

Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects

Tormod Næs¹ | Rosaria Romano² | Oliver Tomic³ | Ingrid Måge¹ |
Age Smilde⁴ | Kristian H. Liland³

¹Raw Materials and Process, Nofima AS, Ås, Norway

²Department of Economics and Statistics, University of Napoli, Federico II, Naples, Italy

³Realtek, The Norwegian University of Life Sciences, Ås, Norway

⁴Biosystems Data Analysis, Faculty of Sciences, University of Amsterdam, Amsterdam, The Netherlands

Correspondence

Tormod Næs, Raw Materials and Process, Nofima AS, Osloveien 1, Box 210, N-1431 Ås, Norway.

Email: tormod.naes@nofima.no

Funding information

Research Council of Norway

Abstract

This paper is about the use of the multiblock regression method sequential and orthogonalized partial least squares (SO-PLS) for path modeling. The paper is a follow up of previously published papers on the same topic and presents a number of new results for the method. First of all, the paper discusses more thoroughly the aspect of how to incorporate blocks in the models and relates this to standard concepts in the area of graphical modeling. Second, the paper defines the concept of direct and indirect effects more precisely in terms of population parameters and shows how they are related to the additional effect in SO-PLS modeling. The paper illustrates the theory by simple graphs, simulations, and a real example from process monitoring.

KEYWORDS

common components, graphical modelling, path analysis, SEM, SO-PLS

1 | INTRODUCTION

Regression analysis has always been one of the most widely used statistical methodologies for analyzing the dependence of a response variable from one or more predictors. Despite its crucial role, the underlying model represents only a partial interpretation of reality. In fact, a regression model focuses on the relations between dependent and independent variables, worrying less about the mechanisms that may exist among the independent ones. This issue was addressed at the beginning of the 20th century by the biogeneticist Wright¹ who developed the basics of the so-called *path analysis*. Wright also invented a specific graphical representation called the *path diagram* to visualize the direct and indirect effects of one variable on another. Path analysis was later extended to behavioral sciences by various authors.^{2–4} The approach was also generalized to relations between multivariate blocks of data, giving rise to the so-called structural equation models (SEMs).^{5,6} The SEM combines the principles of path analysis with those of factor analysis.⁷ Specifically, the factor analysis model is used, for each block, to analyze the relationships between the observed variables and the respective latent variables, while the path analysis (structural approach) is used to analyze the relationships between the underlying latent variables in the different blocks. In SEM, the two aspects of the joint model are frequently referred to as the *measurement* or *outer model* and the *structural* or *inner model*.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. Journal of Chemometrics published by John Wiley & Sons Ltd

SEM has become very important in for instance consumer science and psychology. In these sciences, each individual is typically presented with a number of batteries of questions (so-called questionnaires) corresponding to the phenomena one is interested in. Then the theorized links between the data blocks are set up, and the relations are estimated. In for instance consumer science, the different data blocks may represent information about demography, attitudes, and habits and possibly also preferences for products.⁸ A major goal is to understand what influences preferences and in which way.

Many methods have been put forward for SEM. The most well-known traditions are the LISREL⁵ and the partial least squares (PLS) path modeling (PLS-PM) traditions,^{9,10} based on quite different philosophies. The LISREL method is based on estimating the parameters in a full covariance matrix by maximum likelihood or generalized least squares methodology. PLS on the other hand is based on an algorithmic approach, which focuses on maximizing covariances of latent variables. PLS-PM is a very flexible approach that allows for different number of blocks and different relations between the blocks and suggests different ways to calculate parameters in both inner and outer relations.

As a contrast to these well-established strategies, a more explorative alternative was suggested in Næs et al.¹¹ using the sequential and orthogonalized SO-PLS regression as a building block.¹² The main advantage of the method is that it handles blocks with more than one latent variable more easily than standard methods.¹³ The method is based on fitting, by the use of SO-PLS methodology, each endogenous block to all blocks that have a path, either directly or indirectly into it. Therefore, each estimation step of the method operates locally and does not depend on paths and blocks in other parts of the path diagram. A number of advantages are gained by this approach: First of all, it handles blocks with different dimensionality, both in terms of the number of variables involved and in terms of underlying dimensionality (rank). Second, it is invariant to the relative scaling of the blocks. It should be noted that the SO-PLS approach is different from standard approaches in the sense that it does not establish one latent variable for each block; instead, it provides one set of latent variables for each of the input blocks for each of the models estimated.

Measures of direct and indirect effects are defined using explained variance as a criterion.¹³ These effects can be plotted in a similar way as the direct effects and path coefficients are plotted for the classical approaches PLS-PM and LISREL. The modes A and B that play a central, but also quite complex, role for PLS path modeling¹⁰ are handled directly/implicitly due to the flexibility in the number of components selected for each of the blocks.

This paper is about properties of the SO-PLS method for path modeling that have not been published before. In particular, we will focus on linking the proposed order of the blocks¹³ to graphical modeling.¹⁴ We will show that the order of blocks proposed corresponds to a topological order of the blocks. Next, we will define direct, indirect, and total effects in terms of population parameters and also show how they relate to each other and to the concept of additional effect.¹³ Relations to common and distinct components as defined in Smilde et al.¹⁵ will also be underlined. The methodology will be illustrated by graphical and numerical illustrations as well as calculations on real and simulated data. In one of the simulations, SO-PLS will also be compared with PLS-PM.⁹

The main novelties of the paper are improved definitions of direct and indirect effects, a more precise and unified definition of the order of blocks in SO-PLS, new viewpoints on the relations to established methodology in path modeling and multiblock analysis, and a novel application in process modeling.

2 | GRAPHICAL MODELING

2.1 | Directed acyclic graphs

A directed acyclic graph (DAG^{14,16}) is defined as a graphical model/diagram with directed paths and no cyclic structure. The acyclic requirement means that it is not possible to come back to a block by following a series of paths. Three examples of DAGs are given in Figure 1A-C, two based on real data (Figure 1B,C) from the literature and one constructed for the purpose of illustration (Figure 1A).

In the following, all blocks will be represented by a letter, A, B, C, D, etc. Later on, we will also use symbols \mathbf{X}_A , \mathbf{X}_B , and so forth, representing the data matrices with variables as columns. A direct relation between two blocks will be named an edge or an arrow (since arrows are used in the graphical display to denote direct relations). The concept of a path will be used more generally and also comprise a series of edges that lead from one block to another.

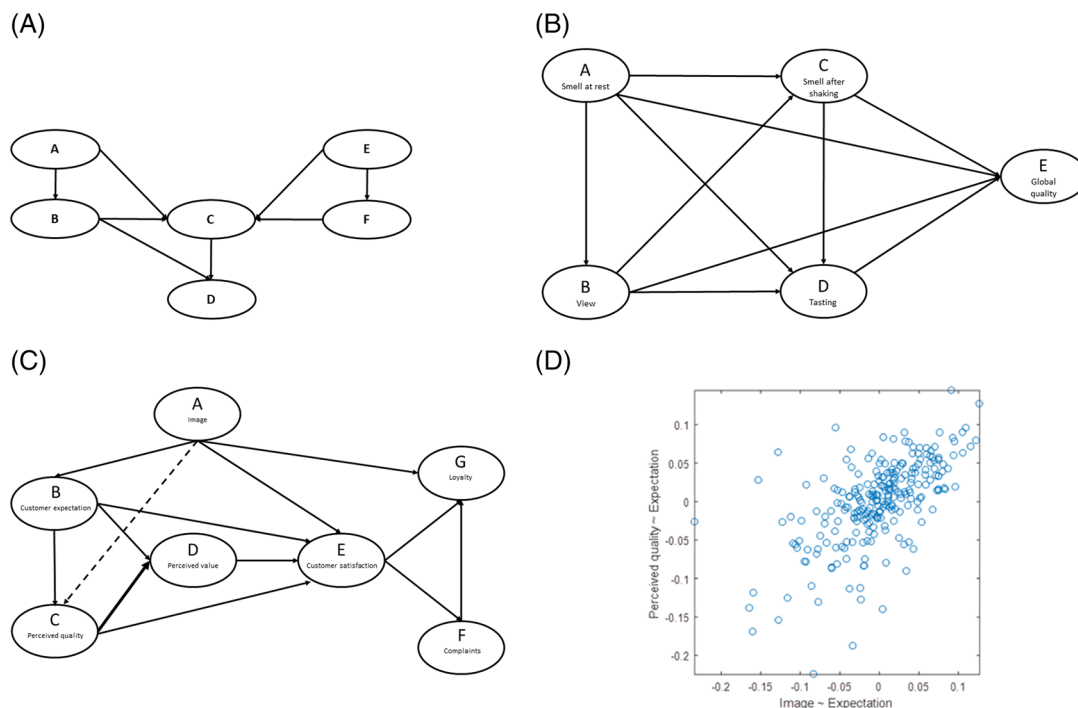


FIGURE 1 Three illustrating examples. (A) A path diagram to be used as an example for describing nonunique topologies, transitive closure, and transitive reduction. Two topological orders possible here are **A B E F C D** and **E F A B C D**. The boldface letters are inserted to indicate a point of missing reachability. The A and B in the first order are not reachable from E and F, and E and F are not reachable from A and B in the second order. C and D are reachable from all 4, that is, from A, B, E, and F. For both orders, one can calculate the direct, indirect, and total effects of the pairs (A,B), (A,C), (A,D), (B,C), (B,D), (E,F), (E,C), (E,D), (F,C), (F,D), and (C,D). For instance, (A,E) and (A,F) are not present in this group of relations since they lack a reachable relation. (B) The example is based on tasting of French wines and is the same as presented in for instance Romano et al.¹³ The different blocks represent different stages in the sensory evaluation of wines. In this case, the topological order is unique; A, B, C, D, E. (C) The blocks represent different aspect of consumer attitudes related to their loyalty for a mobile telephone provider and is the same as the one used in Romano et al.¹³ The dashed line shows the new path identified by the partial correlation analysis shown in Figure 1D. (D) Partial correlation plot for the blocks A and C conditioned on block B in Figure 1C

2.2 | Topological order

A reachable relationship between two blocks A and B is a relation in which one can reach one of the blocks (B) from the other (A) by passing through a number of edges.

Blocks in a given path diagram can always be presented by letters on a line from left to right. If we add the arrows between them and if all point from left to right, the order of the letters corresponds to a so-called topological order of the DAG. For DAGs, there is always a topological order,¹⁴ but it is not necessarily unique.

