde{\mathbf{w}}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{ii}^{-1}\widetilde{\mathbf{w}}_i}} = \mathbf{w}_i \tag{4.43}$$

and

$$\frac{\widetilde{\mathbf{w}}_i}{\sqrt{\widetilde{\mathbf{w}}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{ii}^{-1}\widetilde{\mathbf{w}}_i}} = \mathbf{L}_i \tag{4.44}$$

4.3.3 Estimation of the Structural Equations

Given the estimate of $\mathbf{R}_c$ we now focus on the estimation of $\mathbf{B}\mathbf{c}_{\text{endo}} = \mathbf{C}\mathbf{c}_{\text{exo}} + \mathbf{z}$. We have exclusion constraints for the structural form matrices $\mathbf{B}$ and $\mathbf{C}$, i.e., certain coefficients are a priori known to be zero. There are no restrictions on $\text{cov}(\mathbf{z})$, or if there are, we will ignore them here (for convenience, not as a matter of principle). This seems to exclude Wold's recursive system where the elements of $\mathbf{B}$ on one side of the diagonal are zero, and the equation-residuals are *un*correlated. But we can always regress the first endogenous composite $c_{\text{endo},1}$ on $\mathbf{c}_{\text{exo}}$, and $c_{\text{endo},2}$ on $[c_{endo,1}; \mathbf{c}_{\text{exo}}]$, and $c_{\text{endo},3}$ on $[c_{\text{endo},1}; c_{\text{endo},2}; \mathbf{c}_{\text{exo}}]$ et cetera. The ensuing residuals are *by construction* uncorrelated with the explanatory variables in their corresponding equations, and by implication they are mutually uncorrelated. In a sense, there are no assumptions here, the purpose of the exercise (prediction of certain variables using a specific set of predictors) determines the regression to be performed; there is also no identifiability issue.[15]

[14] A working paper version of this paper said that the elements of the mode A loading vector would always be "larger" than the corresponding true values. I am obliged to Michel Tenenhaus for making me realize that the statement was not true.

[15] See Dijkstra (2014) for further discussion of Wold's approach to modeling. There is a subtle issue here. One could generate a sample from a system with $\mathbf{B}$ lower-triangular, a full matrix $\mathbf{C}$

Now consider $\mathbf{P}$, the regression matrix obtained from regressing the (estimated) endogenous composites on the (estimated) exogenous composites. It estimates $\mathbf{\Pi}$, the reduced form matrix $\mathbf{B}^{-1}\mathbf{C}$. We will use $\mathbf{P}$, and possible other functions of $\mathbf{R}_c$, to estimate the free elements of $\mathbf{B}$ and $\mathbf{C}$. There is no point in trying when $\mathbf{\Pi}$ is compatible with different values of the structural form matrices. So the crucial question is whether $\mathbf{\Pi} = \boldsymbol{B}^{-1}\mathbf{C}$, or equivalently $\mathbf{B}\Pi = \mathbf{C}$, can be solved uniquely for the free elements of $\mathbf{B}$ and $\mathbf{C}$. Take the ith equation[16]

$$\mathbf{B}_{i\cdot}\mathbf{\Pi} = \boldsymbol{C}_{i\cdot} \tag{4.45}$$

where the ith row of $\mathbf{B}$, $\mathbf{B}_{i\cdot}$, has 1 in the ith entry (normalization) and possibly some zeros elsewhere, and where the ith row of $\mathbf{C}$, $\mathbf{C}_{i\cdot}$, may also contain some zeros. The free elements in $\mathbf{C}_{i\cdot}$ are given when those in $\mathbf{B}_{i\cdot}$ are known, and the latter are to be determined by the zeros in $\mathbf{C}_{i\cdot}$. More precisely

$$\mathbf{B}_{(i,k:B_{ik}\text{ free or unit})} \cdot \mathbf{\Pi}_{(k:B_{ik}\text{ free or unit},\, j:C_{ij}=0)} = 0 \tag{4.46}$$

So we have a submatrix of $\mathbf{\Pi}$, the rows correspond with the free elements (and the unit) in the ith row of $\mathbf{B}$, and the columns with the zero elements in the ith row of $\mathbf{C}$. This equation determines $\mathbf{B}_{(i,k:B_{ik}\text{ free or unit})}$ uniquely, apart from an irrelevant nonzero multiple, *when and only when* the particular submatrix of $\mathbf{\Pi}$ has a rank equal to its number of rows minus one. This is just the number of elements to be estimated in the ith row of $\mathbf{B}$. To have this rank requires the submatrix to have at least as many columns. So a little thought will give that a necessary condition for unique solvability, *identifiability*, is that we must have as least as many exogenous composites in the system as coefficients to be estimated in any one equation. We emphasize that this *order condition* as it is traditionally called is indeed nothing more than necessary.[17] The *rank condition* is both necessary and sufficient.

---

and a full, non-diagonal covariance matrix for $\mathbf{z}$. Then no matter how large the sample size, we can never retrieve the coefficients (apart from those of the first equation which are just regression coefficients). The regressions for the other equations would yield values different from those we used to generate the observations, since the zero correlation between their equation-residuals would be incompatible with the non-diagonality of cov($\mathbf{z}$).

[16] What follows will be old hat for econometricians, but since non-recursive systems are relatively new for PLS-practitioners, some elaboration could be meaningful.

[17] As an example consider a square $\mathbf{B}$ with units on the diagonal but otherwise unrestricted, and a square $\mathbf{C}$ of the same dimensions, containing zeros only except the last row, where all entries are free. The order condition applies to all equations but the last, but *none* of the coefficients can be retrieved from $\mathbf{\Pi}$. This matrix is, however, severely restricted: it has rank one. How to deal with this and similar situations is handled by Bekker et al. (1994).

