A note on how to make Partial Least Squares (mode A) consistent[1]

0. Introduction

Our setup will be the 'basic design', a factor model that satisfies the 'fundamental principle of soft modeling': all information between the blocks is conveyed by the latent variables. This entails that the covariance matrix between the indicators of any two blocks has rank one. We assume the availability of a consistent estimator[2] for the covariance matrix of the indicators on which we apply mode A. It is well known that this mode produces weight vectors for the blocks with probability limits proportional to the loadings. If we let Σ represent the covariance matrix of *any* block, λ the corresponding loading vector and $\overline{w}$ the probability limit (*plim*) of the estimated weight vector $\widehat{w}$, we have

$$\overline{w} = \text{plim } \widehat{w} = \lambda \div (\lambda^{\intercal}\Sigma\lambda)^{\frac{1}{2}} . \tag{1}$$

Based on another property of the basic design, I have shown (Dijkstra 1981, 2010) how to estimate the proportionality factor for each weight vector consistently. With the proportionality factors we can adjust the output of mode A and obtain consistent estimators for all parameters of the factor model. The property referred to is the assumed zero correlation between the errors *within* each block. So

$$\Sigma = \lambda\lambda^{\intercal} + \Theta \tag{2}$$

where the diagonal matrix Θ stands for the covariance matrix of the errors. Since the off-diagonal elements of Σ are products of the loadings, $\sigma_{ij} = \lambda_i\lambda_j$, the idea was that the off-diagonal elements s_{ij} of the consistent estimator S for Σ can help adjust the scale of $\widehat{w}$ and thereby obtain a consistent estimator for λ. To this end we multiply $\widehat{w}$ by a scalar c and choose its value such that the difference between s_{ij} and $(c\widehat{w}_i) \cdot (c\widehat{w}_j)$ is 'as small as possible'. In (Dijkstra 1981, 2010) this was simply made operational by minimizing

$$\sum_{i \neq j} [s_{ij} - (c\widehat{w}_i) \cdot (c\widehat{w}_j)]^2 \tag{3}$$

[1] I assume familiarity with PLS, its modes and its notions. See *The Handbook of Partial Least Squares* by Esposito Vinzi et al (2010) e.g. or *Systems Under Indirect Observation*, part II, by (Jöreskog and Wold, 1982), for background and elaboration. It may be appropriate here to point out that similar approches have been outlined for mode B, (Dijkstra 1981, 2010). Our focus here is on mode A.

[2] An estimator S is said to be a consistent estimator for Σ when S is arbitrarily close to Σ, with arbitrarily high probability, provided the sample size is sufficiently large. This is abbreviated to plim $S = \Sigma$.

as a function of c. The solution $\widehat{c}$ is

$$\widehat{c} = \left[\frac{\sum_{i \neq j} \widehat{w}_i \widehat{w}_j s_{ij}}{\sum_{i \neq j} \widehat{w}_i^2 \widehat{w}_j^2} \right]^{\frac{1}{2}} \tag{4}$$

provided the expression in brackets is positive. A special case is obtained for just two indicators. Then[3]

$$\widehat{c} = \sqrt{\frac{s_{12}}{\widehat{w}_1 \widehat{w}_2}}. \tag{5}$$

Clearly, $(\widehat{c}\widehat{w}_1) \cdot (\widehat{c}\widehat{w}_2) = s_{12}$ so that plim $((\widehat{c}\widehat{w}_1) \cdot (\widehat{c}\widehat{w}_2))$ =plim $s_{12} = \lambda_1 \lambda_2$ and $\widehat{c}\widehat{w}$ is consistent for λ. We note for further reference that for the general case $\text{plim}(\widehat{c}) = (\lambda^\intercal \Sigma \lambda)^{\frac{1}{2}}$.

One purpose of this note is to slightly extend the least squares approach to the case where some of the errors are correlated, and by introducing weights. Another purpose is to propose more ways of defining suitable scaling factors like $\widehat{c}$, who may turn out to be useful as well.