In Figure 1A, there are a number of different topological orders for the same path diagram. Two examples are (A **B E F C D**) and (E **F A B C D**); if arrows are added between the capital letters in these two orders according to the arrows in Figure 1A, one can see that all of them point to the right. Note that there may be blocks with no arrow between them, for instance, between F and A in the last of the two topologies. The boldface letters are inserted to indicate such a point. An example of an order of letters, which is not corresponding to a topological order, is (A B F E D C).

In the wine example in Figure 1B and the mobile example in Figure 1C, the topological orders, (A B C D E) and (A B C D E F), respectively, are unique. In both cases, it is impossible to find any other ordering of the blocks (represented by letters) without breaking the “arrow-to-the-right” property.

2.3 | Transitive closure of a DAG

There is another aspect of a DAG that will play a central role when discussing the SO-PLS method below, namely, transitive closure. Transitive closure refers to the saturated path diagram, that is, the diagram with the most edges/arrows

representing the same reachability relation as the original. In other words, the transitive closure of a DAG is obtained by adding all possible arrows without changing the reachability structure, that is, without adding new reachable relations.

Transitive reduction on the other hand is the opposite, namely, the leanest version attainable without changing the reachability structure between any two of the blocks.

Usually, a true path diagram will be in between the two extremes, transitive closure and transitive reduction, which is the case for instance in the mobile example in Romano et al.¹³ (see also Figure 1C). As can be seen, there are lots of direct effects, which are not added. In the wine example (see Figure 1B) in Romano et al.¹³ on the other hand, the true path diagram is the same as the transitive closure; all edges possible are in the diagram. In other words, the original diagram is identical to its transitive closure.

We refer to Figure 2A,B for an illustration of the transitive closure and transitive reduction of the diagram in Figure 1A. The dotted lines in Figure 2A are added to obtain the transitive closure of the original path diagram.

The transitive closure and transitive reduction are important concepts in path modeling because of possible pruning of the diagram and for checking whether important edges have been left out¹³ (see also Section 6.2).

Note in particular that in the case of more than one exogenous block, that is, blocks with no input path, a relation between these blocks cannot be added without changing the reachability structure.

3 | SO-PLS FOR PATH MODELING

3.1 | SO-PLS as a multiblock regression method

We will start by describing the basic SO-PLS regression method, which is the basis for the path modeling approach to be discussed here. In this section, we let \mathbf{Y} denote the output matrix and $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K$ the data blocks to be used for input. All blocks are centered throughout the paper. No scaling will be done; see below for a discussion of invariance with respect to between block scaling.

The SO-PLS is based on the general linear multiblock regression model

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\Theta}_1 + \mathbf{X}_2\boldsymbol{\Theta}_2 + \dots + \mathbf{X}_K\boldsymbol{\Theta}_K + \mathbf{E}, \quad (1)$$

where \mathbf{Y} can be multivariate and the $\boldsymbol{\Theta}_k$ represent the regression coefficients. Note that SO-PLS is only developed for one \mathbf{Y} -block. The SO-PLS fitting of this model is based on a sequential strategy iterating between PLS regression and

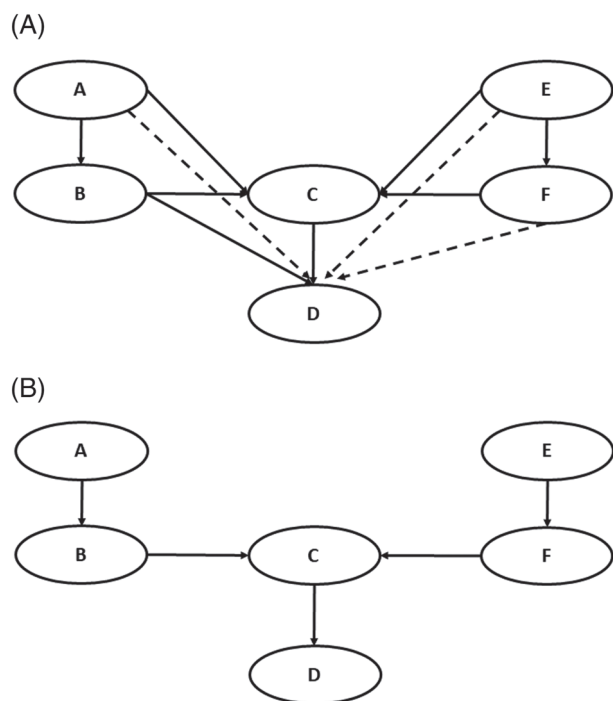


FIGURE 2 Transitive reduction and closure of the example in Figure 1A. A, The transitive closure, that is, the saturated diagram for the path diagram in Figure 1A. No more paths can be added without changing the reachability structure. Note that the set of direct/indirect effects corresponds to all direct paths in the transitive closure. The dotted lines are those needed to establish the transitive closure. B, The transitive reduction of the diagram in Figure 1A. This is the leanest possible path diagram with the same reachability structure as the original

orthogonalization with respect to previously modeled blocks. The first step is to fit \mathbf{Y} to \mathbf{X}_1 by PLS regression, then \mathbf{X}_2 and \mathbf{Y} are orthogonalized with respect to \mathbf{X}_1 , that is, with respect to the PLS scores of \mathbf{X}_1 , and then the deflated \mathbf{Y} is fitted to the deflated/orthogonalized \mathbf{X}_2 . The process continues until all blocks have been fitted. The order of the blocks is sometimes obvious, while in other cases, it must be chosen. Strategies for incorporating blocks have also been developed.^{17,18} The SO-PLS can also be used for design matrices \mathbf{X} ,¹⁹ but this is seldom relevant for path modeling. The orthogonalization is in the framework of path modeling needed for calculating the additional effect of a block, as will be discussed below.

The number of components is usually determined by cross-validation (CV) using the Måge plot,¹² often referred to as *global* model optimization. With this approach, all combinations of numbers of components from all blocks (up to a set limit) are tested, and the root-mean-squared errors of prediction are plotted on the vertical axis in a type of scree plot with the sum of the components across X-blocks on the horizontal axis. The *sequential* approach is another and more conservative alternative for model selection, where one determines the number of components sequentially as new blocks are incorporated. Although the global approach can be computationally intensive, it is often preferred over the sequential approach because it is less sensitive to the order of the blocks and reveals (through visual inspection of the Måge plot) if there are alternative models with approximately the same prediction ability. For the path modeling approach, however, we will use the sequential procedure since the order of the block is important for interpretation. The CV-ANOVA method²⁰ can be useful for assessing the significance of new blocks added. This is a method based on using the paired T test based on cross-validated residuals.

Interpretation is done using scores and plots from the individual PLS models, the additional explained variances of the blocks as they are incorporated and by using the principal components of prediction (PCP²¹) method. The PCP is a method that first applies a PCA on the predicted Y-values and afterwards relates the PCA scores directly to the input variables independently by regression. Cases with only one y-variable need a separate treatment,²¹ but that is not relevant in this study given that Y is multivariate in all examples. PCP provides PCA scores and loading plots plus plots of how the input variables relate to predicted output. One can then simultaneously interpret how the different input variables are related to the predicted output variables and the samples involved. An example of the use of PCP in path modeling can be found in Romano et al.¹³

The SO-PLS is invariant, due to the orthogonalization, to between block scaling (but not within blocks scaling since PLS is used) and allows for different underlying dimensionalities in the blocks, which is a general advantage that also carries over to path modeling. In other words, if one block has many variables of one type (for instance, spectroscopic variables) and another one has only a few variables of another type (for instance process settings), one does not need any special pretreatment regarding the relative scaling of the two. For each block in the sequence, one can concentrate on the optimal number of components needed for that particular block. Since PLS is used for model fitting in each step, the method handles multicollinearity easily. Experience indicates that the order of the blocks is more important for interpretation than for prediction ability.

3.2 | SO-PLS used for path modeling

It must be emphasized that the SO-PLS method for path modeling is only applied for DAGs, that is, noncyclical graphs. Therefore, the method is here named SO-PLS-PM (SO-PLS path modeling).

The SO-PLS-PM^{11,13} consists of two steps:

Step 1

The first step is to fit an SO-PLS model to each endogenous block using all blocks that have a reachable relation to it as input. This means that each model only depends on blocks prior to it in the topological order. How to incorporate blocks sequentially will be discussed thoroughly in Section 4.

This first step gives explicit information about the predictive power at different places in the path diagram and also about the (additional) contributions to prediction of the different blocks as they are incorporated along the topological order (see below). This step is useful for identifying which blocks that actually contribute to the predictions and which blocks that can be meaningfully predicted in the path diagram. The models can be further interpreted using the PLS models for each block in the sequence and the PCP for more condensed interpretation as discussed above. The latter

has the advantage that it gives only one set of plots for each of the endogenous block models and therefore simplifies interpretation.

Step 2

The next step is to estimate the total, direct, and indirect effects of one block on another. These effects are calculated for all combinations of blocks that have a reachable relation, using explained variance as a criterion (see Section 5 for details). The different calculations are also here in practice done by the SO-PLS method since the calculations generally require a multiblock approach. The direct, indirect, and total effects can be visualized using simple graphs of the same style as used in standard PLS-PM for visualizing direct effects (or path coefficients). The precise definitions of the effects and how they are estimated will be given below. Validation is done using CV and the bootstrap for assessing the significance.

As an overview, the different steps are presented in Appendix A. We will now first discuss the first step of the modeling in more detail before turning to the second step in Section 5.

4 | ORDER OF BLOCKS IN SO-PLS AND RELATIONS TO GRAPHICAL MODELING (FIRST MODELING STEP)

4.1 | Order of blocks

For SO-PLS, an order of the blocks in each local model, that is, for each endogenous block, is needed. In Næs et al.¹¹ it was proposed to incorporate blocks according to how far they are from the output¹³; the blocks with the longest way to the output, that is, most paths between input and output, are modelled first, and then the closer blocks are added successively one by one. In this way, one obtains information about how much each of the blocks contributes to prediction ability in addition to the previous blocks further away from the output. In cases where two or more blocks have the same distance to the output (see for instance Figure 1A), the situation is more complicated as will be discussed below.

If the blocks in Figure 1B,C are incorporated according to this distance rule, the blocks are in both cases incorporated in alphabetical order. For instance, for the model for block E in Figure 1B, the blocks are incorporated in the order A, B, C, D.

If block names of all blocks that are input to an endogenous block are placed beside each other on a line in the order they are incorporated in SO-PLS as suggested above, no arrow between the blocks points to the left. This corresponds exactly to the definition of topological order discussed above. In other words, the process of incorporating blocks in a SO-PLS model corresponds exactly to a topological order of the DAG. This observation can then be turned into a definition; the order of blocks in SO-PLS is defined to follow a topological order of the blocks.