A very simple example, which we will use in a small Monte Carlo study in the next subsection is as follows. Let

$$\begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix} \begin{bmatrix} c_{\text{endo},1} \\ c_{\text{endo},2} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & 0 & 0 \\ 0 & 0 & c_{23} & c_{24} \end{bmatrix} \begin{bmatrix} c_{\text{exo},1} \\ c_{\text{exo},2} \\ c_{\text{exo},3} \\ c_{\text{exo},4} \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \tag{4.47}$$

with $1 - b_{12}b_{21} \neq 0$. The order conditions are satisfied: each equation has three free coefficients and there are four exogenous composites.[18] Note that

$$\mathbf{\Pi} = \frac{1}{1 - b_{12}b_{21}} \begin{bmatrix} c_{11} & c_{12} & -b_{12}c_{23} & -b_{12}c_{24} \\ -b_{21}c_{11} & -b_{21}c_{12} & c_{23} & c_{24} \end{bmatrix} \tag{4.48}$$

The submatrix of $\mathbf{\Pi}$ relevant for an investigation into the validity of the rank condition for the first structural form equation is

$$\begin{bmatrix} \Pi_{13} & \Pi_{14} \\ \Pi_{23} & \Pi_{24} \end{bmatrix} = \frac{1}{1 - b_{12}b_{21}} \begin{bmatrix} -b_{12}c_{23} & -b_{12}c_{24} \\ c_{23} & c_{24} \end{bmatrix} \tag{4.49}$$

It should have rank one, and it does so in the generic case, since its first row is a multiple of its second row.[19] Note that we cannot have both c_{23} and c_{24} zero. Clearly, b_{12} can be obtained from $\mathbf{\Pi}$ via $-\Pi_{13}/\Pi_{23}$ or via $-\Pi_{14}/\Pi_{24}$. A similar analysis applies to the second structural form equation. We note that the model imposes two constraints on $\mathbf{\Pi}$: $\Pi_{11}\Pi_{22} - \Pi_{12}\Pi_{21} = 0$ and $\Pi_{13}\Pi_{24} - \Pi_{14}\Pi_{23} = 0$, in agreement with the fact that the 8 reduced form coefficients can be expressed in terms of 6 structural form parameters. For an extended analysis of the number and type of constraints that a structural form imposes on the reduced form see Bekker and Dijkstra (1990) and Bekker et al. (1994).

It will be clear that the estimate $\mathbf{P}$ of $\mathbf{\Pi}$ will not in general satisfy the rank conditions (although we do expect them to be close for sufficiently large samples), and using either $-P_{13}/P_{23}$ or $-P_{14}/P_{24}$ as an estimate for b_{12} will give different answers. Econometric methods construct explicitly or implicitly compromises between the possible estimates. 2SLS, as discussed above is one of them. See Dijkstra and Henseler (2015a,b) for a specification of the relevant formula (formula

[18] With 2SLS $c_{\text{endo},2}$ in the first equation is in the first stage replaced by its regression on the four exogenous variables. In the second stage we regress $c_{\text{endo},1}$ on the replacement for $c_{\text{endo},2}$ and two exogenous variables. So the regression matrix with three columns in this stage is spanned by four exogenous columns, and we should be fine in general. If there were four exogenous variables on the right-hand side, the regression matrix in the second stage would have five columns, spanned by only four exogenous columns, the matrix would not be invertible and 2SLS (and all other methods aiming for consistency) would break down.

[19] For more general models one could ask MATLAB, say, to calculate the rank of the matrices, evaluated for arbitrary values. A very pragmatic approach would be to just run 2SLS. If it breaks down and gives a singularity warning, one should analyze the situation. Otherwise you are fine.

(23)) for 2SLS that honors the motivation via two regressions. Here we will outline another approach based on Dijkstra (1989) that is close to the discussion about identifiability.

Consider a row vector[20] with ith subvector $\mathbf{B}_{i\cdot}\mathbf{P} - \mathbf{C}_{i\cdot}$. If $\mathbf{P}$ would equal $\mathbf{\Pi}$ we could get the free coefficients by making $\mathbf{B}_{i\cdot}\mathbf{P} - \mathbf{C}_{i\cdot}$ zero. But that will not be the case. So we could decide to choose values for the free coefficients that make each $\mathbf{B}_{i\cdot}\mathbf{P} - \mathbf{C}_{i\cdot}$ as "close to zero as possible." One way to implement that is to minimize a suitable quadratic form subject to the exclusion constraints and normalizations. We take

$$\left(\text{vec}\left[(\mathbf{BP} - \mathbf{C})^{\intercal}\right]\right)^{\intercal} \cdot \left(\mathbf{W} \otimes \widehat{\mathbf{R}}_{\text{exo}}\right) \cdot \text{vec}\left[(\mathbf{BP} - \mathbf{C})^{\intercal}\right] \tag{4.50}$$