1. Correlated errors and weighted least squares

Taking account of correlated errors is actually quite trivial, when we know which errors (which elements of ε) are correlated. Let U be the set of *un*correlated pairs: $U := \{(i,j) \,|\text{corr}\,(\varepsilon_i, \varepsilon_j) = 0\}$ assumed to be nonempty. An immediate extension is to minimize with respect to c:

$$\sum_{i,j \in U} \left[s_{ij} - c^2 \widehat{w}_i \widehat{w}_j \right]^2 \tag{6}$$

which produces

$$\widehat{c} = \left[\frac{\sum_{i,j \in U} \widehat{w}_i \widehat{w}_j s_{ij}}{\sum_{i,j \in U} \widehat{w}_i^2 \widehat{w}_j^2} \right]^{\frac{1}{2}}. \tag{7}$$

The least squares criterion implicitly regards all differences $s_{ij} - c^2 \widehat{w}_i \widehat{w}_j$ equally informative. But one could claim that this is too simple. Arguing heuristically, using facts from the asymptotic theory of sample moments from multinormal distributions, we know that the higher the correlation σ_{ij}, the more accurate s_{ij} will be (we will work with standardized variables as

[3] Typically, 'genuine' indicators can be assumed to be positively related to the underlying latent variables, so all three terms under the square root sign will be positive with high probability when the sample is sufficiently large. But loadings with opposite signs present no problem either, since numerator and denominator will tend to be of the same sign.

is customary in PLS). In fact, the asymptotic variance of s_{ij} is $\left(1-s_{ij}^2\right)^2$ divided by the number of observations. Perhaps one should weigh the difference $s_{ij} - c^2\widehat{w}_i\widehat{w}_j$ by the inverse of $\left(1-s_{ij}^2\right)$ in the criterion. If we keep treading along this path, (asymptotic) covariances between s_{ij} and s_{kl} come into the picture as well, leading us to proper weighted least squares. And while we are at it, distribution-free estimates for these covariances are available and could be used instead of the values based on an assumed normality. But it is fair to say that the latter two suggestions will be of some value only when the sample size is 'very large', if experience with these weights in WLS in SEM counts for anything. The first suggestion, where cross-products are ignored, may have some merit though. The adjustment is again straightforward. Let W_{ij} be the weight of the difference $s_{ij} - c^2\widehat{w}_i\widehat{w}_j$, say $1 \div \left(1-s_{ij}^2\right)$, or something else. Then we minimize with respect to c

$$\sum_{i,j\in U}\left[\left(s_{ij} - c^2\widehat{w}_i\widehat{w}_j\right)\cdot W_{ij}\right]^2 \tag{8}$$

and get

$$\widehat{c} = \left[\frac{\sum_{i,j\in U}\widehat{w}_i\widehat{w}_j s_{ij} W_{ij}}{\sum_{i,j\in U}\widehat{w}_i^2\widehat{w}_j^2 W_{ij}}\right]^{\frac{1}{2}}. \tag{9}$$

Finally, as a bridge to the next section, we note that the variance stabilizing transformation of a correlation coefficient (from a normal sample) could also help to 'equalize the importances' of the differences in the criterion: the asymptotic variance of

$$\frac{1}{2}\log\left(\frac{1+s_{ij}}{1-s_{ij}}\right) \tag{10}$$

is just one over the number of observations, independent of σ_{ij}. So we could contemplate to minimize

$$\sum_{i,j\in U}\left[\frac{1}{2}\log\left(\frac{1+s_{ij}}{1-s_{ij}}\right) - \frac{1}{2}\log\left(\frac{1+c^2\widehat{w}_i\widehat{w}_j}{1-c^2\widehat{w}_i\widehat{w}_j}\right)\right]^2 \tag{11}$$

which however will require an iterative procedure for its solution. In the next section we will introduce a family of nonlinear fitting functions, some of which will also allow of an explicit, easy solution.

2. A large class of fitting functions suitable for ratios

Instead of looking at differences, we could take the ratios, assumed to be positive, as our raw material:

$$\frac{s_{ij}}{c^2\widehat{w}_i\widehat{w}_j} \text{ or } \frac{c^2\widehat{w}_i\widehat{w}_j}{s_{ij}} \tag{12}$$

and minimize a criterion of these ratios whose probability limit attains a unique global minimum at the point where all ratios equal one. In other words, based on probability limits $\{\sigma_{ij}\}_{i\neq j}$ and $\overline{w}$, minimization would yield the correct value. As an example, consider a real function $f(x)$ defined for positive real x as

$$f(x) := x - \log(x) - 1. \tag{13}$$

It is clear that $f(x) \geq 0$, and f is zero when and only when $x = 1$. Now minimize with respect to c (or c^2) the criterion

$$\sum_{i\neq j} f\left(\frac{c^2\widehat{w}_i\widehat{w}_j}{s_{ij}}\right) = \sum_{i\neq j}\frac{c^2\widehat{w}_i\widehat{w}_j}{s_{ij}} - \log\left(\frac{c^2\widehat{w}_i\widehat{w}_j}{s_{ij}}\right) - 1. \tag{14}$$

The optimal solution $\widehat{c}$ is the square root of the *harmonic mean* of the $\frac{s_{ij}}{\widehat{w}_i\widehat{w}_j}$'s. Replacing sample values by probability limits, $\text{plim}(s_{ij}) = \lambda_i\lambda_j$ and $\text{plim}(\widehat{w}_i\widehat{w}_j) = \lambda_i\lambda_j \div \lambda^\intercal\Sigma\lambda$ gives $\text{plim}(\widehat{c}) = (\lambda^\intercal\Sigma\lambda)^{\frac{1}{2}}$, exactly as required.