The relation between SO-PLS order and topological order holds both for nonunique and unique order of blocks in SO-PLS, but for the unique case, the situation is simpler since there is only one possibility. For the examples in Figure 1B,C, the topological order is unique, and only one option exists. The nonunique case is more difficult and will be discussed below. Note that in Næs et al.¹¹ this way of setting up the block names with arrows from the path diagram between them was called a dependence diagram. An example of this diagram for Figure 1B is given in Figure 3.

Note also that an imposed ordering based on the criterion above implies that the order of two different input blocks, for instance, A and B, will be the same in all possible models with A and B as input (given that the output is reachable from A and B). For instance, in the dependence diagram for Figure 1B (see Figure 3), the order of the input blocks A B

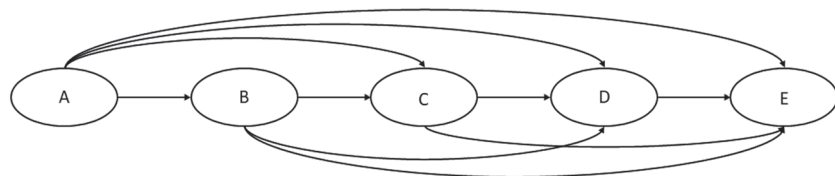


FIGURE 3 Dependence diagram for wine data in Figure 1B. The topological order is here unique and the transitive closure is the same as the original

C in the model for D will be the same as the order of A B C in the larger model for E. The order of the input blocks for all output/endogenous blocks in Figure 1C will also follow the alphabetical order.

4.2 | Nonunique topological order

In the case of a nonunique order, the incorporation of blocks as suggested in previous papers (see Section 4.1. above) leaves it to the user to select order for those places where a choice has to be made. This definition is somewhat diffuse, so from now on, we simply define the SO-PLS order of the blocks to follow one of the topological orders possible. This definition is precise, but the choice between the topological orders is still left to the user. There is, however, some advice regarding choice of topological order that can be given.

For instance, in the topology in Figure 1A, there are two “branches” that meet at the “stem” C, and there are a number of topological orders as discussed above and also illustrated in Figure 4A-C. For such cases, we recommend choosing an order that follows a branch until the stem before starting to model the other branch. This means that one models for instance both A and B before E and F for models with C and D as output even though E should come before B according to the general distance rule. This example corresponds to selecting the topological order A B E F C D in Figure 1A. Note that E F A B C D also shares this property. The choice between the two is up to the user and depends on which one he/she finds most natural to bring into the model before the other.

Another typical situation is depicted in Figure 5 where two branches (A→B→D and A→C→D) are totally separated but end up in the same block. The dashed line in the figure represents the extra arrow needed to establish the transitive closure. The choice of topological order is then up to the user and will depend on which path he/she finds most natural to incorporate as a basis and then investigate how much extra info that is obtained by the other blocks. The user is recommended to set up the dependence diagrams for all topological orders and choose the one that is most natural from a substantive matter point of view.

To sum up, for both unique and nonunique topological order, the input order of the blocks in SO-PLS-PM follows a topological order.

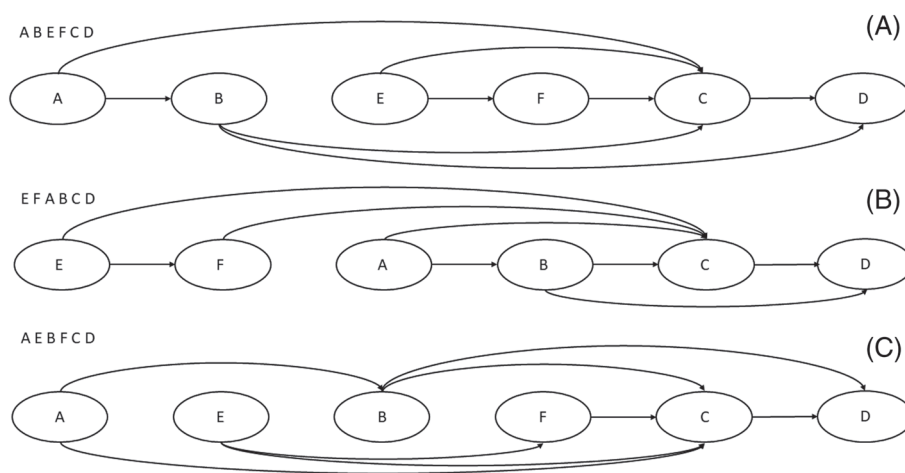


FIGURE 4 Dependence diagrams for the example in Figure 1A. Three possible topological orders A, B, and C, are presented for the diagram in Figure 1A

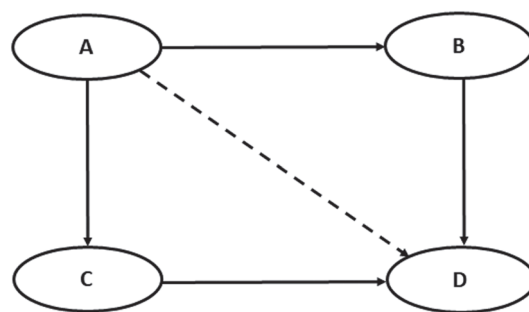


FIGURE 5 Parallel and merging paths. The dotted line represents the path that must be added to obtain the transitive closure of the diagram represented by the solid lines. The two possible topological orders are A B C D and A C B D. The reachable relations that are the basis for the calculations of the effects are A→B, A→C, A→D, B→D, and C→D. The C→B and B→C are not possible since they do not represent reachable relations. No transitive reduction is possible here

4.3 | Relations between SO-PLS modeling and transitive closure

Since for each SO-PLS model (i.e., each endogenous block), all possible blocks that have a reachable relation to the output are incorporated one by one, this means that SO-PLS estimation is not influenced by the transitive closure process, which amounts to adding all possible direct arrows in reachable relations. In other words, the SO-PLS-PM approach operates in the transitive closure of a DAG. An illustration of this is given by comparing one of the topological orders in Figures 1A and 2A (for instance, A B E F C D). They share the same topological order and dependence diagram, but the dependence diagram is now extended by adding all possible direct arrows in reachable relations. The SO-PLS modeling is the same in both cases; that is, the order of the blocks is A, B, E, F, and C for predicting D.

5 | DIRECT, INDIRECT, ADDITIONAL, AND TOTAL EFFECTS DEFINED IN TERMS OF EXPLAINED VARIANCE (FIRST AND SECOND MODELING STEPS)

This section is devoted to more detailed and precise definitions of the concepts of additional, direct, and indirect effects as described in steps 1 and 2 of the SO-PLS-PM. We will first use the simple path diagram in Figure 6 for illustrating the ideas.

5.1 | Background

In Romano et al.¹³ the additional, total, direct, and indirect effects are defined in terms of explained variance. This is different from standard path modeling methodology where these concepts are, although closely related to concepts of explained variance, defined in terms of regression coefficients.²² The explained variance solution was chosen for SO-PLS since the regression coefficient approach is not possible to use for multidimensional blocks, that is, for more than one variable in each of the blocks to be related to each other. The two definitions were compared empirically for one-dimensional blocks in Romano et al.¹³ The structure of the results was quite similar showing that they in the cases presented represent similar phenomena.

We will now first discuss the definitions of the different effects using the simple diagram in Figure 6 and only in terms of population parameters in noise-free data (deterministic model where the responses are 100% accounted for by the given input block). This will establish the concepts, and later on, we will describe how the definitions are extended to more complex diagrams and how the effects can be estimated in practice using CV.

We will describe the definitions in terms of variance contributions of the different model terms to the overall variability in \mathbf{X}_C , that is, the data matrix corresponding to the letter C in the diagram. In most cases in practice, however, one will use the relative value of the variance obtained by dividing by the overall variance $tr(\mathbf{X}_C^T \mathbf{X}_C)$:

$$\text{Relative Explained Variance} = \frac{\text{Variance estimate}}{tr(\mathbf{X}_C^T \mathbf{X}_C)}. \quad (2)$$

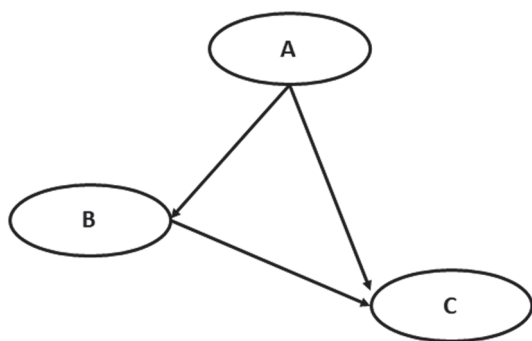


FIGURE 6 Motivating example for defining the direct, indirect, and total effects

An advantage of the relative explained variances (often called explained variance only) is that it gives values between 0% and 100%. When referring to effects in this section, we will refer to variances, not the relative values. In the empirical section of the paper, we will, however, divide the variances by the overall variability to obtain the explained variances as given by formula 2.

We will now first define total effect of \mathbf{X}_A on \mathbf{X}_C and then additional effect of \mathbf{X}_B on \mathbf{X}_C before we split the former into direct and indirect effect of \mathbf{X}_A .

5.2 | The model

The multiblock model for the diagram in Figure 6 for prediction of \mathbf{X}_C in noise-free data can be written as

$$\mathbf{X}_C = \mathbf{X}_A \mathbf{D} + \mathbf{X}_B \mathbf{F}, \quad (3)$$

where \mathbf{D} and \mathbf{F} are regression coefficients. We will assume that \mathbf{X}_A and \mathbf{X}_B have full column rank. In practice, however, all effects will be calculated by PLS regression, for which rank is not an issue.

For defining the different effects and for studying how they are related to each other, we will prefer the equivalent orthogonalized model (see Appendix B):

$$\mathbf{X}_C = \mathbf{X}_A \tilde{\mathbf{D}} + \mathbf{X}_B^{\text{orthA}} \mathbf{F}, \quad (4)$$

where

$$\mathbf{X}_B^{\text{orthA}} = \left(\mathbf{I} - \mathbf{X}_A (\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \right) \mathbf{X}_B. \quad (5)$$

Equation 4 corresponds exactly to the model used for the step 1 in the SO-PLS approach; that is, the \mathbf{X}_A is fitted first, and the second block \mathbf{X}_B is orthogonalized with respect to the first before fitting to \mathbf{X}_C .

Note that \mathbf{F} is the same in the two models, while \mathbf{D} is changed (see Appendix B for detail). The $\tilde{\mathbf{D}}$ can be written as

$$\tilde{\mathbf{D}} = \mathbf{D} + (\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \mathbf{X}_B \mathbf{F}. \quad (6)$$

Note that in the standard SO-PLS regression, one orthogonalizes with respect to the PLS components of block \mathbf{X}_A . In this situation with noise-free data considered here, the PLS needs the maximum number of components, and therefore, orthogonalization with respect to \mathbf{X}_A will be the same as orthogonalization with respect to the PLS scores of \mathbf{X}_A .