Here $\otimes$ stands for Kronecker's matrix multiplication symbol, $\widehat{\mathbf{R}}_{\text{exo}}$ is the estimated p.d. correlation matrix of the estimated exogenous composites, $\mathbf{W}$ is a p.d. matrix with as many rows and columns as there are endogenous composites, and the operator "vec" stacks the columns of its matrix-argument one underneath the other, starting with the first. If we take a diagonal matrix $\mathbf{W}$ the quadratic form disintegrates into separate quadratic forms, one for each subvector, and minimization yields in fact 2SLS estimates. A non-diagonal $\mathbf{W}$ tries to exploit information about the covariances between the subvectors. For the classical econometric simultaneous equation model it is true that $\text{vec}\left[(\mathbf{BP} - \mathbf{C})^{\intercal}\right]$ is asymptotically normal with zero mean and covariance matrix $\text{cov}(\mathbf{z}) \otimes \mathbf{R}_{\text{exo}}^{-1}$ divided by the sample size, adapting the notation somewhat freely. General estimation theory tells us to use the inverse of an estimate of this covariance matrix in order to get asymptotic efficiency. So $\mathbf{W}$ should be the inverse of an estimate for $\text{cov}(\mathbf{z})$. The latter is traditionally estimated by the obvious estimate based on 2SLS. Note that the covariances between the structural form residuals drive the extent to which the various optimizations are integrated. There is no or little gain when there is no or little correlation between the elements of $\mathbf{z}$. This more elaborate method is called 3SLS.

We close with some observations. Since the quadratic form in the parameters is minimized subject to zero constraints and normalizations only, there is an explicit solution, see Dijkstra (1989, section 5), for the formulae.[21] If the fact that the weights are estimated can be ignored, there is also an explicit expression for the asymptotic covariance matrix, both for 2SLS and 3SLS. But if the sampling variation in the weights does matter, this formula may not be accurate and 3SLS may not be more efficient than 2SLS. Both methods are essentially non-iterative and very fast, and therefore suitable candidates for bootstrapping. One potential advantage of 2SLS over 3SLS is that it may be more robust to model specification errors, because as opposed to its competitor, it estimates equation by equation, so that an error in one equation need not affect the estimation of the others.

[20] This is in fact, see below: $\left(\text{vec}\left[(\mathbf{BP} - \mathbf{C})^{\intercal}\right]\right)^{\intercal}$.

[21] For the standard approach and the classical formulae, see, e.g., Ruud (2000)

4.3.4 Some Monte Carlo Results

We use the setup from Dijkstra and Henseler (2015a,b) adapted to the present setting. We have

$$\begin{bmatrix} 1 & -0.25 \\ -0.50 & 1 \end{bmatrix} \begin{bmatrix} c_{\text{endo},1} \\ c_{\text{endo},2} \end{bmatrix} = \begin{bmatrix} -0.30 & 0.50 & 0 & 0 \\ 0 & 0 & 0.50 & 0.25 \end{bmatrix} \begin{bmatrix} c_{\text{exo},1} \\ c_{\text{exo},2} \\ c_{\text{exo},3} \\ c_{\text{exo},4} \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \tag{4.51}$$

All variables have zero mean, and we will take them jointly normal. $\text{Cov}(\mathbf{c}_{\text{exo}})$ has ones on the diagonal and 0.50 everywhere else; the variances of the endogenous composites are also one and we take $\text{cov}(c_{\text{endo},1}, c_{\text{endo},2}) = \sqrt{0.50}$. The values as specified imply for the covariance matrix for the structural form residuals $\mathbf{z}$:

$$\text{cov}(\mathbf{z}) = \begin{bmatrix} 0.5189 & -0.0295 \\ -0.0295 & 0.1054 \end{bmatrix} \tag{4.52}$$

Note that the correlation between z_1 and z_2 is rather small, -0.1261, so the setup has the somewhat unfortunate consequence to potentially favor 2SLS. The R-squared for the first reduced form equation is 0.3329 and for the second reduced form equation this is 0.7314.

Every composite is built up by three indicators, with a covariance matrix that has ones on the diagonal and 0.49 everywhere else. This is compatible with a one-factor model for each vector of indicators but we have no use nor need for that interpretation here.

The composites $(c_{\text{exo},1}, c_{\text{exo},2}, c_{\text{exo},3}, c_{\text{exo},4}, c_{\text{endo},1}, c_{\text{endo},2})$ need weights. For the first and fourth we take weights proportional to $[1, 1, 1]$. For the second and fifth the weights are proportional to $[1, 2, 3]$ and for the third and sixth they are proportional to $[1, 4, 9]$. There are no deep thoughts behind these choices.

We get the following weights (rounded to two decimals for readability): $[0.41, 0.41, 0.41]$ for blocks one and four, $[0.20, 0.40, 0.60]$ for blocks two and five, and $[0.08, 0.33, 0.74]$ for blocks three and six.

The loadings are now given as well: $[0.81, 0.81, 0.81]$ for blocks one and four, $[0.69, 0.80, 0.90]$ for blocks two and five, and $[0.61, 0.74, 0.95]$ for blocks three and six.

One can now calculate the 18 by 18 covariance/correlation matrix $\mathbf{\Sigma}$ and its unique p.d. matrix square root $\mathbf{\Sigma}^{1/2}$. We generate samples of size 300, which appears to be relatively modest given the number of parameters to estimate. A sample of size 300 is obtained via $\mathbf{\Sigma}^{1/2} \times \text{randn}(18, 300)$. We repeat this ten thousand times, each time estimating the weights via MAXVAR,[22] the loadings

[22] One might as well have used mode B of course, or any of the other canonical variables approaches. There is no *fundamental* reason to prefer one to the other. MAXVAR was available, and is essentially non-iterative.

via regressions and the correlations in the obvious way, and all structural form parameters via 2SLS and 3SLS using standardized indicators.[23]

The loadings and weights are on the average slightly underestimated, see Dijkstra (2015) for some of the tables: when rounded to two decimals the difference is at most 0.01. The standard deviations of the weights estimators for the endogenous composites are either the largest or the smallest: for the weights of $c_{\text{endo},1}$ we have resp. $[0.12, 0.12, 0.11]$ and for $c_{\text{endo},2}$ $[0.04, 0.04, 0.04]$; the standard deviations for the weights of the exogenous composites are, roughly, in between. And similarly for the standard deviations for the loadings estimators: for the loadings on $c_{\text{endo},1}$ we have resp. $[0.08, 0.07, 0.05]$ and for $c_{\text{endo},2}$ $[0.05, 0.04, 0.01]$; the standard deviations for the loadings on the exogenous composites are again, roughly, in between.