Sustituting $1/x$ for x yields $1/x + \log(x) - 1$ which is also nonnegative and zero only for $x = 1$. Minimization of the corresponding criterion yields the square root of the *arithmetic mean* of the $\frac{s_{ij}}{\widehat{w}_i\widehat{w}_j}$'s. It also has the desired probability limit of course. If we take $g(x) := \frac{1}{2}f(x) + \frac{1}{2}f(1/x) = \frac{1}{2}(x + 1/x) - 1$ the induced optimal correction factor equals

$$\sqrt{\frac{\sum_{i\neq j}\frac{s_{ij}}{\widehat{w}_i\widehat{w}_j}}{\sum_{i\neq j}\frac{\widehat{w}_i\widehat{w}_j}{s_{ij}}}}. \tag{15}$$

This is the square root of the geometric mean of the previous factors (the square roots of the arithmetic mean and the harmonic mean of the $\frac{s_{ij}}{\widehat{w}_i\widehat{w}_j}$'s), a neat compromise between the other solutions. The function $g(x)$ satisfies by construction $g(x) = g(1/x)$, so it does not matter on which of the ratios in (12) one focusses. As another example with the same property, consider a real function $h(x)$ defined for positive real x as[4]

$$h(x) := \frac{1}{2}(\log(x))^2. \tag{16}$$

Again, $h(x) \geq 0$, and h is zero when and only when $x = 1$. The value of c that minimizes the corresponding criterion is simply [5]

$$\sqrt{\left(\prod_{i\neq j}\left(\frac{s_{ij}}{\widehat{w}_i\widehat{w}_j}\right)\right)^{1/\#}}, \tag{17}$$

[4]The factor $\frac{1}{2}$ is an innocent normalization, yielding $f^{(2)}(1) = 1$ for the second derivative, which we will impose on all functions of x below.

[5]The symbol '#' stands for the number of different pairs (i, j)).

the square root of the *geometric mean* of the $\frac{s_{ij}}{\widehat{w}_i \widehat{w}_j}$'s. Also with the correct probability limit.

Incidentally, recall that the geometric, harmonic and arithmetic means are members of the family of power means. So any square root of a power mean will be a contender. Since the power means can be ordered (e.g. the harmonic mean $\leq$ the geometric mean $\leq$ the arithmetic mean), it is clear that the choice is real. We have as yet no clear guide as to which is best.

The functions $\frac{1}{4}(1-x)^2 + \frac{1}{4}(1-1/x)^2$ and $\left[\frac{1}{2}(x^r + x^{-r}) - 1\right]/r^2$ for real $r \geq 1$ are obviously feasible choices as well. We can generate an infinite number of candidates by taking any smooth, strictly convex function h with $h(y) = h(-y)$ for all real y and a unique minimum of zero at $y = 0$, and then take $h(\log(x))$ for positive x. By construction, $h(\log(x)) = h(\log(1/x))$. This family of functions has been studied in another context, multicriteria decision analysis (the AHP-method (Dijkstra, 2011)). Arbitrary members do not allow of explicit minimizers of the induced criteria, they typically require iterative techniques. But their probability limits all equal $(\lambda^\intercal \Sigma \lambda)^{\frac{1}{2}}$, as can be verified with some work based on general minimum distance estimation theory.

As before we can take correlated errors into account, and introduce weights.

A final remark: which function, which approach works best is as yet an entirely open question. It may well be that the approach first tried (equation (4)), and its variations, that led to PLSc, is a dependable workhorse and is good enough for most purposes, but the others may have merit also. Any guesses/advice, anyone?

Theo K. Dijkstra, April 29^{th} 2013.

References

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

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

[3] Dijkstra, T. K. (2011). On the extraction of weights from pairwise comparison matrices. *Central European Journal of Operations Research,* DOI 10.1007/s10100-011-0212-9

[4] Esposito Vinzi, V., Chin, W. W., Henseler, J., Wang, H. (eds.), *Handbook of Partial Least Squares,* Springer-Verlag, Heidelberg.

[5] Jöreskog, K. G. & Wold, H. (eds.),(1982). *Systems under Indirect Observation*, part II, North-Holland, Amsterdam.

[6] Steele, J. Michael (2004). *The Cauchy-Schwarz Master Class,* Cambridge University Press (The Mathematical Association of America)