5.3 | Total and additional effects

The two contributions in Equation 4 are orthogonal, which means that they in practice can be fitted separately. Since they are orthogonal, the two variance contributions can therefore be calculated separately and added afterwards.

The variance corresponding to the first, \mathbf{X}_A , is called the total effect of block \mathbf{X}_A on block \mathbf{X}_C and is defined as the variance of $\mathbf{X}_A \tilde{\mathbf{D}}$, that is, as

$$\text{tot}_{\mathbf{X}_A} = \text{tr} \left(\tilde{\mathbf{D}}^T \mathbf{X}_A^T \mathbf{X}_A \tilde{\mathbf{D}} \right). \quad (7)$$

The variance contribution of \mathbf{X}_B on \mathbf{X}_C after orthogonalization with respect to \mathbf{X}_A is called the additional effect and is defined as

$$\text{additional}_{\mathbf{X}_B} = \text{tr}(\mathbf{F}^T (\mathbf{X}_B^{\text{OrthA}})^T \mathbf{X}_B^{\text{OrthA}} \mathbf{F}). \quad (8)$$

In other words, the additional effect is the effect of B that comes in addition to A in describing C. Note that the concept of additional effect is not present in standard path modeling approaches.

Because of orthogonality, we obtain

$$\text{tr}(\mathbf{X}_C^T \mathbf{X}_C) = \text{tr}(\tilde{\mathbf{D}}^T \mathbf{X}_A^T \mathbf{X}_A \tilde{\mathbf{D}}) + \text{tr}(\mathbf{F}^T (\mathbf{X}_B^{\text{OrthA}})^T \mathbf{X}_B^{\text{OrthA}} \mathbf{F}) = \text{tot}_{\mathbf{X}_A} + \text{additional}_{\mathbf{X}_B}. \quad (9)$$

The total and additional effects are crucial in step 1 for interpreting where prediction ability is located in the path diagram.

5.4 | The direct and indirect effects of \mathbf{X}_A on \mathbf{X}_C

The intuitive idea behind the concepts of direct and indirect effects (step 2) of a block A on another block C is that the total effect of A on C as defined above can be split up in an effect/influence that goes directly from A to C and one that goes via a block B lying between the two in the topological order. In order to make the definition precise, a simple rewriting of the model is needed. Note that using regression coefficients between latent variables as done in standard path modeling is generally not possible here because of the multidimensional nature of the blocks.

The model 3 can be modified in the same way as in Equation 4 by changing the order of the two blocks, that is, as

$$\mathbf{X}_C = \mathbf{X}_B \tilde{\mathbf{F}} + \mathbf{X}_A^{\text{OrthB}} \mathbf{D}. \quad (10)$$

The direct effect of \mathbf{X}_A on \mathbf{X}_C is then defined as the variance contribution of the last term in Equation 10, that is, as

$$\text{direct}_{\mathbf{X}_A} = \text{tr}(\mathbf{D}^T (\mathbf{X}_A^{\text{OrthB}})^T \mathbf{X}_A^{\text{OrthB}} \mathbf{D}). \quad (11)$$

In other words, the direct effect is obtained as the effect of \mathbf{X}_A on \mathbf{X}_C , which is based on the part of A that is orthogonal to \mathbf{X}_B , that is, uncorrelated with B. The direct effect is the effect \mathbf{X}_A that is not related to \mathbf{X}_B or does not go via \mathbf{X}_B .

Note that this is the same as the variance contribution obtained by regressing \mathbf{X}_C onto the residual of \mathbf{X}_A when fitted to \mathbf{X}_B (i.e., $\mathbf{X}_A^{\text{OrthB}}$) using the equation:

$$\mathbf{X}_C = \mathbf{X}_A^{\text{OrthB}} \mathbf{D}. \quad (12)$$

Note that \mathbf{X}_C here comprises more than the term to the right of the equality sign, namely, $\mathbf{X}_B \tilde{\mathbf{F}}$, but this is not needed here since the two contributions are orthogonal. Since $\mathbf{X}_A^{\text{OrthB}}$ is not always full rank, for instance, a Moore-Penrose solution is needed here (or PLS which is usually used for estimation). Note that \mathbf{X}_C could also be orthogonalized without changing the value of the direct effect.

This definition of direct effect is directly inspired by partial correlation in the one-dimensional case, which is essentially a correlation between residuals of \mathbf{X}_A and \mathbf{X}_C after fitting to \mathbf{X}_B . It is possible also here to use residuals for both \mathbf{X}_A and \mathbf{X}_C in the regression (projection is idempotent), but the explained variance must then be multiplied by the ratio of the variance of \mathbf{X}_C and the residual in order to obtain the correct direct effect of \mathbf{X}_A on \mathbf{X}_C (and not the effect on the residual of \mathbf{X}_C).

The indirect effect is then simply defined as the difference between the total and direct effect, that is, as

$$\text{indirect}_{\mathbf{X}_A} = \text{tot}_{\mathbf{X}_A} - \text{direct}_{\mathbf{X}_A}. \quad (13)$$

In other words, the indirect effect is defined as the variance contribution to \mathbf{X}_C from \mathbf{X}_A that is not already taken care of by the direct effects, that is, the effect that goes through other blocks.

Note that there is no guarantee that the indirect effect is positive. We have never seen it for real data, but in simulations, we have been able to create a negative indirect effect. A negative indirect effect will in this case mean that some parts of the direct and indirect contribution of \mathbf{X}_A on \mathbf{X}_C pull in different directions where one cancels the other. This effect is closely linked to the fact that a partial correlation between two variables can be larger than the direct correlation between them. A very simple example of this is given in Appendix C.

5.5 | Relations between all effects in Figure 6

Equations 9 and 13 together lead to

$$\text{tr}(\mathbf{X}_C^T \mathbf{X}_C) = \text{tot}_{\mathbf{X}_A} + \text{additional}_{\mathbf{X}_B} = \text{direct}_{\mathbf{X}_A} + \text{indirect}_{\mathbf{X}_A} + \text{additional}_{\mathbf{X}_B} \quad (14)$$

binding together all the different effects for the model structure in 3. See Section 6.4 for further discussion of the interpretation of this.

It is important to stress that this type of relation is not universally valid for several blocks in between \mathbf{X}_A and \mathbf{X}_C (see below) and for nonunique topological orders. For instance, in the diagram in Figure 1A, the additional effect of blocks \mathbf{X}_E and \mathbf{X}_F in the topological order (A B E F C D) is calculated for the output block \mathbf{X}_D even if they are not part of the reachable relation between instance \mathbf{X}_A and \mathbf{X}_D . Therefore, the direct effect of \mathbf{X}_A on \mathbf{X}_D in this case is obtained by conditioning only on \mathbf{X}_B and \mathbf{X}_C and without taking \mathbf{X}_E and \mathbf{X}_F into account (see Section 5.6).

In particular, it can be seen that for a given overall variance explained by \mathbf{X}_A and \mathbf{X}_B on \mathbf{X}_C and a given direct effect of \mathbf{X}_A on \mathbf{X}_C , the sum of the additional and indirect variance is constant. This means that the more of \mathbf{X}_A that is transmitted through \mathbf{X}_B , the less extra information the unique part of the \mathbf{X}_B block has in predicting \mathbf{X}_C .

A simple numerical example of the different effects is given in Appendix D.

5.6 | More complex path models

This section concentrates on the extension of total, indirect, and direct effects to situations when there are more blocks “between” \mathbf{X}_A and \mathbf{X}_C than in the example provided in Figure 6.

5.6.1 | Additional and total effects in more complex relations

Equations 7 and 8 hold for the simple situation depicted in Figure 6 and are useful for establishing the concept. When there are more blocks involved, however (as in for instance Figure 1), the additional effects are defined by sequentially orthogonalizing blocks in the topological order, starting with the block to the left in the dependence diagram. When for instance modeling of \mathbf{X}_D using the topological order represented in Figure 1B, one starts fitting \mathbf{X}_A , then orthogonalizes \mathbf{X}_B with respect to \mathbf{X}_A , before orthogonalizing \mathbf{X}_C with respect to \mathbf{X}_A and \mathbf{X}_B , and so forth. The additional effects are defined as the incremental variance contributions obtained as each new block in the order is orthogonalized and fitted.

Note that the variance contribution of the first step of an SO-PLS model is what is called the total effect of that block on the output (see step 2 and Section 5.5). Note also that the additional effect contributions can be obtained theoretically (in the population as above) by sequential use of the same trick as used when going from model 3 to model 4, but the formulas become quite complex and do not shed light on the concept. They are therefore not presented here.

5.6.2 | Direct and indirect effects for more blocks than one between X_A and X_C

In this paper, we propose to calculate the direct and indirect effects of a block (input block, X_A) on another (output block, X_C) in the following way: The conditioning used for the direct effects is now with respect to all blocks between X_A and X_C in a reachable relation. There could for instance be two or even more blocks between them in the topological order.

As an example, let us first consider Figure 1B with a unique topological order. The direct effect of for instance X_A on X_D is here defined conditioning on all blocks between the two in the topological order, that is, X_B and X_C . An example of a nonunique order is given in Figure 2A, which is the transitive closure of Figure 1A. In this case, the direct effect of X_A on X_D is obtained by conditioning only on X_B and X_C . The X_E and X_F are kept out since they do not belong to the reachable relations between X_A and X_D .

Note that these definitions are independent on the first interpretation stage of SO-PLS (step 1 in Appendix A).

5.7 | Estimating the effects

In practice, the calculation of the additional effect comes from stage 1 of the SO-PLS-PM (see Appendix A). The total and direct effects can be calculated by simply regressing X_C onto X_A and on $X_A^{\text{orth}B}$, respectively. The indirect effect is then as above calculated as the difference between total and direct effects.

When using SO-PLS in the conditioning process with several more intermediate blocks between X_A and X_C , one will again normally incorporate the blocks according to how far they are from the input in terms of how many paths one needs to reach X_A . Since interpretation along the way is generally not done and experience indicates that block ordering in SO-PLS does not necessarily have any strong effect on predictions (and consequently on the residuals), it is less important here which order is chosen.

In the example below, sequential CV is used to find the number of components from each block. The validated explained variances for the various effects are then estimated from the models, and bootstrapping is applied to assess the variability of the effects.¹³ In practice, one will estimate the effects using CV.¹³ This may possibly lead to more conservative estimates of the direct and indirect effects than one will obtain by the standard PLS-PM methodology since these are based on direct fitting. Also, one loses the exact relations between the effects in Equation 14 and ends up with approximate relations between effects.