The following table gives the results for the coefficients in **B** and **C**, rounded to two decimals:

	Value	Mean 2SLS	Mean 3SLS	std 2SLS	std 3SLS
b_{12}	−0.25	−0.26	−0.26	0.08	0.08
b_{21}	−0.50	−0.50	−0.50	0.05	0.05
c_{11}	−0.30	−0.29	−0.28	0.05	0.05
c_{12}	+0.50	+0.49	+0.49	0.06	0.06
c_{23}	+0.50	+0.49	+0.49	0.03	0.03
c_{24}	+0.25	+0.25	+0.25	0.03	0.03

Clearly, for the model at hand 3SLS has nothing to distinguish itself positively from 2SLS[24] (its standard deviations are only smaller than those of 2SLS when we use three decimals). This might be different when the structural form residuals are materially correlated.

We also calculated, not shown, for each of the 10,000 samples of size 300 the theoretical (asymptotic) standard deviations for the 3SLS estimators. They are all on the average 0.01 smaller than the values in the table, they are relatively stable, with standard deviations ranging from 0.0065 for b_{12} to 0.0015 for c_{24}. They are not perfect but not really bad either.

It would be reckless to read too much into this small and isolated study, for one type of distribution. But the approach does appear to be feasible.

[23] The whole exercise takes about half a minute on a slow machine: 4CPU 2.40 Ghz; RAM 512 MB.

[24] It is remarkable that the accuracy of the 2SLS and 3SLS estimators is essentially as good, in three decimals, as those reported by Dijkstra and Henseler (2015a,b) for Full Information Maximum Likelihood (FIML) for the same model in terms of latent variables, i.e., FIML as applied to the *true* latent variable scores. See Table 2 on p. 18 there. When the latent variables are not observed directly but only via indicators, the performance of FIML clearly deteriorates (stds are doubled or worse).

4.4 Testing the Composites Model

In this section we sketch four more or less related approaches to test the appropriateness or usefulness of the model. In practice one might perhaps want to deploy all of them. Investigators will easily think of additional, "local" tests, like those concerning the signs or the order of magnitude of coefficients et cetera.

A thorny issue that should be mentioned here is *capitalization on chance*, which refers to the phenomenon that in practice one runs through cycles of model testing and adaptation until the current model tests signal that all is well according to popular rules-of-thumb.[25] This makes the model effectively stochastic, random. Taking a new sample and going through the cycles of testing and adjusting all over again may well lead to another model. But when we give estimates of the distribution functions of our estimators we imply that this helps to assess how the estimates will vary when other samples of the same size would be employed, while keeping the model *fixed*. It is tempting, but potentially very misleading, to ignore the fact that the sample (we/you, actually) favored a particular model after a (dedicated) model search, see Freedman et al. (1988), Dijkstra and Veldkamp (1988), Leeb and Pötscher (2006), and Freedman (2009)[26]. It is not clear at all how to properly validate the model on the very same data that gave it birth, while using *test* statistics as *design* criteria.[27] Treating the results conditional *on the sample at hand*, as purely descriptive (which in itself may be rather useful, Berk 2008), or testing the model on a fresh sample (e.g., a random subset of the data that was kept apart when the model was constructed), while bracing oneself for a possibly big disappointment, appear to be the best or most honest responses.

[25] "Capitalization on chance" is sometimes used when "small-sample-bias" is meant. That is quite something else.

[26] Freedman gives the following example. Let the 100×51 matrix $[\mathbf{y}, \mathbf{X}]$ consists of independent standard normals. So there is no (non-) linear relationship whatsoever. Still, a regression of $\mathbf{y}$ on $\mathbf{X}$ can be expected to yield an R-square of 0.50. On the average there will be 5 regression coefficients that are significant at 10%. If we keep the corresponding $\mathbf{X}$-columns in the spirit of "exploratory research" and discard the others, a regression could easily give a decent R-square and "dazzling t-statistics" (Freedman 2009, p.75). Note that here the "dedicated" model search consisted of merely two regression rounds. Just think of what one can accomplish with a bit more effort, see also, e.g., Dijkstra (1995).

[27] At one point I thought that "a way out" would be to condition on the set of samples that favor the chosen model using the same search procedure (Dijkstra and Veldkamp 1988): if the model search has led to the simplest true model, the conditional estimator distribution equals, asymptotically, the distribution that the practitioner reports. This conditioning would give substance to the retort given in practice that "we always condition on the given model." But the result referred to says essentially that we can ignore the search *if* we know it was *not* needed. So much for comfort. It is even a lot worse: Leeb and Pötscher (2006) show that convergence of the conditional distribution is only pointwise, not uniform, not even on compact subsets of the parameter space. The bootstrap cannot alleviate this problem, Leeb and Pötscher (2006), Dijkstra and Veldkamp (1988).

4.4.1 Testing Rank Restrictions on Submatrices

The covariance matrix of any subvector of $\mathbf{y}_i$ with any choice from the other indicators has rank one. So the corresponding regression matrix has rank one. To elaborate a bit, since $\mathrm{E}(c_1|c_2, c_3, \ldots, c_N)$ is a linear function of $\mathbf{y}$ the formula $\mathrm{E}(\mathbf{y}_1|\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N) = \mathbf{L}_1\mathrm{E}(c_1|c_2, c_3, \ldots, c_N)$ tells us that the regression matrix is a column times a row vector. Therefore its $p_1 \cdot (\ p - p_1)$ elements can be expressed in terms of just $(p - 1)$ parameters (one row of $(p - p_1)$ elements plus $(p_1 - 1)$ proportionality factors). This number could be even smaller when the model imposes structural constraints on $\mathbf{R}_c$ as well. A partial check could be performed using any of the methods developed for restricted rank testing. A possible objection could be that the tests are likely to be sensitive to deviations from the Gaussian distribution, but jackknifing or bootstrapping might help to alleviate this. Another issue is the fact that we get many tests that are also correlated, so that simultaneous testing techniques based on Bonferroni or more modern approaches are required.[28]

4.4.2 Exploiting the Difference Between Different Estimators

We noted that a number of generalized canonical variable programs yield identical results when applied to a $\boldsymbol{\Sigma}$ satisfying the composites factor model. But we expect to get different results when this is not the case. So, when using the estimate for $\boldsymbol{\Sigma}$ one might want to check whether the differences between, say PLS mode B and MAXVAR (or any other couple of methods), are too big for comfort. The scale on which to measure this could be based on the probability (as estimated by the bootstrap) of obtaining a larger "difference" than actually observed.

4.4.3 Prediction Tests, via Cross-Validation

The path diagram might naturally indicate composites and indicators that are most relevant for prediction. So it would seem to make sense to test whether the model's rank restrictions can help improve predictions of certain selected composites or indicators. The result will not only reflect model adequacy but also the statistical phenomenon that the imposition of structure, even when strictly unwarranted, can help in prediction. It would therefore also reflect the sample size. The reference for an elaborate and fundamental discussion of prediction and cross-validation in a PLS-context is Shmueli et al. (2016).

[28] See, e.g., chapter 34 from DasGupta (2008).

4.4.4 Global Goodness-of-Fit Tests

In SEM we test the model by assessing the probability value of a distance measure between the sample covariance matrix $\mathbf{S}$ and an estimated matrix $\widehat{\boldsymbol{\Sigma}}$ that satisfies the model. Popular measures are

$$\frac{1}{2}\text{tr}\left(\mathbf{S}^{-1}\left(\mathbf{S}-\widehat{\boldsymbol{\Sigma}}\right)\right)^2 \tag{4.53}$$

and

$$\text{tr}\left(\mathbf{S}\widehat{\boldsymbol{\Sigma}}^{-1}\right)-\log\left(\det\left(\mathbf{S}\widehat{\boldsymbol{\Sigma}}^{-1}\right)\right)-p \tag{4.54}$$

They belong to a large class of distances, all expressible in terms of a suitable function f:

$$\sum_{k=1}^{p} f\left(\gamma_k\left(\mathbf{S}^{-1}\widehat{\Sigma}\right)\right). \tag{4.55}$$

Here $\gamma_k(\cdot)$ is the kth eigenvalue of its argument, and f is essentially a smooth real function defined on positive real numbers, with a unique global minimum of zero at the argument value 1.The functions are "normalized," $f''(1)=1$, entailing that the second-order Taylor expansions around 1 are identical.[29] For the examples referred to we have $f(\gamma)=\frac{1}{2}(1-\gamma)^2$ and $f(\gamma)=1/\gamma+\log(\gamma)-1$, respectively. Another example is $f(\gamma)=\frac{1}{2}(\log(\gamma))^2$, the so-called geodesic distance; its value is the same whether we work with $\mathbf{S}^{-1}\widehat{\boldsymbol{\Sigma}}$ or with $\mathbf{S}\widehat{\boldsymbol{\Sigma}}^{-1}$. The idea is that when the model fits perfectly, so $\mathbf{S}^{-1}\widehat{\boldsymbol{\Sigma}}$ is the identity matrix, then all its eigenvalues equal one, and conversely. This class of distances was first analyzed by Swain (1975).[30] Distance measures outside of this class are those induced by WLS with general fourth-order moments based weight matrices,[31] but also the simple ULS: $\text{tr}\left(\mathbf{S}-\widehat{\boldsymbol{\Sigma}}\right)^2$. We can take any of these measures, calculate its value, and use the bootstrap to estimate the corresponding probability value. It is important to pre-multiply the observation vectors by $\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\mathbf{S}^{-\frac{1}{2}}$ before the bootstrap is implemented, in order to ensure that their empirical distribution has a covariance matrix that agrees with the assumed model.

[29] The estimators based on minimization of these distances are asymptotically equivalent. The value of the third derivative of f appears to affect the bias: high values tend to be associated with small residual variances. So the first example, "GLS," with $f'''(1)=0$, will tend to underestimate these variances more than the second example, "LISREL," with $f'''(1)=-4$. See Swain (1975).

[30] Swain (1975). See also Dijkstra (1990).

[31] The manual of EQS, Bentler (2006) is a treasure trove with information on goodness-of-fit testing with WLS, and Structural Equations Modeling generally. For related discussions, see Bentler and Dijkstra (1985) and Wansbeek and Meijer (2000).