6 | MISCELLANEOUS ASPECTS OF THE SO-PLS-PM METHOD

6.1 | Graphical representation

All effects can be visualized graphically,¹³ which helps interpretation and gives a good overview. These plots are based on the same path diagram as the one used as point of departure, but now, there are two sets of lines between blocks, one set for the direct effects and one set for the indirect with the nonsignificant ones eliminated. The thickness of the lines is related to the level of significance of the effect; the thicker the line is, the more important is the path. Significance is here defined using the bootstrap.¹³ An example of this will be shown in the process monitoring example below (Figure 10B).

6.2 | Transitive closure and checking for new or unnecessary paths

Since SO-PLS works in the transitive closure of the path diagram and direct effects are calculated for all combinations in reachable relations, this opens the opportunity to check for possible omitted paths in the original diagram. This aspect was illustrated and utilized for the mobile data in Romano et al.¹³ in order to detect a new edge not added in the original diagram. It was shown by calculating all direct effect that one of the paths not present in the original diagram was in fact quite important (Figure 1C,D). This aspect has no effect on the SO-PLS solution since it already operates in the transitive closure, but it was shown that it had a quite substantial effect on the PLS-PM results. In other words, the PLS-PM model depends heavily on this type of correctness of the full path model, which is not the case for SO-PLS-PM.

The question is then whether one should rely on the original model or on the effects seen in the data; our general preference will then be for data. This result indicates a possible robustness when using SO-PLS-PM instead of the standard PLS-PM.

If, on the other hand, elimination of nonsignificant direct effects is considered, the transitive reduction will be the final step possible within the path model considered. Such a reduction can possibly be useful for interpretation, but note that such a reduction has no effect on SO-PLS-PM that operates in the closure.

6.3 | Different components for prediction and to be predicted

Since more than one component is allowed in each of the blocks, a possible scenario is that the block \mathbf{X}_B (see Figure 6) has a component that can be predicted from \mathbf{X}_A and another component that can be used for predicting \mathbf{X}_C . In such a case, both the direct effect of blocks \mathbf{X}_A on \mathbf{X}_B and the additional effect of block \mathbf{X}_B on \mathbf{X}_C are present, while the indirect effect of \mathbf{X}_A is equal to 0. See also Figure 7 for an illustration of this phenomenon. We also refer to the simulation below for a numerical illustration and to Menichelli et al.⁸ for discussion and a demonstration of this issue.

More generally one can say that \mathbf{X}_B may contain a part (component) that transmits information from \mathbf{X}_A to \mathbf{X}_C and one part which is unique for \mathbf{X}_B in predicting \mathbf{X}_C .

6.4 | Relations to common and distinct component definitions for symmetric methods

In Smilde et al.¹⁵ a number of definitions related to common and distinct components among sets of blocks, \mathbf{X}_A and \mathbf{X}_B , were proposed (without reference to regression). The common components were for noise-free data defined as linear combinations of the columns of the blocks \mathbf{X}_A and \mathbf{X}_B with correlation (canonical correlation) equal to 1. Distinct subspaces of \mathbf{X}_A and \mathbf{X}_B were then defined using the concept of orthogonal complements of \mathbf{X}_A and \mathbf{X}_B with respect to the common space (spanned by the common components). Distinct components can be defined for instance by orthogonalizing the subspaces of \mathbf{X}_A and \mathbf{X}_B with respect to the common subspace and applying PCA to obtain the components. Various implementations of these concepts for real data were also proposed and discussed in Smilde et al.¹⁵

Equation 14 refers to prediction; that is, how well the different parts of \mathbf{X}_A and \mathbf{X}_B explain \mathbf{X}_C , but apart from that, they bear a resemblance with the definitions of common and distinct components. The first term in 14 measures the contribution on \mathbf{X}_C from the part of \mathbf{X}_A that is orthogonal to \mathbf{X}_B , and the last term measures how much of the variation in \mathbf{X}_C is explained by \mathbf{X}_B after orthogonalization with respect to \mathbf{X}_A . This strongly resembles the definition of distinct components in Smilde et al.¹⁵; the only difference is that here we orthogonalize with respect to the whole opposite block (\mathbf{X}_A or \mathbf{X}_B), while in Smilde et al.¹⁵ orthogonalization is done with respect to the common components. The term in the middle in Equation 14 represents the part of \mathbf{X}_C that is not predicted by the two orthogonal components and therefore represents a kind of common component contribution. In conclusion, one can say that the decomposition in 14 resembles a decomposition into common and distinct components and how the three parts contribute to the variability in \mathbf{X}_C .

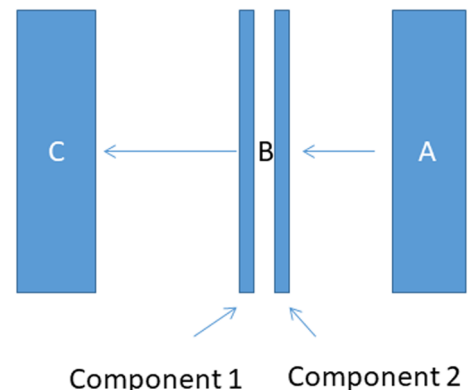


FIGURE 7 Illustration of a situation where block A influences one of the components in block B, while another component in block B influences block C

6.5 | Relations between definitions of direct and indirect effects for SO-PLS-PM and PLS-PM

In the definitions above, the variance explained is used as criterion for all estimated effects. The standard definitions used in LISREL and PLS-PM for the effects are, however, based on regression coefficients. The direct effects are defined as the regression coefficients for the direct paths, and the indirect effects are obtained by multiplying the regression coefficients for the indirect paths. Total effects are defined as sums of the two. In Romano et al.¹³ the two sets of definitions were compared empirically, and it was found that they were quite well correlated in situations where they could be compared (i.e., for one-dimensional blocks), indicating that they at least in some cases measure similar phenomena.

An important difference between the two definitions is that while the SO-PLS definitions are valid for any dimensionality of the blocks. For methods depending on a one-dimensional representation of the latent variable in each block (for instance PLS-PM), this is not the case. Another difference is that the focus in the definitions is slightly different. Look for instance at the blocks A, B, C diagram in Figure 6. The direct effect of \mathbf{X}_B on \mathbf{X}_C is here defined without taking \mathbf{X}_A , or maybe other blocks with paths into \mathbf{X}_C , into account. For the standard method, all paths leading into C are considered simultaneously in a multivariate regression approach. In other words, all our definitions are only focused on effects in reachable relations without taking other factors/blocks into account. Put another way, only the input and output block and blocks between them in a reachable relation are involved.

The SO-PLS-PM method introduces the concept of additional effect, which is not considered in standard methods. This effect provides information about the additional contribution of the block \mathbf{X}_B in predicting \mathbf{X}_C when \mathbf{X}_A is already in the model. It represents a measure of the incremental contribution to the explained variance (or R^2 in the standard PLS-PM framework). This effect plays an important role in model building, model interpretation, and for determining which blocks are important in the path diagram¹³ (see examples below).

Another important difference is that the definitions for SO-PLS are made for the real manifest data and not for the latent variables as is the case for PLS-PM.

6.6 | Relations to causal model building and partial RV

The above definitions of direct effects are based on explained variances obtained from orthogonalized blocks. In the case of one-dimensional blocks, this is more or less equivalent to calculating partial correlation. The only difference is that instead of calculating explained variance, one looks at the correlation between residuals. In the multivariate case, there are a number of measures for measuring similarity between matrices. One of these is the SMI,²³ another one is the RV coefficient.²⁴ The relation between these similarity measures and explained variance is the same as between explained variance and partial correlation in the univariate case. The partial RV coefficients obtained on residuals are used in Aben et al.²⁵ for inferring causality relations among the blocks.

7 | SIMULATED EXAMPLE

The main goal of this example is to show the importance of taking multidimensionality into account and the possible dangers of not doing it.

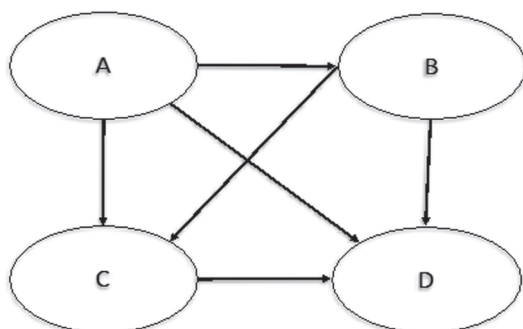


FIGURE 8 Path diagram for the simulated example

The path diagram for the simulation is given in Figure 8. All blocks are designed to have two components. The data were obtained by first generating the components of blocks \mathbf{X}_A at random and then constructing the components of \mathbf{X}_B , \mathbf{X}_C , and \mathbf{X}_D by using linear combinations of \mathbf{X}_A . An exception is for block \mathbf{X}_B where the component 1 in \mathbf{X}_B is generated as a linear function of \mathbf{X}_A , while component 2 is generated at random and is the only part of \mathbf{X}_B that influence blocks \mathbf{X}_C . There is no link between \mathbf{X}_A and \mathbf{X}_C through \mathbf{X}_B . Once the components for the four first blocks were obtained, these were multiplied by random loadings (10 for each block), thus obtaining the manifest variables. Random noise was added to each block. The unique topological order is A, B, C, D. The diagram is the transitive closure of itself.

The SO-PLS-PM results for step 1 (see Appendix A) are presented in Table 1. As can be seen, all blocks have an effect. In particular, one can see that only one component is needed in \mathbf{X}_B to predict \mathbf{X}_C , which corresponds to how it was constructed. We can also see that \mathbf{X}_A has a strong effect on \mathbf{X}_B and that \mathbf{X}_B has a clear additional effect of $20.8 = 53.80 - 33.00$ for predicting \mathbf{X}_C , which also corresponds to how the data are generated.

The total, direct, and indirect effects are presented in Table 2. As can be seen, the indirect effect of \mathbf{X}_A on \mathbf{X}_C is quite small despite the fact that \mathbf{X}_A is important for \mathbf{X}_B and \mathbf{X}_B is important for predicting \mathbf{X}_C . The reason is that there are two different components of \mathbf{X}_B , one that can be predicted by \mathbf{X}_A and one that is used for predicting \mathbf{X}_C . Therefore, only one component is needed in \mathbf{X}_B , as described above.