For $\widehat{\boldsymbol{\Sigma}}$ one could take in an obvious notation $\widehat{\boldsymbol{\Sigma}}_{ii} := \mathbf{S}_{ii}$ and for $i \neq j$

$$\widehat{\boldsymbol{\Sigma}}_{ij} := \widehat{r}_{ij} \cdot \mathbf{S}_{ii}\widehat{\mathbf{w}}_i \cdot \widehat{\mathbf{w}}_j^{\mathsf{T}}\mathbf{S}_{jj}. \tag{4.56}$$

Here $\widehat{r}_{ij} = \widehat{\mathbf{w}}_i^{\mathsf{T}}\mathbf{S}_{ij}\widehat{\mathbf{w}}_j$ if there are no constraints on $\mathbf{R}_c$, otherwise it will be the ijth element of $\widehat{\mathbf{R}}_c$. If $\mathbf{S}$ is p.d., then $\widehat{\boldsymbol{\Sigma}}$ is p.d. (as follows from the appendix) and $\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\mathbf{S}^{-\frac{1}{2}}$ is well-defined.

4.5 Some Final Observations and Comments

In this chapter we outlined a model in terms of observables only while adhering to *the soft modeling principle* of Wold's PLS. Wold developed his methods against the backdrop of a particular *latent variables model*, the basic design. This introduces N additional *un*observable variables which by necessity cannot in general be expressed unequivocally in terms of the "manifest variables," the indicators. However, we can construct composites that satisfy the same structural equations as the latent variables, in an infinite number of ways in fact. Also, we can design composites such that the regression of the indicators on the composites yields the loadings. But in the regular case *we cannot have both*.

Suppose $\mathbf{y} = \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon}$ with $\mathrm{E}\mathbf{f}\boldsymbol{\varepsilon}^{\mathsf{T}} = 0$, $\boldsymbol{\Theta} :=\mathrm{cov}(\boldsymbol{\varepsilon}) > 0$, and $\boldsymbol{\Lambda}$ has full column rank. The p.d. $\mathrm{cov}(\mathbf{f})$ will satisfy the constraints as implied by identifiable equations like $\mathbf{B}\mathbf{f}_{\mathrm{endo}} = \mathbf{C}\mathbf{f}_{\mathrm{exo}} + \boldsymbol{\zeta}$ with $\mathrm{E}\mathbf{f}_{\mathrm{exo}}\boldsymbol{\zeta}^{\mathsf{T}} = 0$. All variables have zero mean. Let $\widehat{\mathbf{f}}$, of the same dimension as $\mathbf{f}$, equal $\mathbf{F}\mathbf{y}$ for a fixed matrix $\mathbf{F}$. If the regression of $\mathbf{y}$ on $\widehat{\mathbf{f}}$ yields $\boldsymbol{\Lambda}$ we must have $\mathbf{F}\boldsymbol{\Lambda} = \mathbf{I}$ because then

$$\boldsymbol{\Lambda} = \mathrm{E}\left[\mathbf{y}\,(\mathbf{F}\mathbf{y})^{\mathsf{T}}\right] \cdot [\mathrm{cov}\,(\mathbf{F}\mathbf{y})]^{-1} = \mathrm{cov}\,(\mathbf{y})\,\mathbf{F}^{\mathsf{T}}[\mathbf{F}\mathrm{cov}\,(\mathbf{y})\,\mathbf{F}^{\mathsf{T}}]^{-1} \tag{4.57}$$

Consequently

$$\widehat{\mathbf{f}} = \mathbf{F}\,(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\varepsilon}) = \mathbf{f} + \mathbf{F}\boldsymbol{\varepsilon} \tag{4.58}$$

and $\widehat{\mathbf{f}}$ has a larger covariance matrix then $\mathbf{f}$ (the difference is p.s.d., usually p.d.). One example is[32] $\mathbf{F} = \left(\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\Theta}^{-1}$ with $\mathrm{cov}\left(\widehat{\mathbf{f}}\right) - \mathrm{cov}(\mathbf{f}) = \left(\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}\right)^{-1}$.

So, generally, if the regression of $\mathbf{y}$ on the composites yields $\boldsymbol{\Lambda}$, the covariance matrices cannot be the same, and the composites cannot satisfy the same equations

[32] One can verify directly that the regression yields $\boldsymbol{\Lambda}$. Also note that here $\mathbf{F}\boldsymbol{\Lambda} = \mathbf{I}$.

as the latent variables $\mathbf{f}$.[33] Conversely, if $\text{cov}\left(\widehat{\mathbf{f}}\right) = \text{cov}(\mathbf{f})$, then the regression of $\mathbf{y}$ on the composites cannot yield $\mathbf{\Lambda}$.

If we minimize $\text{E}(\mathbf{y}-\mathbf{\Lambda}\mathbf{F}\mathbf{y})^{\intercal}\,\mathbf{\Theta}^{-1}\,(\mathbf{y}-\mathbf{\Lambda}\mathbf{F}\mathbf{y})$ subject to $\text{cov}(\mathbf{F}\mathbf{y}) = \text{cov}(\mathbf{f})$ we get the composites that LISREL reports. We can generate an infinite number of alternatives[34] by minimizing $\text{E}(\mathbf{f}-\mathbf{F}\mathbf{y})^{\intercal}\,\mathbf{V}\,(\mathbf{f}-\mathbf{F}\mathbf{y})$ subject to $\text{cov}(\mathbf{F}\mathbf{y}) = \text{cov}(\mathbf{f})$ for any conformable p.d. $\mathbf{V}$. Note that each composite here typically uses *all* indicators. Wold takes composites that combine the indicators per block. Of course, they also cannot reproduce the measurement equations and the structural equations, but the parameters can be obtained (consistently estimated) using suitable corrections (PLSc.[35])

Two challenging research topics present themselves: *first*, the extension of the approach to more dimensions/layers, and *second*, the imposition of sign constraints on weights, loadings, and structural coefficients, while maintaining as far as possible the numerical efficiency of the approach.

Appendix

Here we will prove that $\mathbf{\Sigma}$ is positive definite when and only when the correlation matrix of the composites, $\mathbf{R}_c$, is positive definite. The "only when"-part is trivial. The proof that $\{\mathbf{R}_c$ is p.d.$\}$ implies $\{\mathbf{\Sigma}$ is p.d.$\}$ is a bit more involved. It is helpful to note *for that purpose* that we may assume that each $\mathbf{\Sigma}_{ii}$ is a unit matrix (pre-multiply and post-multiply by a block-diagonal matrix with $\mathbf{\Sigma}_{ii}^{-\frac{1}{2}}$ on the diagonal, and redefine $\mathbf{w}_i$ such that $\mathbf{w}_i^{\intercal}\mathbf{w}_i = 1$ for each i). So if we want to know whether the eigenvalues of $\mathbf{\Sigma}$ are positive it suffices to study the eigenvalue problem $\widetilde{\mathbf{\Sigma}}\mathbf{x} = \gamma\mathbf{x}$:

$$\begin{bmatrix} \mathbf{I}_{p_1} & r_{12}\mathbf{w}_1\mathbf{w}_2^{\intercal} & r_{13}\mathbf{w}_1\mathbf{w}_3^{\intercal} & \cdot & r_{1N}\mathbf{w}_1\mathbf{w}_N^{\intercal} \\ & \mathbf{I}_{p_2} & r_{23}\mathbf{w}_2\mathbf{w}_3^{\intercal} & \cdot & r_{2N}\mathbf{w}_2\mathbf{w}_N^{\intercal} \\ & & \cdot & \cdot & \cdot \\ & & & \mathbf{I}_{p_{N-1}} & r_{N-1,N}\mathbf{w}_{N-1}\mathbf{w}_N^{\intercal} \\ & & & & \mathbf{I}_{p_N} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdot \\ \mathbf{x}_{N-1} \\ \mathbf{x}_N \end{bmatrix} = \gamma \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdot \\ \mathbf{x}_{N-1} \\ \mathbf{x}_N \end{bmatrix} \tag{4.59}$$

[33] One may wonder about the "best linear predictor" of $\mathbf{f}$ in terms of $\mathbf{y}$: $\text{E}(\mathbf{f}|\,\mathbf{y})$. Since $\mathbf{f}$ equals $\text{E}(\mathbf{f}|\,\mathbf{y})$ plus an uncorrelated error vector, $\text{cov}(E\,(\mathbf{f}|\,\mathbf{y}))$ is not "larger" but "smaller" than $\text{cov}(\mathbf{f})$. So $\text{E}(\mathbf{f}|\,\mathbf{y})$ satisfies neither of the two desiderata.

[34] Dijkstra (2015).

[35] PLSc exploits the lack of correlation between some of the measurement errors *within* blocks. It is sometimes equated to a *particular* implementation (e.g., assuming all errors are uncorrelated, and a specific correction), but that is selling it short. See Dijkstra (2011, 2013a,b) and Dijkstra and Schermelleh-Engel (2014).

with obvious implied definitions. Observe that every nonzero solution of

$$\begin{bmatrix} \mathbf{w}_1^{\mathsf{T}} & \mathbf{0} & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2^{\mathsf{T}} & \mathbf{0} & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \mathbf{0} & \mathbf{w}_{N-1}^{\mathsf{T}} & \mathbf{0} \\ \mathbf{0} & \cdot & \cdot & \mathbf{0} & \mathbf{w}_N^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdot \\ \mathbf{x}_{N-1} \\ \mathbf{x}_N \end{bmatrix} = \mathbf{0} \tag{4.60}$$

corresponds with $\gamma = 1$, and there are $\sum_{i=1}^{N} p_i - N$ linearly independent solutions. The multiplicity of the root $\gamma = 1$ is therefore $\sum_{i=1}^{N} p_i - N$ and we need to find N more roots. By assumption $\mathbf{R}_c$ has N positive roots. Let $\mathbf{u}$ be an eigenvector with eigenvalue μ, so $\mathbf{R}_c\mathbf{u} = \mu \cdot \mathbf{u}$. We have

$$\widetilde{\boldsymbol{\Sigma}} \begin{bmatrix} u_1\mathbf{w}_1 \\ u_2\mathbf{w}_2 \\ \cdot \\ u_N\mathbf{w}_N \end{bmatrix} = \begin{bmatrix} (u_1 + r_{12}u_2 + \cdot + r_{1N}u_N)\,\mathbf{w}_1 \\ (r_{21}u_1 + u_2 + \cdot + r_{2N}u_N)\,\mathbf{w}_2 \\ \cdot \\ (r_{N1}u_1 + r_{N2}u_2 + \cdot + u_N)\,\mathbf{w}_N \end{bmatrix} = \mu \begin{bmatrix} u_1\mathbf{w}_1 \\ u_2\mathbf{w}_2 \\ \cdot \\ u_N\mathbf{w}_N \end{bmatrix} \tag{4.61}$$

In other words, the remaining eigenvalues are those of $\mathbf{R}_c$, and so all eigenvalues of $\widetilde{\boldsymbol{\Sigma}}$ are positive. Therefore $\boldsymbol{\Sigma}$ is p.d., as claimed.