In order to illustrate that PLS-PM¹⁰ is not usable for this data set, we did a standard analysis without checking for unidimensionality (Table 3, which of course never should be done in practice). One can see that there is for instance a

TABLE 1 Explained variances for SO-PLS-PM for the simulated data (step 1, Appendix A) The number of components used are given in parenthesis

Block	Model for B	Model for C	Model for D
A	21.95 (1)	33.00 (2)	73.48 (1)
B		53.80 (1)	76.98 (1)
C			80.66 (2)

Note. Each column represents a model, that is, for the endogenous blocks B, C, and D. The rows represent the cumulated explained variances as the different blocks are incorporated. For the first model, first column, only A is needed. For the next columns, both A and B are used as input, and for the last column, the three blocks A, B, and C are used as input.

Abbreviation: SO-PLS-PM: sequential and orthogonalized partial least squares-path modeling.

TABLE 2 Direct, indirect, and total effects for the SO-PLS-PM model for the simulated data (step 2, Appendix A) The standard errors and the number of components used are given in parenthesis

	Direct Effect	Indirect Effect	Total Effect
A→B	21.95 (1.24, 1)	0 (0)	21.95 (1.24, 1)
A→C	25.01 (0.96, 2)	7.99 (1.41)	33.00 (1.66, 2)
A→D	18.60 (0.78, 2)	56.40 (1.05)	75.00 (1.12, 2)
B→C	27.74 (1.92, 2)	0 (0)	27.74 (1.92, 2)
B→D	1.76 (0.88, 1)	0 (0)	1.76 (0.88, 1)
C→D	66.34 (1.75, 2)	0 (0)	66.34 (1.75, 2)

Abbreviation: SO-PLS-PM: sequential and orthogonalized partial least squares-path modeling.

TABLE 3 Direct, indirect, and total effects for the PLS-PM for the simulated data

	Direct	Indirect	Total
A > B	0.52	0.00	0.52
A > C	-0.89	0.02	-0.87
A > D	-0.18	-0.56	-0.74
B > C	0.04	0.00	0.04
B > D	0.35	0.04	0.38
C > D	0.84	0.00	0.84

Abbreviation: PLS-PM: partial least squares-path modeling.

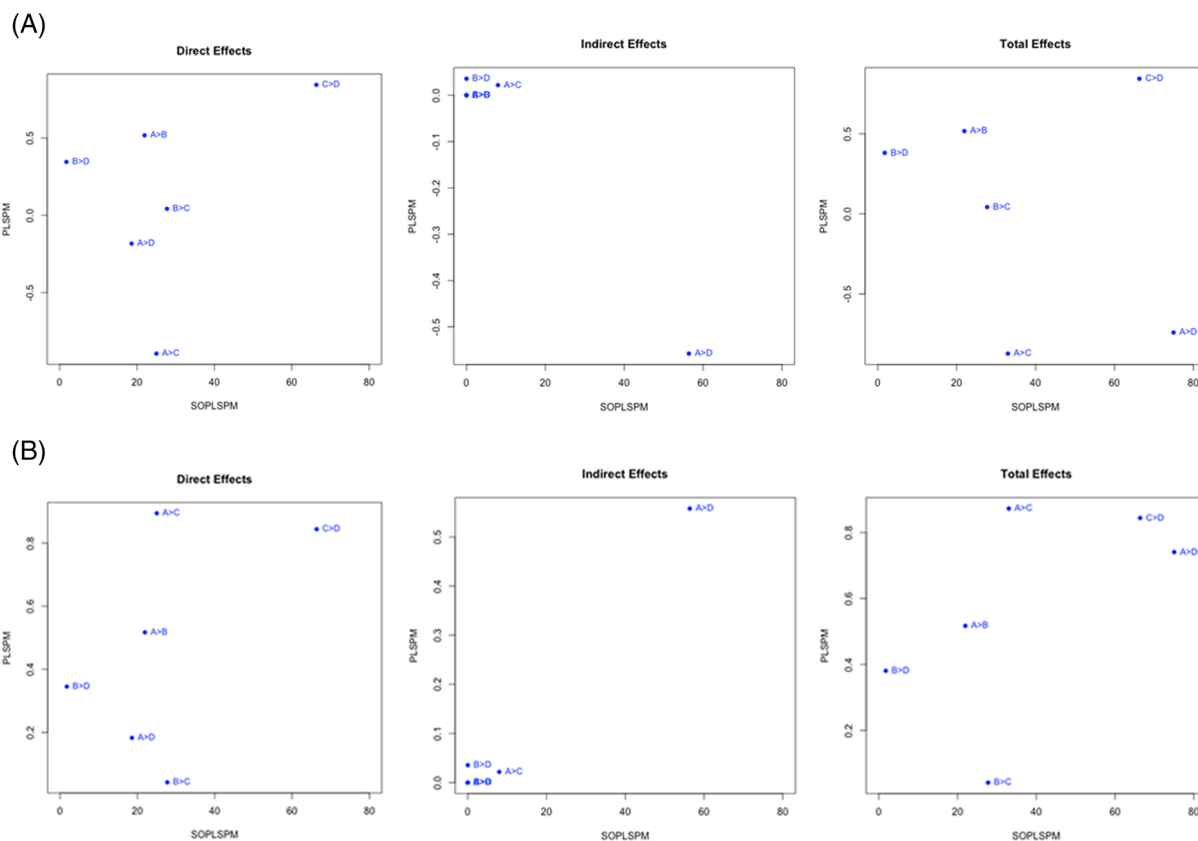


FIGURE 9 Simulated data. Plot of relation between direct, indirect, and total effects for sequential and orthogonalized partial least squares-path modeling (SO-PLS-PM) and for PLS-PM. (A) The true values; (B) The absolute values of the values in Figure 9A

very low direct effect of \mathbf{X}_B on \mathbf{X}_C , which means that the method is not able to detect the special structure of \mathbf{X}_B . In other words, the solution is clearly misleading in this case.

The plot of the direct, indirect, and total effects for the two methods are presented in Figure 9A. Since the PLS-PM also can give negative values, the absolute values were also compared (Figure 9B). As can be seen, in this case, there is little correspondence between the two definitions, which is different from the results obtained in Romano et al.¹³ where most of the blocks were one-dimensional. One can, however, see that the relation is slightly stronger if we disregard the sign of the effects. According to the interpretation above, the SO-PLS-PM solution is clearly more relevant in this case since the blocks are multidimensional.

For comparison, also the PLS-PM solution for the subdiagram with only the blocks \mathbf{X}_A , \mathbf{X}_B , and \mathbf{X}_C was calculated. The results are presented in Table 4. As can be seen, some elements of this are extremely different from the part of the table above that corresponds to \mathbf{X}_A , \mathbf{X}_B , and \mathbf{X}_C . In particular, $A \rightarrow B$ and $B \rightarrow C$ are different in the two tables. This shows that the solution for multidimensional data can be very dependent on the whole diagram if PLS-PM is used uncritically. In other words, the solution changes dramatically for the comparable parts depending on which other blocks that are incorporated. This is not the case for SO-PLS-PM where the results will be the same for the full diagram and the \mathbf{X}_A , \mathbf{X}_B , \mathbf{X}_C diagram.

TABLE 4 Direct, indirect, and total effects for the reduced model based only on A, B, C using PLS-PM for the simulated data

	Direct	Indirect	Total
A > B	-0.75	0.00	-0.75
A > C	-1.01	0.26	-0.76
B > C	-0.34	0.00	-0.34

Abbreviation: PLS-PM: partial least squares-path modeling.

8 | REAL EXAMPLE BASED ON PROCESS DATA

The example is taken from production of a Gouda-type cheese. In short, the cheese is produced by first pasteurizing the raw milk and standardizing the fat/protein ratio. Then, starter culture and rennet are added, making the milk coagulate. Then the curd is cut and heated, the whey is drained out, and the curd pieces are pressed into blocks.

The raw milk has a substantial variation in composition, depending on for instance how the cow is fed, which again varies between farms and seasons. The starter culture is produced on-site from milk and specific bacteria cultures and varies in activity. Currently, some process parameters are adjusted manually based on lab measurements of starter activity, observations of the curd and end-product quality measurements of previous batches made from the same milk and starter (in a feed-backward manner). This procedure is, however, not optimal, as the delay is substantial, and the adjustments are subjective and highly dependent on the experience of the operator. Also, an undesired variation in end-product quality is observed despite efforts to keep it constant. The main barrier for developing a better control strategy is that the relationships between the sources of variation and end-product quality are not fully understood. An important task in cheese production is therefore to interpret and quantify the effect of each production step on the end-product quality, in order to increase process knowledge and identify important control points.

Data from 795 batches over a time span of almost 4 years were collected from the most important process steps:

- A: Raw milk (six variables representing chemical composition of milk, storage time, and storage temperature)
- B: Production of starter culture (28 variables, including pH and temperatures during production, measured activity, and storage time)
- C: Controllable settings in the cheese vat (43 variables representing times and temperatures during, for example, mixing, coagulation, cutting, and heating)
- D: On-line measurements (three variables representing pH at different time points from pasteurized milk to fresh cheese)
- E: End-product quality. For simplicity, we have chosen one quality characteristic in this example.

The path diagram is given in Figure 10A. This is an example of nonunique topological order. The topological orders possible here are as follows:

- A B C D E
- A C B D E
- B A C D E
- B C A D E
- C A B D E
- C B A D E

Step 1 of the SO-PLS method

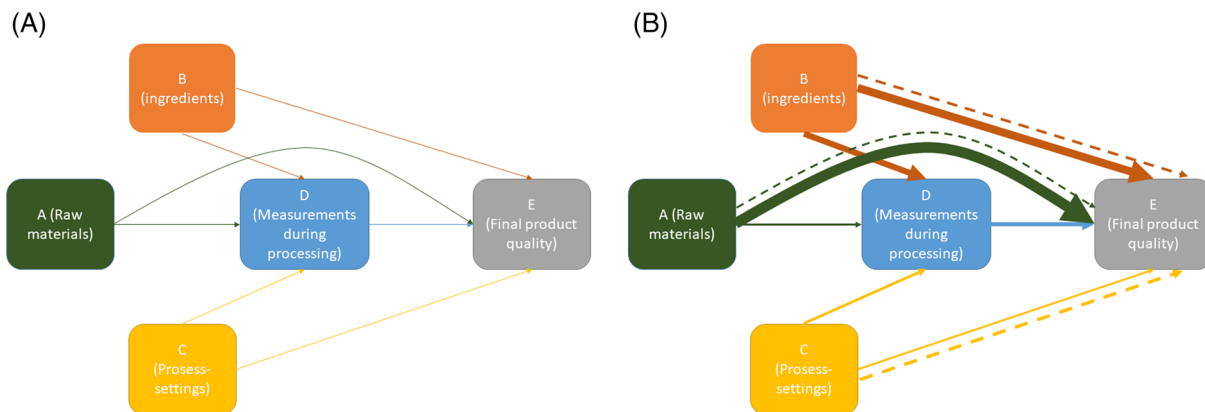


FIGURE 10 Path diagrams for the process control example. A, The basic path diagram for the study. B, The graphical illustration of the effects obtained by the sequential and orthogonalized partial least squares-path modeling (SO-PLS-PM) method. Only significant effects as determined by the bootstrap are presented. The solid lines represent the direct effects and the dotted lines the indirect effects

For the step 1 of the method, see Appendix A; all blocks are involved in modeling. In this case, we will in particular be interested in the first topological order (A B C D E) above since raw materials and ingredients are natural to consider before considering the process settings. Therefore, two models are considered, $\mathbf{X}_A, \mathbf{X}_B, \mathbf{X}_C$ versus \mathbf{X}_D and $\mathbf{X}_A, \mathbf{X}_B, \mathbf{X}_C, \mathbf{X}_D$ versus \mathbf{X}_E . The blocks $\mathbf{X}_A, \mathbf{X}_B$, and \mathbf{X}_C are exogenous blocks, and no model is possible for them. The results from the calculations are presented in Table 5 with standard errors (obtained by the bootstrap). The number of components is presented in parentheses. The additional effects can be read directly out of the table, for instance, the additional effect of adding block C to the model for block E when blocks A and B are already in the model is equal to 1.92.