Note for the determinant of $\boldsymbol{\Sigma}$ that

$$\det(\boldsymbol{\Sigma}) = \det(\mathbf{R}_c) \times \det(\boldsymbol{\Sigma}_{11}) \times \det(\boldsymbol{\Sigma}_{22}) \times \det(\boldsymbol{\Sigma}_{33}) \times \ldots \times \det(\boldsymbol{\Sigma}_{NN}) \tag{4.62}$$

and so the Kullback–Leibler' divergence between the Gaussian density for block-independence and the Gaussian density for the composites model is $-\frac{1}{2}\log(\det(\mathbf{R}_c))$. It is well known that $0 \leq \det(\mathbf{R}_c) \leq 1$, with 0 in case of a perfect linear relationship between the composites, so Kullback–Leibler divergence is infinitely large, and 1 in case of zero correlations between all composites, with zero divergence.

References

Bekker, P. A., & Dijkstra, T. K. (1990). On the nature and number of the constraints on the reduced form as implied by the structural form. *Econometrica, 58*(2), 507–514

Bekker, P. A., Merckens, A., & Wansbeek, T. J. (1994). *Identification, equivalent models and computer algebra*. Boston: Academic.

Bentler, P. M., & Dijkstra, T. K. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis* (Chap 2, pp. 9–42). Amsterdam: North-Holland.

Bentler, P. M. (2006). EQS 6 structural equations program manual. Multivariate Software Inc.

Berk, R. A. (2008). *Statistical learning from a regression perspective.* New York: Springer.

Boardman, A., Hui, B., & Wold, H. (1981). The partial least-squares fix point method of estimating interdependent systems with latent variables. *Communications in Statistics-Theory and Methods, 10*(7), 613–639.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. New York: Springer,
Dijkstra, T. K. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics, 22*(1/2), 67–90 (Invited contribution to the special issue on the Interfaces between Econometrics and Psychometrics).
Dijkstra, T. K. (1981). *Latent variables in linear stochastic models* (PhD thesis, University of Groningen). Available on Research Gate.
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*(2), 327–336.
Dijkstra, T. K. (1995). Pyrrho's Lemma, or have it your way. *Metrika, 42*(1), 119–125.
Dijkstra, T. K. (2010). Latent variables and indices: Herman Wold's basic design and partial least squares. In V. E. Vinzi, W. W. Chin, J. Henseler & H. Wang (Eds.), *Handbook of partial least squares, concepts, methods and applications* (Chap, 1, pp. 23–46). Berlin: Springer.
Dijkstra, T. K. (2011). *Consistent partial least squares estimators for linear and polynomial factor models*. Technical Report. Research Gate. doi:10.13140/RG.2.1.3997.0405.
Dijkstra, T. K. (2013a). *A note on how to make PLS consistent*. Technical Report. Research Gate, doi:10.13140/RG.2.1.4547.5688.
Dijkstra, T. K. (2013b). *The simplest possible factor model estimator, and successful suggestions how to complicate it again.* Technical Report. Research Gate. doi:10.13140/RG.2.1.3605.6809.
Dijkstra, T. K. (2014). PLS' Janus face. *Long Range Planning, 47*(3), 146–153.
Dijkstra, T. K. (2015). *All-inclusive versus single block composites.* Technical Report. Research Gate. doi:10.13140/RG.2.1.2917.8082.
Dijkstra, T. K., & Henseler, J. (2015a). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics and Data Analysis, 81*, 10–23.
Dijkstra, T. K., & Henseler, J. (2015b). Consistent partial least squares path modeling. *MIS Quarterly, 39*(2), 297–316.
Dijkstra, T. K., & Schermelleh-Engel, K. (2014). Consistent partial least squares for nonlinear structural equation models. *Psychometrika, 79*(4), 585–604 [published online (2013)].
Dijkstra, T. K., & Veldkamp, J. H. (1988). Data-driven selection of regressors and the bootstrap. In T. K. Dijkstra (Ed.), *On model uncertainty and its statistical implications* (Chap. 2, pp. 17–38). Berlin: Springer.
Freedman, D. A. (2009). *Statistical models, theory and practice*. Cambridge: Cambridge University Press. Revised ed.
Freedman, D. A., Navidi, W., & Peters, S. C. (1988). On the impact of variable selection in fitting regression equations. In T. K. Dijkstra (Ed.), *On model uncertainty and its statistical implications* (Chap. 1, pp. 1–16). Berlin: Springer.
Haavelmo, T. (1944). The probability approach in econometrics. PhD-thesis *Econometrica 12*(Suppl.), 118pp. http://cowles.econ.yale.edu/
Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics, 34*(5), 2554–2591.
Ruud, P. A. (2000). *Classical econometric theory*. New York: Oxford University Press.
Shmueli, G., Ray, S., Velasquez Estrada, J. M., & Chatla, S. (2016). The elephant in the room: Predictive performance of PLS models. *Journal of Business Research, 69*, 4552–4564.
Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika, 58*(3), 433–451.
Pearl, J. (2009). *Causality—models, reasoning and inference*. Cambridge: Cambridge University Press.
Swain, A. J. (1975). A class of factor analysis estimation procedures with common asymptotic sampling properties. *Psychometrika, 40*, 315–335.
Wansbeek, T. J. & Meijer, E. (2000). *Measurement error and latent variables in econometrics*. Amsterdam: North-Holland.
Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. In F. N. David (Ed.), *Research papers in statistics. Festschrift for J. Neyman* (pp. 411–444). New York: Wiley.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock et al. (Eds.), *Quantitative sociology* (Chap. 11, pp. 307–358). New York: Academic.

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation, Part II* (Chap. 1, pp. 1–54). Amsterdam: North-Holland.