From Table 5, we can see that block \mathbf{X}_C (process conditions) adds very little in both models and that the end-product block \mathbf{X}_E is described better than block \mathbf{X}_D (measurements taken during processing). For the \mathbf{X}_E model, the block \mathbf{X}_A (milk quality) is the most important. Block \mathbf{X}_D adds nothing when $\mathbf{X}_A, \mathbf{X}_B$, and \mathbf{X}_C are already in the model for \mathbf{X}_E . In other words, little extra information is gained from measurements taken during processing.

Step 2 of the SO-PLS method

Note that only the three first letters change the order between them in the topological orders above; \mathbf{X}_D and \mathbf{X}_E are always last. Note also that for the three first letters, there is no reachable relation between them, and therefore, no model can be made for $\mathbf{X}_A, \mathbf{X}_B$, and \mathbf{X}_C (they are exogenous blocks). Only the models for \mathbf{X}_D and \mathbf{X}_E are of interest. For \mathbf{X}_D , only direct effect on \mathbf{X}_E can be calculated. For calculating the direct, indirect, and total effects of blocks $\mathbf{X}_A, \mathbf{X}_B$, and \mathbf{X}_C on \mathbf{X}_E , the three paths represented by $\mathbf{X}_A, \mathbf{X}_B$, and \mathbf{X}_C must be treated separately. The reason for this is that only blocks in between the two in a reachable relation are involved in these calculations. For instance, when considering the indirect effect of \mathbf{X}_A on \mathbf{X}_E , the blocks \mathbf{X}_B and \mathbf{X}_C are not in between the two in a reachable relation; only \mathbf{X}_D is. They are therefore not used in the conditioning.

TABLE 5 SO-PLS-PM results for process modeling example. The standard errors and the number of components used are given in parenthesis

Used as Input	D as Output	E as Output
A	3.6 (0.92, 2)	29.54 (2.28, 2)
B	10.6 (0.81, 3)	14.4 (2.08, 4)
C	2.91 (1.41, 3)	1.92 (1.9, 1)
D		0 (1.83, 0)
Total	17.11 (0.92)	45.86 (2.28)

Note. Explained variances results for step 1 in Appendix A. The numbers presented are the additional effects of the blocks as they are added in the two models for D and E.

Abbreviation: SO-PLS-PM: sequential and orthogonalized partial least squares-path modeling.

TABLE 6 SO-PLS-PM results for process modeling example. The standard errors and the number of components used are given in parenthesis

	Direct Effect	Indirect Effect	Total Effect
A→D	4.07 (0.96, 5)	0	4.07 (0.96, 5)
A→E	25.24 (2.29, 5)	5.14 (0.83)	30.39 (2.52, 5)
D→E	10.05 (2.00, 3)	0	10.05 (2.00, 3)
B→D	13.04 (1.20, 5)	0	13.04 (1.20, 5)
B→E	18.88 (2.03, 5)	5.89 (1.16)	24.78 (2.30, 5)
D→E	10.05 (2.00, 3)	0	10.05 (2.00, 3)
C→D	6.62 (1.27, 5)	0	6.62 (1.27, 5)
C→E	5.86 (2.10, 5)	8.38 (1.30)	14.24 (2.50, 5)
D→E	10.05 (2.00, 3)	0	10.05 (2.00, 3)

Note. Effects obtained for step 2 in Appendix A.

Abbreviation: SO-PLS-PM: sequential and orthogonalized partial least squares-path modeling.

From Table 6, we can see that there are generally small indirect effects through \mathbf{X}_D , at least for the models based on \mathbf{X}_A and \mathbf{X}_B . This supports the results in Table 5, which say that measurements taken during processing add quite little (as compared with direct effects) to explaining this particular quality property. As can also be seen, block \mathbf{X}_A has a larger effect than the blocks \mathbf{X}_B and \mathbf{X}_C on \mathbf{X}_E ; that is, milk properties are more important than processing for the end-product variation. The largest direct effect on \mathbf{X}_D is \mathbf{X}_B , which is natural since the starter culture (represented by \mathbf{X}_B) is an acid, which makes the pH (represented by \mathbf{X}_D) drop. The block \mathbf{X}_D has a clear but moderate effect on \mathbf{X}_E .

Contributions from individual variables within the blocks are given by the loadings, which can be plotted in scatter plots and interpreted as for regular PLS regression. In this case, the loading plots (not shown) obtained from the total effect calculations revealed that the fat and protein content are the most important characteristics of the raw milk (block A), in combination with storage temperature. For block B, the most important variables are related to the pH trajectory during production of starter culture. This suggests that reducing the variation in these specific parameters may result in a more stable end-product quality.

Figure 10B gives an overview of the results using the same diagram setup as in Figure 10A. The dotted lines represent the indirect effects and the solid lines the direct effect. The thickness of the lines represents the degree of significance of the effects. Nonsignificant effects are not represented.

9 | CONCLUSIONS

In this paper, we have linked the SO-PLS-PM path modeling to the theory of DAGs. In particular, this was used to define precisely the order of the blocks to incorporate in the different SO-PLS models. It was also shown that SO-PLS-PM is not influenced by the process of generating transitive closure of a path diagram. In addition, definitions of direct and indirect effects were made precise and discussed in terms of definitions related to common and distinct components in multiblock analysis. Relations between the different effects were clarified. The fact that some components/dimensions in a block \mathbf{X}_A may be influenced by a block \mathbf{X}_B while other dimensions of the same block \mathbf{X}_A may influence another block \mathbf{X}_C was highlighted and illustrated in an example. Both real data from process modeling as well as simulated data were used for illustrating the theory.

ACKNOWLEDGEMENTS

We would like to thank the Research Council of Norway for financial support and University of Naples, Federico II, for providing a travel grant.

ORCID

Tormod Næs  <https://orcid.org/0000-0001-5610-3955>

Ingrid Måge  <https://orcid.org/0000-0003-0364-0225>

Age Smilde  <https://orcid.org/0000-0002-3052-4644>

Kristian H. Liland  <https://orcid.org/0000-0001-6468-9423>

REFERENCES

1. Wright S. On the nature of size factors. *Genetics*. 1918;3(4):367-374.
2. Duncan OD. Path analysis: sociological examples. *Am J Sociol*. 1966;72(1):1-16.
3. Finney JM. Indirect effects in path analysis. *Socio Meth Res*. 1972;1(2):175-186.
4. Greene VL. An algorithm for total and indirect causal effects. *Polit Anal*. 1977;369-381.
5. Jöreskog KG. Structural analysis of covariance and correlation matrices. *Psychometrika*. 1978;43(4):443-477.
6. Graff J, Schmidt P. A general model for decomposition of effects. In: Jöreskog K, Wold H, eds. *Systems Under Indirect Observation: Causality, Structure, Prediction*. Amsterdam: North-Holland; 1982:131-148.
7. Spearman C. "General Intelligence," objectively determined and measured. *Am J Psychol*. 1904;15(2):201-292.
8. Menichelli E, Almøy T, Tomic O, Olsen NV, Næs T. SO-PLS as an exploratory tool for path modelling. *Food Qual Prefer*. 2014;36:122-134.
9. Wold H. Partial least squares. In: Kotz S, Johnson NL, eds. *Encyclopedia of Statistical Sciences*. New York: Wiley; 1985:581-591.
10. Tenenhaus M, Vinzi VE, Chetelin Y-M, Lauro C. PLS path modeling. *Comput Stat Data Anal*. 2005;48:159-205.
11. Næs T, Tomic O, Mevik BH, Martens H. Path modelling by sequential PLS regression. *J Chemometr*. 2011;25(1):28-40.

12. Næs T, Tomic O, Afseth NK, Segtnan V, Måge I. Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemom Intel Lab Syst.* 2013;124:32-42.
13. Romano R, Tomic O, Liland KH, Smilde A, Næs T. A comparison of two PLS based approaches to structural equation modeling. *J Chemometr.* 2019;33(3). <https://doi.org/10.1002/cem3105>
14. Edwards D. *Introduction to graphical modelling.* NY: Springer; 2000.
15. Smilde AK, Måge I, Næs T, et al. Common and distinct components in data fusion. *J Chemometr.* 2017;31(7). <https://doi.org/10.1002/cem2900>
16. Wikipedia: https://en.wikipedia.org/wiki/Directed_acyclic_graph
17. Niimi J, Tomic O, Næs T, Jeffrey DW, Bastian SEP, Boss PK. Application of sequential and orthogonalised-partial least squares (SO-PLS) regression to predict sensory properties of Cabernet Sauvignon wines from grape chemical composition. *Food Chem.* 2018;256:195-202.
18. Lauzon-Gauthier J, Manolescu P, Duchesne C. The sequential multi-block PLS algorithm (SMB-PLS). Comparison of performance and interpretability. *Chemom Intel Lab Syst.* 2018;180:72-83.
19. Jørgensen K, Næs T. A design and analysis strategy for situations with uncontrolled raw material variation. *J Chemometr.* 2004;18:45-52.
20. Indahl UG, Næs T. Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling. *J Chemometr.* 1998;12(4):261-278.
21. Langsrud Ø, Næs T. Optimised score plot by principal components of prediction. *Chemom Intel Lab Syst.* 2003;68:61-74.
22. Bollen KA. *Structural Equations with Latent Variables.* NY: Wiley; 1989.
23. Indahl UG, Næs T, Liland KH. A similarity index for comparing coupled matrices. *J Chemometr.* 2018;32:10. <https://doi.org/10.1002/cem3049>
24. Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: the *RV*-coefficient. *Appl Stat.* 1976;25(3):257-265.
25. Aben N, Westerhuis JA, Song Y, et al. iTOP: inferring the topology of omics data. *Bionformatics.* 2018;34:i988-i996.

How to cite this article: Næs T, Romano R, Tomic O, Måge I, Smilde A, Liland KH. Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. *Journal of Chemometrics.* 2020;e3243. <https://doi.org/10.1002/cem.3243>

APPENDIX A: DESCRIPTION OF SO-PLS-PM PROCEDURE

Step 1

Scopes

- Identify how well the different endogenous blocks can be predicted and which blocks that contribute to the predictions.
- Interpret how the input blocks relate to the output block

Steps taken to reach the goal

- For each endogenous block establish the dependence diagrams; choose the most appropriate order if the topological order is not unique.
- For each endogenous block estimate the prediction ability and additional effects of the blocks in the model using SO-PLS regression
- For each endogenous block calculate the PCP plot.

Step 2

Scopes

- Estimate, present, and interpret direct, indirect, and total effects. The number of components is determined independently from step 1 since different blocks are involved.

Steps taken to reach the goal

- For each combination of blocks in a reachable relation, the output block is regressed onto the input block and the explained variance using CV is calculated. This is the total effect.
- For each combination of blocks in a reachable relation, the dependent block is regressed onto the residual of the input block when regressed onto all blocks between the two in a reachable relation. Also, here, the CV is used. This gives the direct effect.
- The indirect effect is calculated as the difference between the two.
- The effects are presented in the path diagram, and the standard errors are calculated by the bootstrap.

The two steps are linked through the topological order of the blocks, which is underlying the estimation process in step 1.

APPENDIX B: ALTERNATIVE FORMULA FOR D IN EQUATION 6

The original linear model in 3 can alternatively be written as

$$\mathbf{X}_C = \mathbf{X}_A \tilde{\mathbf{D}} + \mathbf{X}_B^{\text{orthA}} \mathbf{F} \tag{B1}$$

where \mathbf{X}_B is orthogonalized with respect to \mathbf{X}_A . This model is obtained by splitting the \mathbf{X}_B in model 3 up in the projection onto \mathbf{X}_A and the one orthogonal to it, that is,

$$\mathbf{X}_C = \mathbf{X}_A \mathbf{D} + \left(\mathbf{X}_A (\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \right) \mathbf{X}_B \mathbf{F} + \left(\mathbf{I} - \mathbf{X}_A (\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \right) \mathbf{X}_B \mathbf{F} = \mathbf{X}_A \tilde{\mathbf{D}} + \mathbf{X}_B^{\text{OrthA}} \mathbf{F}, \tag{B2}$$

where

$$\tilde{\mathbf{D}} = \mathbf{D} + \left(\mathbf{X}_A^T \mathbf{X}_A \right)^{-1} \mathbf{X}_A^T \mathbf{X}_B \mathbf{F}. \tag{B3}$$

The two terms in B1 are orthogonal by definition and can therefore be treated separately in regression.

y_1	y_2	t
3	1	2
9	1	5
9	-1	4
4	0	2
6	0	3
2.5	-0.5	1
7	1	4
9	1	5

APPENDIX C: CORRELATION VERSUS PARTIAL CORRELATION

If we define $y_1 = t + e_1$ and $y_2 = t + e_2$, where e_2 is equal to $-e_1$, this will be a situation where the residuals have a negative correlation (correlation equal to -1), while the “systematic part” t is the same in both responses (correlation equal to 1). In this case, the systematic part and the random part will go in different direction and then cancel each other. For instance, the data set constructed in this way gives correlation between y_1 and y_2 equal to 0.15 and partial correlation equal to -1 .

APPENDIX D: NUMERICAL ILLUSTRATION OF THE EFFECTS AND THEIR RELATION

D.1 | Data used in simulation

In this section, we give a numerical illustration of the above effects and their relations. Let us consider the same simple model as in Equation 3 *without noise* where \mathbf{X}_C is the response block, \mathbf{X}_A and \mathbf{X}_B are the input blocks, and \mathbf{D} and \mathbf{F} are the regression coefficients.

$$\mathbf{X}_C = \mathbf{X}_A \mathbf{D} + \mathbf{X}_B \mathbf{F}. \quad (\text{D1})$$

The input blocks are generated using random normal numbers with mean 0 and standard deviation equal to 1, of dimensions (10 * 3) and (10 * 4), respectively. The values thus generated were then rescaled by subtracting the column average from the corresponding columns so that the column means have effectively a value of zero. The \mathbf{D} and \mathbf{F} matrices of regression coefficients of dimension (3 * 2) and (4 * 2), respectively, are generated by a uniform distribution on the interval [0.05, 1.05]. It follows that \mathbf{X}_C will be a matrix of dimensions (10 * 2):

$$\mathbf{X}_A = \begin{bmatrix} -0.635 & 1.015 & -0.643 \\ -0.305 & 0.151 & 0.207 \\ 1.484 & 0.192 & -0.601 \\ -0.004 & -0.098 & -0.304 \\ 0.055 & -0.764 & -0.200 \\ 1.640 & 1.578 & -1.262 \\ 0.386 & 0.289 & 1.262 \\ -1.340 & -2.175 & 0.578 \\ -0.761 & 0.493 & -0.714 \\ -0.520 & -0.681 & 1.678 \end{bmatrix}, \mathbf{X}_B = \begin{bmatrix} -0.361 & -0.359 & -0.396 & -0.812 \\ 0.257 & 0.121 & 0.271 & -0.646 \\ -1.286 & -1.069 & -0.156 & 0.971 \\ -1.133 & 0.214 & 0.639 & -1.285 \\ 1.409 & -0.681 & 0.360 & -0.577 \\ -0.704 & 0.481 & 0.294 & 0.777 \\ 1.554 & 0.246 & 0.099 & 0.107 \\ 0.855 & 1.357 & -0.479 & 1.838 \\ 0.184 & -0.123 & 0.066 & -0.319 \\ -0.774 & -0.187 & -0.697 & -0.055 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} 1.039 & 0.120 \\ 0.448 & 0.294 \\ 0.166 & 0.842 \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 0.272 & 0.494 \\ 0.074 & 0.184 \\ 0.257 & 0.441 \\ 0.266 & 0.419 \end{bmatrix}, \mathbf{X}_C = \begin{bmatrix} -0.754 & -1.079 \\ -0.238 & 0.179 \\ 1.317 & -0.766 \\ -0.567 & -1.063 \\ -0.048 & 0.100 \\ 2.328 & -0.206 \\ 1.234 & 2.095 \\ -1.572 & 0.919 \\ -0.716 & -0.584 \\ -0.985 & 0.404 \end{bmatrix}.$$

D.2 | Modifying the model for the purpose of calculating effects

For studying the different effects, one needs to rewrite the model as discussed above, that is, as

$$\mathbf{X}_C = \mathbf{X}_A \tilde{\mathbf{D}} + \mathbf{X}_B^{\text{orthA}} \mathbf{F}, \quad (\text{D2})$$

where $\tilde{\mathbf{D}}$ is obtained as in Equation 6

$$\tilde{\mathbf{D}} = \mathbf{D} + (\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \mathbf{X}_B \mathbf{F}, \quad (\text{D3})$$

$$\tilde{\mathbf{D}} = \begin{bmatrix} 1.133 & 0.246 \\ 0.232 & -0.063 \\ 0.155 & 0.835 \end{bmatrix}.$$

D.3 | The different effects and their relation

D.3.1 | Overall variance of \mathbf{X}_C

The overall variance of the contribution of \mathbf{X}_A and \mathbf{X}_B on \mathbf{X}_C is given by

$$\text{tr}(\mathbf{X}_C^T \mathbf{X}_C) = 22.28. \quad (\text{D4})$$

D.3.2 | The total effect of \mathbf{X}_A on \mathbf{X}_C

We calculate the variance corresponding to the total effect of \mathbf{X}_A on \mathbf{X}_C , defined as the explained variance contribution of $\mathbf{X}_A \tilde{\mathbf{D}}$:

$$\text{tot}_{\mathbf{X}_A} = \text{tr}(\tilde{\mathbf{D}}^T \mathbf{X}_A^T \mathbf{X}_A \tilde{\mathbf{D}}) = 17.42. \quad (\text{D5})$$

D.3.3 | The additional effect of \mathbf{X}_B

The additional effect of \mathbf{X}_B on top of \mathbf{X}_A can be calculated as the variance contribution of $\mathbf{X}_B^{\text{orthA}} \mathbf{F}$:

$$\text{additional}_{\mathbf{X}_B} = \text{tr}(\mathbf{F}^T (\mathbf{X}_B^{\text{OrthA}})^T \mathbf{X}_B^{\text{orthA}} \mathbf{F}) = 4.86. \quad (\text{D6})$$

As can be seen, the total variance of both \mathbf{X}_A and \mathbf{X}_B is exactly equal to the sum of the two, which is should be. Note that the additional and total effect sum to the overall variance.

D.3.4 | The direct effect of \mathbf{X}_A

We can calculate the direct effect as the variance contribution of $\mathbf{X}_A^{\text{orthB}} \mathbf{D}$ obtained by fitting \mathbf{X}_A to $\mathbf{X}_A^{\text{orthB}}$, that is, as

$$\text{direct}_{\mathbf{X}_A} = \text{tr}(\mathbf{D}^T (\mathbf{X}_A^{\text{OrthB}})^T \mathbf{X}_A^{\text{orthB}} \mathbf{D}) = 9.20. \quad (\text{D7})$$

The other way of calculating the direct effect is described in Section 5.4 and gives the same result 9.20 as it should be.

D.3.5 | The indirect effect of \mathbf{X}_A

The indirect effect can be calculated by the differences between tot and direct:

$$\text{indirect}_{\mathbf{X}_A} = \text{tot}_{\mathbf{X}_A} - \text{direct}_{\mathbf{X}_A} = 17.42 - 9.20 = 8.22. \quad (\text{D8})$$

D.3.6 | Relations between all effects

If we add the direct, indirect, and additional effects, we obtain the formula:

$$\text{tr}(\mathbf{X}_C^T \mathbf{X}_C) = 22.28 = \text{direct}_{\mathbf{X}_A} + \text{indirect}_{\mathbf{X}_A} + \text{additional}_{\mathbf{X}_B} = 9.20 + 8.22 + 4.86. \quad (\text{D9})$$